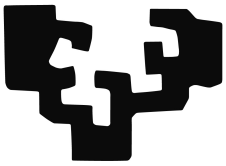
eman ta zabal zazu



EUSKAL HERRIKO UNIBERTSITATEA
Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Doktoretza-tesia

Adimen Artifizialeko metodoak gizarte ikerkuntzarako: analisi demografikoa, jarreren detekzioa eta joera politikoen identifikazioa

Joseba Fernandez de Landa Aguirre

2024

EUSKAL HERRIKO UNIBERTSITATEA
Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Adimen Artifizialeko metodoak gizarte ikerkuntzarako: analisi demografikoa, jarreren detekzioa eta joera politikoen identifikazioa

Joseba Fernandez de Landa Aguirrek Rodrigo Agerriren zuzendaritzapean eginiko tesitxostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Donostia, 2024ko Uztaila.

La utopía está en el horizonte.
Camino dos pasos,
ella se aleja dos pasos
y el horizonte se corre
diez pasos más allá.
¿Entonces para que sirve la utopía?
Para eso, sirve para caminar.

Eduardo Galeano

I have an excellent idea!
Let's change the subject.

The March Hare

Esker onak

Eskerrik asko...

... Rodriri ikerkuntzaren munduan norabidea mantentzen laguntzeagatik.

... Iñaki, Kepa eta Anjelesi ikerkuntzarako ateak zabaltzeagatik.

... Ixa taldeari. Taldekide izandakoei eta direnei, zuen laguntza eta momentu on guztiengatik. Sagardotegixa guztiengatik eta goxo mugitzeagatik. Surf eta eskalada selekzioei.

... Arkaitz for giving me the chance to collaborate with you and hosting me at QMUL. Thanks to all the researchers from CogSci lab specially to Weihe and Andrea.

... ama, aita, Julen eta familia guztiari momentu txar eta onetan hor egoteagatik. Denagatik.

... a Elisa, por animarme siempre y por todo el apoyo incondicional.

... Kutxikume edo, hobeto esanda, Kutxiko fosilei, nire erokeriak aguantatzeagatik.

Lan hau UPV/EHUk diruz lagunduta burutu da (UPV/EHU-PIF19/208).

Abstract

This thesis dissertation explores the intersection of social research and artificial intelligence (AI), investigating how AI technology can be leveraged to enhance the methodology and outcomes of social science studies. The research explores the capabilities of AI, particularly Machine Learning and Natural Language Processing (NLP), to analyze large datasets, identify patterns and infer features that would be challenging to obtain through traditional methods. To achieve this goal, we develop methodologies to automatically characterize social media users leveraging their text content and user interactions thereby enabling more accurate and generalizable predictions. The developed methodologies are then applied to three main applications including demographic characteristic identification, stance detection and political leaning inference. First, this thesis presents the first large scale computational approach for demographic analysis to characterize Basque social media users, including automatic age and community prediction. Second, we exploit the effectiveness of both textual and interaction data to perform stance detection on social media with state-of-the-art results. Specifically, we build the *VaxxStance* dataset, the first crosslingual dataset for stance detection which includes interaction and text data. Furthermore, we present the *Relational Embedding* (RE) interaction-based user representation method which enables to capture user-based features with optimal performance not just for stance detection but also for political leaning inference task. Third, REs outperform every other interaction-based method for multi-class political leaning inference across diverse contexts, allowing to distinguish with high-accuracy between users with different levels of political engagement. Finally, the ability of REs to be effectively combined with textual features demonstrates their robustness and adaptability to perform AI-based social research.

Note for non-Basque speaking readers: Whereas the first half of the thesis is in Basque, the second half and the related articles are in English (Eranskinak). Non-Basque speaking readers are recommended to read the Conclusions (Section 6) to get an overview of the main contributions made in this thesis.

Laburpena

Tesi honek ikerketa sozialaren eta Adimen Artifizialaren (AA) arteko elkarrekintza aztertzen du, AA teknologiak nola baliatu daitezkeen ikertuz gizarte zientzietako ikerkuntza metodologia berritzaileak proposatzeko. Ikerketak AAren gaitasunetan sakontzen du, bereziki ikasketa automatikoa eta hizkuntzaren prozesamendua baliatuta datu multzo handien azterketa, patroien identifikazioa eta ezaugarriak aurreikuspenak lantzeko, metodo tradizionalen bidez burutzea bereziki zaila izango litzatekeena. Horretarako, sare sozialetako testu eta interakzio datuak baliatuta erabiltzaileak automatikoki ezaugarritzeko metodologiak garatu dira. Metodologiek datuen erauzketa eta erabiltzaileen errepresentazioa bilatuko dute, orokortu daitezkeen eta zehatzagoak diren iragarpenak egiteko asmoarekin. Metodo hauen erabilgarritasuna frogatzeko, kasu-azterketak egin dira aplikazio praktikoak burutuz, ezaugarri demografikoen identifikazio, jarreren detekzio eta joera politikoaren inferentzia atazetan. Lehenbizi, euskal erabiltzaileen adina edo komunitateak bezalako ezaugarri sozialak automatikoki iragartzeko eskala handiko lehen hurbilpen konputazionala aurkezten dugu. Bigarrenik, testu-datuen zein interakzio-datuen eraginkortasuna aintzat hartuta, sare sozialetan oinarritutako jarreren detekzio ataza bi datu motak erabilita burutzea erabaki da. Hortaz, jarreren detekzioa hainbat gai eta hizkuntzetan burutzeko asmoarekin, interakzio eta testu datuez osatuta dagoen *VaxxStance* hizkuntza arteko datu-multzoa eta *Relational Embedding* (RE) erabiltzaileen errepresentazio metodoa garatu dira. RE metodoak, hizkuntzarekiko independenteak diren interakzioak baliatzen dituenez, hainbat hizkuntza eta gai ezberdinetan emaitza onenak lortzeko ahalmena dauka. Hirugarrenik eta azkenik, RE metodoa baliatu dugu alderdien araberako joera politikoen identifikazioa burutzeko, interakzioetan oinarritutako bestelako metodoak gaindituz. Gainera, REak errepresentazio testualekin konbinatzean errendimendua hobetzen dela frogatu da, metodoaren moldagarritasuna eta orokortzeko gaitasuna erakutsiz.

Gaien aurkibidea

Taulen zerrenda

Irudien zerrenda

1. KAPITULUA

Sarrera

Azken hamarkadetan, gizarte aldaketa nabarmenak gertatzen ari dira batez ere informazio eta komunikazioen arloan emandako aurrerapen teknologikoek eraginda. Euskal Autonomi Erkidegoan herritarren % 86,3 interneteko erabiltzailea da, azken 10 urteetan baino ia 20 puntu gehiago (Eustat 2022). Sare sozialen erabilerari begira, herritarren % 56,5-ak parte hartzen du bertan, gazteen artean kopuru hori % 90,8-raino ailegatzen delarik (Eustat 2022). Europar Batasunaren kasuan, populazioaren % 90-a dago internetera konektatua eta % 65-a sare sozialen erabiltzaile da (Eurostat 2023b), gazteen kasuan erabilerak % 96 eta % 84-koak izanik hurrenez hurren (Eurostat 2023a). Gizakion artean erlazionatzeko bide berriak ireki dira, oztopo espazial zein denboralak hautsi egin dira eta etengabeko konexioa ahalbidetu da komunitatearekin, nonahi eta noiznahi komunikatuta egoteko aukera irekiz (Castells 2023). Horrela, informazioak funtsezko papera betetzen du ekonomia, ekoizpen eta kontsumoan ez ezik, harreman sozialak eta praktika kulturalak ere moldatzen ditu (Webster 2014). Mundu mailako elkarrekintza sozialen areagotzearekin tokian tokiko gertakariak kilometro askotara gertatzen diren gertaeren arabera moldatzen hasiak dira, espazio eta denbora kontzeptuen disgregazioa sortuz (Giddens 1990). Erronka global hauek estatuen babes tradizionalaren muga gainditu dute, mundu mailan ematen diren arazo politiko zein sozialak geurera ekarriz (Beck 1992). Gizartea entitate dinamiko eta zatikatu batean bilakatzen ari da, non lotura eta identitate tradizionalak harreman arin eta aldakorretan transformatzen ari diren (Bauman 2000).

Eguneroko bizitzan nahi gabe ere egiten ditugun ekintza ugari aldaketa teknologiko esanguratsu hauen emaitza dira: telefono mugikorretik deiak egin, kreditu-

txartelekin erosketak egin, ordenagailuetatik lana egin, bideojokoetara jolastu, oporretako argazkiak publikatu, bideoak ikusi edota konpartitu, medikuaren diagnostikoa ikusi, errezeta bat begiratu, lankideak zoriondu, ezeagunak iraindu, klaseko lanak berridatzi, azterketetarako ikasi, iritzia eman... Ekintza guzti hauek arrasto digitala uzten dute, norbanako zein komunitatearen portaera irudikatzeko baliatu daitezkeenak (Lazer *et al.* 2009). Honen eraginez, milioika pertsonen jarduera sozial, ekonomiko, politiko eta kulturalak modu eraginkorrean digitalizatzen ari dira, datu-multzo zabal eta ugariak sortuz (Hofman *et al.* 2021). Bizitza sozialaren digitalizazio masibo horrek aukera berriak eskaintzen ditu ikerketa sozialerako, baina potentzial hori aprobetxatzeko, gizarte zientzialariek beren metodologiak osatu eta garatu beharko dituzte datu-zientzian garatutako ikuspegiekin (Salganik 2019). Horrela, gizarte zientzietan oraindik zabaltzen ari den "iraultza konputazional" baten hastapenak sortzen ari dira (Lazer *et al.* 2009: 2020; Salganik 2019).

Gizarte zientzia konputazional bat sortzen ari da, datuak bildu eta aztertzeko gaitasuna aurrekaririk gabeko planoetara eramaten ari dena (Lazer *et al.* 2009). Gizarte zientzia konputazionala, beraz, giza jokabidea aztertzen duen diziplina arteko arloa da, horretarako sare sozialetako, interneteko edo digitalizatutako bestelako datu-multzo handietan teknika konputazionalak aplikatzen dituena (Edelmann *et al.* 2020). Adimen Artifiziala gizarte dinamikak testuinguru ezberdinetan nola sortu eta garatzen diren aztertzeko metodologia indartsuak eskaintzen hasi da (Edelmann *et al.* 2020). Metodologia erabilienen artean, ikasketa automatikoan oinarritutako testuen sailkapen eta generazioa daukagu, aurrez anotatutako datuak erabilita datu berrietara proiekzioak egiten dituena (Ziems *et al.* 2023). Arlo honek ospe handia hartu du azken urteetan, ikertzaileentzat garai batean bideraezinak edo erabilgarriak ez ziren eskala handiko datuak eta diseinu esperimentalak erabiliz, milaka ikerketa bultzatuz (Lazer *et al.* 2020).

Gizarte azterketa helburu duten hurbilpen konputazionalek kalitate handiko proiekzio demografikoak egiteko erabiltzen dira batzuetan, datu kantitate handiak eskaintzen baitituzte. Hori dela eta, digitalki sortutako datuak biztanleriaren estimazioak egiteko erabiltzen dira, batez ere estatistika ofiziala fidagarria ez den eremuetan (Cesare *et al.* 2018; Eagle *et al.* 2010; Palmer *et al.* 2013). Datu hauek ere, migrazio prozesuen estimazio zehatzagoak (Palmer *et al.* 2013; Zagheni *et al.* 2017) eta migrazio eta genero bereizketak aztertzeko (Stewart *et al.* 2019; Fatehkia *et al.* 2018) erabiliak izan dira. Aztarna digitalak, osasun publikoari buruzko ikerketak aurrera eramateko datuak ere eskaintzen ditu, hala nola, tabakoaren erabilera aztertzeko (Myslín *et al.* 2013; Huang *et al.* 2014) edo txertoekiko jarrerak ezagutzeko (Salathé and Khandelwal 2011; Larson *et al.* 2013), osasun

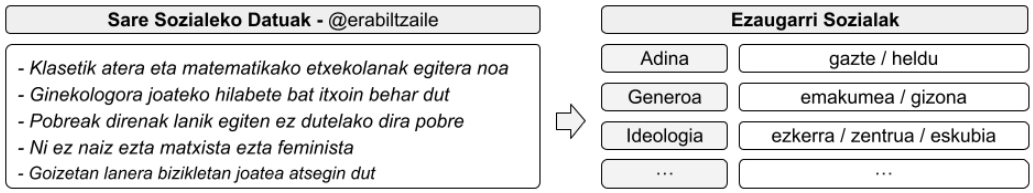
mentalari buruzko ikerketa egiteko (De Choudhury *et al.* 2013; Chancellor and De Choudhury 2020) eta baita gaixotasun infekziosoen inguruko azterketak egiteko ere (Lampos and Cristianini 2010; Corley *et al.* 2010; Culotta 2010; Salathé *et al.* 2013; Santillana *et al.* 2015; Ayyoubzadeh *et al.* 2020; Wicke and Bolognesi 2020). Erabiltzaileen demografiari buruzko ikerketa gehienak Twitter (gaur egun X) sare sozialeko datuak erabiliz egiten dira, ikasketa automatiko gainbegiratuko metodoak baliatuta (Cesare *et al.* 2017). Twitterreko datuen erabileraren arrakasta argitalpenak publikoak zirelako eta, duela gutxi arte, datuen bilketa ahalbidetzen zelako izan da.

Sare sozialetatik erauzitako informazioa datu tradizionalak (inkestak, errolda, erregistroak, elkarrizketak...) osatzeko erabili daiteke, saretik kanpoko prozesu guztiak atzeman ez arren informazio baliotsua eskaintzen baitu. Hala ere, informazio iturri osagarri hauek gizarte-mugimendu askorentzat iturri nagusi gisa balio dute (Zhang and Pan 2019) hauen sareko eta ohiko jardueraren arteko korrespondentzia handia izanik (Abul-Fottouh and Fetner 2018; Hanna 2013). Twitter bezalako sareek gainera datu-multzo handiak eskaintzen dituzte, informazioa bera sareetan nola hedatzen den aztertzeko erabil daitezkeenak (Barberá *et al.* 2015; González-Bailón *et al.* 2013; González-Bailón and Wang 2016). Honekin lotuta, sareko polarizazio politikoa aztertua izaten ari da, fenomeno honek homofiliarekin dituen harremanak ikertuz (Boutyline and Willer 2017; DellaPosta *et al.* 2015). Sareetatik eratorritako testuetan oinarritutako datuen ugaritasunak diskurtso politikoari buruzko ikerketak bultzatu ditu ere, estilo zehatzen eraginkortasuna aztertu (Bail 2015), diskurtso marjinalak arrakastatsu nola bilakatzen diren ikertu (Bonikowski and Gidron 2016) eta migrazioaren kontrako jarrerak nola hedatzen diren analizatuz (Flores 2017).

Gizarte zientzia konputazionalaren jaiotzarekin batera, iraganean balizkoak ziren metodo edota teknikak gaur egungo gizarte dinamiko honetan sinesgarritasuna eta efizientzia galtzen hasiak dira. Adibide bezala, inkesta politikoak jomugan daude, ez direlako gai izan gertatutako hainbat aldaketa aurreikusteko, besteak beste Erresuma Batuko Brexita (Celli *et al.* 2016) edo Estatu Batuetan Donald Trumpen presidentzia (Kennedy *et al.* 2018). Honela, egoera berri eta aldakor honetan, gizartea interpretatu eta ulertzeko tresna berriak proposatzen jarraitu behar direla argi ikusten da. Bereziki garrantzitsua da horrelako ikerkuntza metodologiak garatzea baldintzak azkar aldatzen diren egoeretan, esate baterako, politikagintzan, osasun-krisietan edo gatazka sozialetan. Ikerketa modu azkarrean eta zehaztasunez egin daitezke honelako metodoak aplikatuta. Datu kopuru erraldoiak lortu eta aztertzeko konputazio ahalmena geroz eta handiagoa baita. Bestalde, metodo tradizionalen bidez aztertzeko zailak diren gaiak azter daitezke,

bereziki online munduan ematen diren dinamika geroz eta ohikoagoak. Gainera, ikerketa metodologiak esparru zehatz batean errotuak egon ohi dira, ataza edo ingurune jakin batean emaitza onak lortuz. Hala ere, metodologia hauek bestelako esparru edo domeinuetara moldatzea zaila izaten da, anotatutako datuekiko dependentzia handia baita.

Gabezi horiei aurre egiteko asmoarekin, tesi honetan, gizarte ikerkuntza eta Adimen Artifizialaren arteko konbinaketaz baliatuta, orokortu daitezkeen eta orain artekoa baino zehaztasun altuagoa duten ikerketa metodologiak esploratu nahi dira. Honela, gizakien arteko interakzio birtual eta testu adierazpenen datuekin esperimentazioa eginez, gizakien ezaugarri demografiko edota ideologikoak aurreikusteko metodoak proposatu eta frogatuko dira. Horretarako, sare sozialen inguruneak eskaintzen dituen datu mota ezberdinak erabiliko dira erabiltzaileen ezaugarriak iragarri eta aztertzeko. Ingurune honetan, datuak nola erauzi eta anotatu daitezkeen ikertu da, bertan komunitate zehatzak nola identifikatu eta datuekin nola ordezkatu daitezkeen aztertuz. Ezaugarrien iragarpena ahalbidetzeko, orokortu daitezkeen metodoak aztertu nahi dira, hau da, egoera eta datu ezberdinetan aplikatu daitezkeen metodo dinamikoak, besteak beste, ataza, toki, momentu edota hizkuntza ezberdinetara erraz egokitu daitezkeenak.



1.1 Irudia – Sare sozialetik iragarri daitezken ezaugarriak. Adibidean @erabiltzaile hipotetikoak egindako testu adierazpenak ikusi ditzakegu. Adierazpen hauetatik hainbat ezaugarri sozial aurreikusi daitezke.

Sare sozialetako datuetatik abiatuta, ezaugarri sozial ezberdinak iragarri daitezke, hala nola adina, generoa edota ideologia izanik ohikoenak (Cesare *et al.* 2017). 1.1. irudian ikusi daitekeen moduan, testu datuetatik posiblea da erabiltzaile baten hainbat ezaugarri begi hutsez iragartzea. Helburua ordea, ezaugarri hauek automatikoki iragartzea izango da, ikerkuntza soziala ahalbidetzeko hurbilpenak aztertuz eta garatuz. Horretarako, testu sailkapena oinarri duten atazak erabili eta hobetzen ahalegindu gara, hizkuntzaren prozesamenduan ere aurrera pausuak emanez. Horrela, lan honek hainbat egoera eta ataza ezberdinetan aplikatu daitekeen gizarte ikerketa konputazionaleko teknika berriak nola garatu dai-

tezkeen aztertzen du. Etengabe aldaketan dagoen gizarte digitalizatuaren argazki zehatzagoak eta ulerkorragoak egiten lagunduko duten teknikak hain zuzen ere.

1.1 Helburuak eta ikerketa-lerroak

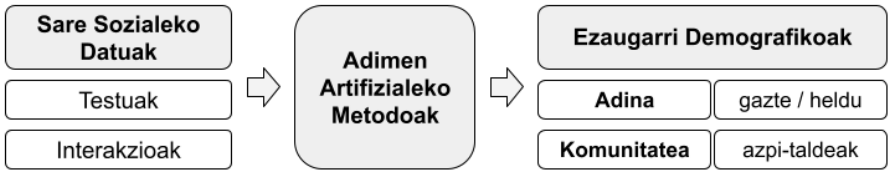
Tesi honen helburu nagusia sare sozialetako erabiltzaileak automatikoki ezaugarritzeko teknikak garatzean zentratuko da. Ezaugarritze horrek erabiltzaileen arteko elkarrekintzak eta publikatutako testuak izango ditu oinarritzat. Datu horiekin erabiltzaileen ezaugarri demografiko zein ideologikoen aurreikuspena burutuko da, errealitate zein egoera desberdinetara moldatzeko ahalmena izango duena.

1.2 Irudia – Erabiltzaileen ezaugarritze automatikoa. Horretarako Adimen Artifizialeko metodoak aplikatu dira. Alde batetik erabiltzaileen ezaugarritzea egin da eta, bestetik, aurrez zehaztutako atazaren ezaugarri sozialen sailkapena.

Horrela, orokortu daitezkeen eta zehatzagoak diren iragarpenak egiteko, ikasketa automatikoa eta hizkuntzaren prozesamenduko metodoak erabiliko dira. Helburu horretara ailegatzeko, pausu hauek jarraitu ditugu: (i) **esplorazio fasean**, erabiltzaileen ezaugarri demografikoen identifikazioa eta komunitate detekzioa landu dira datu jasoketa zehatzak zein testu sailkapen eta erabiltzaileen errepresentazio metodoak baliatuta; (ii) **garapen fasean**, hainbat gai eta hizkuntzetarako erabiltzaile mailako errepresentazio orokor bat lortzeko gai den metodologia garatu da, erabiltzaileen datuetatik informazio sozio-politikoaren errepresentazio bektorialak sortzen dituena; (iii) **aplikazio fasean**, informazio sozio-politikoaren errepresentazioa aplikatu da toki eta egoera ezberdinetan erabiltzaileen joera politikoa zehaztasunez iragartzeko. Zehazki, tesiko ikerketa-lerroak horrela antolatu dira:

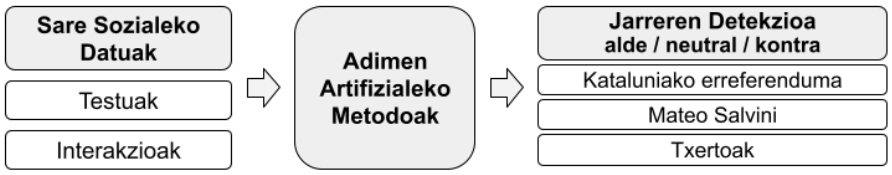
[L1] Ezaugarri demografikoen identifikazioa euskal komunitatean. Gazte euskaldunek sare sozialetan zertaz eta zeinekin aritzen diren aztertzea da ikerketa lerro honen helburua, ezaugarri demografikoen identifikazio atazarekin (§2.2.1) lotuz. Hizkuntza zehatz baten hiztunak identifikatu, era-

biltzaileen adina iradoki eta azpi-komunitateak antzematea hain zuzen ere. Horretarako, Twitter sare sozialean, ezaugarri zehatzak betetzen dituen talde bateko erabiltzaileak nola identifikatu eta hauen datuak nola lortu daitezkeen ikertu da. Erabiltzaileen adina identifikatzeko, testua baliatuta, idazteko modua zein erabiltzaileak anotatu eta hizkuntzaren prozesamenduko metodo aurreratuekin esperimentazioa egin da. Komunitateen identifikaziorako, edukia partekatzean oinarritutako interakzioak erabilita, ikasketa automatiko ez-gainbegiratuarekin esperimentuak egin dira. Laburbilduz, testu eta interakzioak baliatuz, zehaztasun altuko erabiltzaileen ezaugarritzea nola lortu daitekeen landu da.



1.3 Irudia – Ezaugarri demografikoen identifikazio ataza. L1 ikerketa lerroa.

[L2] Jarreren detekzioa hizkuntza eta gai anitzetan. Erabiltzaileen testu eta interakzioak baliatuta jarreren detekzio ataza (§2.2.2) orokortzea da ikerketa lerro honen helburua. Erabiltzaileak ezaugarritzeko orduan testu eta interakzioen arrakasta ikusita, jarrera detekzio ataza testu sekuentzia soiletik atera eta erabiltzaile mailara eraman da. Ataza bera gai edota hizkuntza zehatzetatik independiente egin eta orokortu daitekeen metodologia sendo bat nola garatu landu da. Horretarako, erabiltzaileen errepresentazioan oinarrituko datu bilketa eta sailkapen metodologiak jorratu dira.



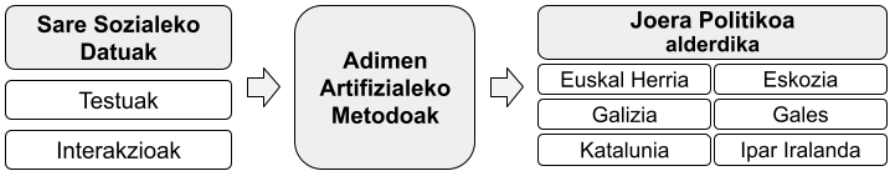
1.4 Irudia – Jarreren detekzio ataza. L2 ikerketa lerroa

[L2.1] Erabiltzaile mailako datu jasoketa hizkuntza anitzetan. Jarreren detekzioa erabiltzaile mailan egiteko, erabiltzaileen interakzioak eta

hizkuntza ezberdinetako testuak jasotzeko metodologia landu da. Horretarako txertoen inguruko jarrera detekzioa aukeratu da, gai berdinaren inguruan datu mota eta hizkuntza ezberdinetan esperimentazioa egiteko helburuarekin.

[L2.2] Interakzioetan oinarritutako ezaugarritzea. Jarrera detekzioan iragarpen zehatzagoak lortzeaz gain, hainbat hizkuntza eta gaietara moldagarriak diren metodologiak landu dira. Ataza honen iragarpenen hobekuntza eta hedakuntzan interakzio datuek eta erabiltzaile mailako hurbilpenak duten eragina aztertu da. Interakzioetan oinarritutako erabiltzaileen errepresentazio aberatsak egiteko metodologia nola garatu ikertu da. Errepresentazio horiek beste testu errepresentazioekin konbinatzeko daukaten ahalmena eta errendimenduan daukaten inpaktua aztertu da.

[L3] Joera politikoaren iragarpena erabiltzaile mailan. Aurreko ikerketa lerroan garatutako errepresentazioak aztertuko dira joera politiko iragarpen ataza (§2.2.3) erabiltzaile mailan burutzeko. Aurretik garatutako datu jasoketa eta erabiltzaileen errepresentazio metodologiak bestelako arloetan aplikatu ahal diren ikertu nahi da. Gainera, joera politikoa, ikuspegi bitar (ezker/eskuin edo liberal/kontserbakor) ohikotik aldendu eta alderdi mailara eraman da, joera anitzetan zein errealitate ezberdinetan aplikagarria izateko. Era honetan analisi sozio-politikoa modu zabalagoan eta zehatzagoan egiteko metodoak landu dira.



1.5 Irudia – Joera politikoaren iragarpen ataza. L3 ikerketa lerroa.

[L3.1] Hurbilpen bitarra eta alderdi politiko anitzetakoa. Joera politikoaren inferentzia modu bitarrean zein alderdietan oinarritutako hurbilpenean aplikatuko da politikoki konplexuak diren lurraldeetan. Erabiltzaileen arteko interakzioak baliatuta, joera politiko bitarra eta alderdietan oinarritutakoaren inferentzia, alderdi politiko ugari dituzten hainbat lurraldeetan aplikatu da. Metodo ezberdinen konparaketa

egin da, gainbegiratze indartsu zein gainbegiratze arineko eszenatokiak kontutan hartuta.

[L3.2] Inplikazio maila desberdinen azterketa. Joera politikoaren inferentzia beste lurralde eta alderdietan aplikatu da, aurreko baldintza berdinez gain, honako honetan inplikazio politikoaren maila ezberdinak ere kontutan hartu dira. Hiru inplikazio maila ezberdin aztertu dira, inplikazio handitik hasi eta inplikazioak txikiagotuz: alderdiko kideak, alderdiko jarraitzaileak eta alderdiko zaleak.

[L3.3] Hurbilpen hibridoa. Joera politikoa aurreko baldintzetan identifikatzeko, interakzio zein testuak erabiltzeak daukan potentziala ikertu da. Datu mota eta ezaugarritze mota ezberdinen arteko konparaketak eta konbinaketak landu dira, hurbilpen hibridoak proposatuz.

1.2 Ekarpen zientifikoak

Atal honetan tesian zehar egin diren ekarpen zientifikoak aurkeztuko ditugu. Egindako ekarpenak landutako atazekin erlazioa izango dute, hau da, ezaugarri demografikoen identifikazioa (§2.2.1), jarreren detekzioa (§2.2.2) eta joera politikoaren iragarpenarekin (§2.2.3) erlazionatuta daude zuzenean. Hala ere, landutako datu bilketa eta erabiltzaile errepresentazio metodoak ataza guztietan zeharka erabiliak izan daitezke, jarraian azalduko den moduan.

- **Ezaugarri demografikoen identifikazioa:** Sare sozialetan euskal hiztunak identifikatu, adina iradoki eta komunitateak antzeman dira. Horretarako *heldugazte-osoa*[1] corpusa argitaratu dugu, erabiltzaile ezberdinen euskarazko 6 milioi publikazioz osatua dagoena. Bestalde, idazkera estiloa (*heldugazte*[2]) eta bizitza etapa (*heldugazte-age*[3]) iragartzeko bi datu-multzo garatu ditugu. Horrela, euskal erabiltzaileak gazteak edo helduak diren identifikatzea ahalbidetuko duten sailkatzaileak entrenatu eta ebaluatu dira. Azkenik, erabiltzaileen interakzioek informazio soziopolitikoa jasotzeko duten potentziala ezagutu dugu. Ekarpen hauek L1 ikerketa-lerroarekin lotuta daude.

- **Jarreren detekzioa:** Gai eta hizkuntza ezberdinetara egokitu daitekeen jarreren detekzioa burutzeko, testu zein erabiltzaile arteko interakzioetan oinarritutako datu-multzo eta erabiltzaileen errepresentazio teknikak garatu dira. Alde batetik, *VaxxStance*[4] datu-multzoa proposatu da, lehenengoa izanik jarreren detekzioa frogatzeko hizkuntza anitzetan eta testu zein interakzio datuak konbinatzen. Bestalde, *Relational Embedding* metodoa proposatu da, interakzioetan oinarritutako erabiltzaileen errepresentazioak baliatuta, hizkuntza eta gai anitzetan jarreren detekzio atazan emaitza onenak lortzen dituena. Bertako ekarpenak L2 ikerketa-lerroarekin lotuta daude.

- **Joera politikoen iragarpena:** Ataza honen ezker-eskubi edo liberal- kontserbakor ikuspegi bitarrarekin amaitu eta berau aztertzeko modu berri bat proposatu da, alderdi politiko instituzionalekin erlazionatzen dena. Era honetan, edozein esparrutara moldagarria izango den joera politikoen iragartze

[1] http://ixa2.si.ehu.es/heldugazte-corpus/heldugazte.osoa.tar.gz
[2] https://github.com/ixa-ehu/heldugazte-corpus
[3] https://github.com/joseba-fdl/heldugazte-age-corpus
[4] https://vaxxstance.github.io/

erreal eta egokitu bat ahalbidetzen da. Horretarako, erabiltzaileen interakzioak baliatuta, testuinguru ezberdinetara moldatu daitezkeen metodologiak frogatu dira, besteak beste *Relational Embedding* metodoa. Lehenik eta behin ikuspegi bitarra (esker-eskubi) eta alderdi anitzekoak konparatu dira erregio ezberdinetan esperimentuak burutuz. Bigarrenik, erabiltzaileen inplikazio politiko ezberdinak kontutan hartuta, erregio berrietan esperimentuak egin dira. Azkenik, interakzioez gain testu datuak ere erabili dira, datu mota eta metodoen arteko konparaketak eginez. Aipatutako pausu bakoitzerako datu-multzo propioak sortu dira, 3 datu-multzo garatuz. Horrez gain, *Basque Twitter Corpus*[5] partekatu da euskarazko 8 milioi publikazioz osatua, sare sozialetako erabiltzaileen testuetatik abiatuta joera politikoa aztertzeko. Aipatutako ekarpenak L3 ikerketa-lerroarekin erlazionatuta daude.

Horrez gain, tesian zehar argitaratutako artikuluak aurkeztuko ditugu. Alde batetik, tesiarekin zuzenean erlazionatutako argitalpenak zerrendatuko dira, eta bestetik, tesiarekin zeharka erlazionatutako artikuluak aurkeztuko dira. Zerrendatutako artikuluak gomendatutako irakurketa ordenaren arabera aurkeztuta daude.

1.2.1 Tesiarekin zuzenean erlazionatutako artikuluak

Atal honetan tesiarekin erlazionatzen diren 9 artikuluak aurkezten dira, baita tesi honen barnean zein ikerketa lerrorekin eta nola erlazionatzen diren ere. Artikuluak eranskinetan aurkitu daitezke, gomendatutako irakurketa ordena jarraituz. Lan guztietatik bost, kongresu edo aldizkari zientifikoetan publikatuta daude eta geratzen diren lauak aldizkari zientifiko ezberdinetara bidalita eta onarpenaren zain daude.

[A.1] Fernandez de Landa *et al.* (2019a)

Fernandez de Landa, J., Agerri, R., & Alegria, I. Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information.* 10 (6), 212. ISSN: 2078-2489.

[5] https://github.com/joseba-fdl/basque_twitter_covid19_corpus

Artikulu hau L1 ikerketa lerroan kokatzen dugu, eta bertan gizarte ikerkuntza burutzeko baliabide eta metodologia berritzaileak ikertu dira, zehazki ezaugarri demografikoen iragarpenarekin lotuz. Era honetan, Adimen Artifiziala ikerketa sozialean nola aplikatu daitekeen ikertzeko lehen urratsak eman dira. Lan honen motibazioa gazte euskaldunen dinamikak aztertzeko nahian oinarritzen da. Horretarako, sare sozialetan komunitate zehatz batekin lotutako erabiltzaileak identifikatzeko eta datuak jasotzeko metodologia proposatzeaz gain, berau aplikatuta *heldugazte-osoa* datu multzo aurkeztu da, 8.000 erabiltzaile eta hauek sortutako euskarazko 6 milioi dokumentu barneratzen dituena. Ondoren, *heldugazte* datu-multzoa proposatu da, testu sekuentziak erregistro informal edo formal arabera anotatuta daudenak. Honekin, testuen erregistroa automatikoki iradokitzeko hainbat sailkatzaile entrenatu eta ebaluatu dira. Gerora, erabiltzaileen idazketa erregistroaren arabera, erabiltzaile gazte eta helduak identifikatu dira, *heldugazte-osoa* multzoko erabiltzaile guztiak automatikoki sailkatuz. Hauen testu eta interakzio datuak erabilita, erabiltzaile gazte zein helduen gai eta erlazionatzeko moduak erauzi dira, horretarako metodo ez-gainbegiratuak aplikatuz eta emaitza kualitatiboki aztertuz. Horrez gain, sortutako datu-multzoak eta sailkatzaileak euskal hizkuntzarako izan arren, proposatutako metodoak baliabide urriko bestelako hizkuntzetan aplikatu daitezke. Guzti honekin, gizarte ikerkuntza Adimen Artifizialeko teknika bitartez burutzea posible dela erakutsi da, egituratu gabeko informazioa kudeatuta interpretagarria den ezagutza sortuz eta ikerketa egiteko modu berriei bide emanez.

[A.2] Fernandez de Landa and Agerri (2021b)

Fernandez de Landa, J., & Agerri, R. (2021). Social analysis of young Basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development.* 1-15. ISSN: 0143-4632 / 1747-7557.

Artikulu hau ere L1 ikerketa lerroan kokatzen dugu, eta bertan aurreko laneko (Fernandez de Landa *et al.* 2019a) metodologien gainean hobekuntzak proposatzeaz gain, baliabide berriak eskaintzen dira. Era honetan, *heldugazte-age* datu-multzoa sortu dugu, erabiltzaile gazte eta helduen arabera hauen euskarazko publikazioak etiketatuz. Datu horiek erabilita, ataza automatikoki burutzeko hizkuntzaren prozesamenduko hainbat sailkatzailerekin esperimentuak egin dira, besteak beste, Transformerretan oinarritutako hizkuntza-eredu elebakar eta elea-

nitzekin (Agerri *et al.* 2020; Devlin *et al.* 2019). Erabiltzaileen adina identifikatze aldera, aurreko laneko eta bertan garatutako metodoak konparatu dira mundu errealeko eszenatokietan, haien errendimendua kualitatiboki aztertzeko eta ebaluatzeko. Gerora, erabiltzaileen interakzioak eta errepresentazio metodo neuronalak erabilita (Grover and Leskovec 2016), erabiltzaileen harremantzeko moduak edo azpitaldeak iragartzea lortu da. Azkenik, interakzio eta errepresentazio metodoen konbinaketari esker, informazio soziopolitikoa jasotzeko daukaten ahalmena ezagutu dugu.

[A.3] Agerri *et al.* (2021)

Agerri, R., Centeno, R., Espinosa, M., Fernandez de Landa, J., & Rodrigo, Á. (2021). VaxxStance@IberLEF 2021: Overview of the Task on Going Beyond Text in Cross-Lingual Stance Detection. *Procesamiento Del Lenguaje Natural*, 67, 173-181. ISSN: 1135-5948 / 1989-7553. *Autoreak alfabetikoki ordenatuta daude.

L2.1 ikerketa lerroan kokatua dagoen artikulu honek, *VaxxStance* ataza eta izen bereko datu-multzoa nola sortu zen deskribatzen du. Atazak, txertoei buruzko adierazpenen jarrera aldekoa, neutrala edo kontrakoa detektatzea proposatzen du. Datu eleaniztunak proposatzen dira, gai berdinari buruz euskarazko eta gaztelaniazko testuak eskainiz. Testuaz gain, interakzio datuak ere jaso dira, informazio sozio-politikoa gehituta (Fernandez de Landa *et al.* 2019a; Fernandez de Landa and Agerri 2021b) ataza honetan emaitza hobeak lortzeko asmoz. Horrela, jarrera detekzio atazarako, eleaniztasuna eta datu mota ezberdinak barnebiltzen dituen lehen datu-multzoa aurkezten dugu. Helburua hizkuntza-arteko ikuspegiak aztertzea da, testuetatik eratorritako informazioa sare sozialetik lortutako erabiltzaileen arteko interakzioekin ere osatuz. Emaitzek frogatzen dute interakzioetatik eratorritako informazio soziala funtsezkoa dela emaitza lehiakorrak lortzeko.

[A.4] Fernandez de Landa and Agerri (2022)

Fernandez de Landa, J., & Agerri, R. Relational Embeddings for Language Independent Stance Detection. *Preprint: arXiv:2210.05715*. Submitted to KBS.

Gerora, L2.2 ikerketa lerroan kokatutako artikulu honetan, jarrera detekzioa landu da, orokortu daitekeen eta zehaztasun gehiagorekin dabilen metodo bat proposatuz. Horretarako, testuaz gain, sare sozialetan eskuragarri dauden interakzio datuen bitartez errepresentazioak lortu dira. Zehatzago esanda, erabiltzaileak jarraitu eta edukia konpartitzea bezalako informazio sozial publikoa baliatu da. Datu hauek erabilita, *Relational Embedding* metodoa proposatzen dugu, interakzio pareak baliatuta errepresentazio bektorial dentso eta adierazgarriak sortzeko gai dena. Metodo hau, hizkuntza eta gai ezberdinetara aplika daiteke eskuzko ingeniaritza handirik gabe. Jarrera detekzio atazan aplikatuz, lau hizkuntza eta hiru gai ezberdin batzen dituzten zazpi datu-multzo publikoetan (*VaxxStance* barne) egindako esperimentuek erakusten dute gure metodoa errepresentazio testualekin konbinatzean errendimendua nabarmen hobetzen dela. Horrela, zazpi ebaluazio multzoetatik seietan puntako emaitzak lortu ditugu, aurre-entrenatutako hizkuntza-eredu (Conneau *et al.* 2020) eta elkarrekintzan oinarritutako DeepWalk (Perozzi *et al.* 2014) edo node2vec (Grover and Leskovec 2016) bezalako metodoak gaindituz.

[A.5] Fernandez de Landa and Agerri (2023)

Fernandez de Landa, J., & Agerri, R. (2023). HiTZ-IXA at PoliticES-IberLEF2023: Document and Sentence Level Text Representations for Demographic Characteristics and Political Ideology Detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org.* ISSN 1613-0073.

Artikulu honetan, L1 eta L3 ikerketa lerroak jorratzen dira, sare sozialetako erabiltzaileen testua baliatuta ezaugarri demografikoak eta joera politikoa iragartzeko metodoak landuz. Bertan, PoliticES 2023 (Garcia-Díaz *et al.* 2023) atazan izan dugun parte-hartzea deskribatzen dugu. Ataza erabiltzaileen publikazio

testualetatik ezaugarri demografiko eta politikoak aurreikustean oinarritzen da: generoa, lanbidea eta lerrokatze politiko bitarra edo lautarra. Horrela, erabiltzaile bakoitzaren hainbat publikazio testual baliatu beharko dira, lau ezaugarri ezberdin aurreikusiz datu berdinetatik abiatuta. Helburu horretarako, bi pausotan banatutako metodologia proposatzen dugu, alde batetik erabiltzaileen errepresentazio orokor bat burutuz eta bigarrenik ezaugarrien araberako berariazko sailkatzaileak entrenatuz ezaugarri sozial bakoitzerako. Testuen errepresentazioak, erabiltzaile eta publikazio mailako errepresentazioak uztartzen ditu, testuetatik jaso daitezkeen alderdi orokorrak (erabiltzaile) eta zehatzak (publikazioa) bateratuz. Metodoaren abantaila gehigarria sendotasun eta orokortzeko gaitasunean datza, errepresentazio berdinak erabilita lehiakortasunez jarduten baitu. Izan ere, proposatutako metodoak bigarren puntuazio altuena lortu du PoliticES 2023 atazarako eta lanbide kategoria aurreikusteko orduan emaitza onenak lortu ditu.

[A.6] Fernandez de Landa *et al.* (2024a)

Fernandez de Landa, J., García-Ferrero, I., Salaberria A., & Campos, J. A. (2024). Uncovering Social Changes of the Basque Speaking Twitter Community During COVID-19 Pandemic. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*. ISBN: 978-2-493814-29-6, ISSN: 2951-2093 (COLING), ISSN: 2522-2686 (LREC).

Artikulu honetan, L3 ikerketa lerroa jorratu da, sare sozialetako erabiltzaileen testuetatik abiatuta, komunitatearen joera politikoaren inguruko analisiak landuz. Hots, ikerketa honen helburu nagusia COVID-19 pandemiak euskarazko Twitter komunitatean dituen ondorioetan sakontzea da, Hizkuntzaren Prozesamenduko teknikak baliatuta. Horretarako, euskarazko 8 milioi publikazioen corpus handia bildu eta partekatu da. Gerora, edukia denboraren arabera nola aldatu den aztertu da, horretarako testuetan azaltzen diren hitz zein emojien aldaketa kuantitatibo zein kualitatiboak aztertuz. Azterketa kuantitatiboan, terminoek garai desberdinetan izan duten maiztasunaren aldaketari erreparatu zaio, maiztasunen erregresio lineala erabiliz. Azterketa kualitatiboan, erabileran gehien handitu diren hitz eta emojien hitz-bektoreen errepresentazioen bilakaera aztertu da, aldaketa semantikoen inguruko analisia burutuz. Hurbilpen hauen bidez, euskal erabiltzaileek pandemiaren zehar erakutsitako joera politikoetan aldaketa nabarmenak aurkitu dira.

[A.7] Fernandez de Landa and Agerri (2024)

Fernandez de Landa, J., & Agerri, R. Political Leaning Inference through Complex Scenarios in Plurinational Spain. *Preprint: arXiv:2406.07964*.

Artikulu hau, berriz, L3.1 ikerketa lerroan kokatua dago eta bertan hedagarria eta zehatza den joera politikoa identifikatzeko metodologia bat proposatzen da. Horrela, alderdi politiko berrien etengabeko agerpenak aro politiko aldakor batean murgildurik gaudela erakusten du, non ezkerra/eskuina edo kontserbadorea/liberala bezalako kategorizazio sinple eta bitarrak jada ez diren baliagarriak. Sinpletasun hori gainditzeko, alderdi politiko anitzetan oinarritutako joera politikoaren identifikazioa proposatzen dugu. Ikuspegi honek, alderdietan oinarrituta izanagatik, eremu eta momentu ezberdinetara moldagarria da. Horretarako, eremu ezberdinak (Euskal Herria, Katalunia eta Galizia) barnebiltzen dituen datumultzoa sortu da, erabiltzaileen orientazio anitza (7 alderdi politiko) zein bitarra (ezker-eskuin) adierazten duten etiketak eta hauen interakzioak jasotzen dituena. Azterketa hau bi pausuetan egin da, lehenbizi interakzioetan oinarritutako erabiltzaileen errepresentazioak burutuz, gerora, erabiltzaile horien joera politiko bitarra zein alderdiak iradokitzeko ereduak sortuz. Horrela, erabiltzaileen errepresentazio metodo ezberdinak frogatu dira, alderdi anitzeko zein ezker-eskuin joera politikoak iragartzeko, haien artean *Relational Embedding* (Fernandez de Landa and Agerri 2022) metodoa ere erabilita. Esperimentuetan ikusi da, *Relational Embedding* metodoa beste errepresentazio metodo batzuekin alderatzean, errendimendua antzekoa dela iragarpen bitarretan, baina oinarrizko metodoen gainbehera handia da alderdi anitzeko iragarpenetan. Izan ere, *Relational Embedding* errepresentazioak erabilita, errendimendu handiko emaitzak lortzen dira 3 eskualde ezberdinetako alderdi anitzeko iragarpenetan baita entrenamendu datu eskasekin ere. Horrez gain, akatsen analisiak eta bistaratzeek erakusten dute *Relational Embedding* errepresentazioak gai direla talde barneko eta talde arteko kidetasun politikoak harrapatzeko.

[A.8] Fernandez de Landa *et al.* (2023)

Fernandez de Landa, J., Zubiaga, A., & Agerri, R. Generalizing Political Leaning Inference to Multi-Party Systems: Insights from the UK Political Landscape. *Preprint: arXiv:2312.01738*. Submitted to IEEE TCSS.

Artikulu hau L3.2 ikerketa lerroan kokatua dago eta aurreko laneko (Fernandez de Landa and Agerri 2024) ikuspegi eta metodoak bestelako eremu batean aplikatzearekin batera, inplikazio politiko desberdinen iragarpena ere proposatzen du. Bertan, Erresuma Batuko hiru naziotan (Eskozia, Gales eta Ipar Irlanda) joera politikoaren inferentzia aztertu da, alderdi anitzez (4 edo 5 alderdi) osaturiko panorama politiko ezberdinak dituzten eskualdeak izanik. Gainera, berrikuntza nabarmen bezala, inplikazio maila ezberdinetako (alderdiko kide, jarraitzaile eta zaleak) erabiltzaileen joerak ere barneratuko dira, inferentzia errealago baten mesedetan. Horretarako, joera politikoaren arabera etiketatutako erabiltzaileek eta elkarren arteko interakzioek osatutako datu multzoa sortu da, baina kasu honetan, inplikazio maila ezberdineko erabiltzaileak gehituta. Erabiltzaileen arteko interakzioak eta *Relational Embedding* metodoa, joera politikoa errepresentatzeko konbinaketa ahaltsua dela frogatu da beste behin ere, emaitza sendoak lortuz alderdi anitzeko sistema politikoak dituzten hiru eskualdeetan, baita entrenamendu datu gutxi dagoenean ere. Horrez gain, inplikazio ezberdinetako joerak aurreikusteko orduan, *Relational Embedding* metodoaren nagusitasuna nabarmena dela ikusi da. Hala ere, inplikazio politiko txikiagoa duten erabiltzaileen joera politikoa iragartzeko orduan hobekuntzarako tartea dagoela ikusi da.

[A.9] Fernandez de Landa *et al.* (2024b)

Fernandez de Landa, J., Zubiaga, A., & Agerri, R. HTIM: Hybrid Text-Interaction Modelling for Broadening Political Leaning Inference in Social Media. *Preprint: arXiv:2406.08201*. Submitted to Plos ONE.

Azkeneko artikulu hau L3.3 ikerketa lerroan kokatua dago eta L3 ikerketa lerroa borobiltzeko asmoa dauka. Hots, interakzioetan oinarritutako datu eta errepresentazioez gain, testuan oinarritutako datuak ere erabiliko dira erabiltzaileen errepresentazioak lortu eta joera politikoa identifikatzeko. Horrela testu eta in-

terakzioetan oinarritutako datu-multzoa sortzen da, aurreko laneko datuak osatuz (Fernandez de Landa *et al.* 2023). Testu eta interakzio datuak eskuragarri edukitzean, datu mota batekiko dependentzia ekiditeaz gain, datu moten arteko konparaketa eta konbinaketa burutzeko aukera ematen du. Esperimentuek erakutsi dute, interakzioetan oinarritutako errepresentazioak, testuan oinarritutako artearen egoerako errepresentazioek baino errendimendu hobea daukatela. Bestalde, datu mota ezberdinen arteko konbinaketa landu da, HTIM proposatuz. Eredu hibrido honek sare sozialetako testu eta interakzioen errepresentazioen fusioa ahalbidetzen du, hainbat lurraldeetako eta alderdi anitzetako joera politikoa zehaztasunez identifikatzeko. Konbinaketa horri esker hobekuntza lortzen da, nabarmenagoa izanik inplikazio politikoa txikiagotzen den heinean, hau da, joera politikoaren identifikazioa zailagoa den heinean.

1.2.2 Tesiarekin zeharka erlazionatutako artikuluak

Ondorengo argitalpenak tesi-garaian idatzitako artikuluak dira, eta tesiaren gai nagusitik aldentzen diren arren, gizarte zientzia konputazionalekin zeharka lotu daitezke.

Fernandez de Landa *et al.* (2019b)

Fernandez de Landa, J., Agerri, R., & Alegria, I. (2019). Euskaldun gazte eta helduen harremanak Twitterren. *III. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Gizarte Zientziak eta Zuzenbidea.* 83-90 or. ISBN: 978-84-8438-681-0.

Fernandez de Landa (2019)

Fernandez de Landa, J. (2019). Gazteak eta euskara sare sozialetan. Zer, nori, nork: euskarazko txio formal eta informalak sailkatuz eta konparatuz. *Eusko Ikaskuntzaren XVIII. Kongresua Geroa Elkar-Ekin: Mendeurreneko Kongresua.* 348-355 or. ISBN: 978-84-8419-293-0.

Fernandez de Landa *et al.* (2021)

Fernandez de Landa, J., García-Ferrero, I., Salaberria A., & Campos, J. A. (2019). Twitterreko Euskal Komunitatearen Eduki Azterketa Pandemia Garaian. *IV. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Ingeniaritza eta Arkitektura.* 137-144 or. ISBN: 978-84-8438-785-5. **Ikergazte 2021 Udalbiltza sari berezia.**

Salaberria *et al.* (2021)

Salaberria, A., Campos, J. A., García-Ferrero, I., & Fernandez de Landa, J. (2021). Itzulpen Automatikoko Sistemen Analisia: Genero Alborapenaren Kasua. *IV. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Ingeniaritza eta Arkitektura.* 153-160 or. ISBN: 978-84-8438-788-6.

Fernandez de Landa and Agerri (2021a)

Fernandez de Landa, J., & Agerri, R. (2021). Euskarazko on-line artikuluetan aipatutako izendun entitate nabarmenen identifikazioa denbora errealean. *Ekaia EHUko Zientzia eta Teknologia aldizkaria* 40, 315-328. ISSN: 0214-9001.

Alkorta *et al.* (2024)

Alkorta, J., Farwell, A., Fernandez de Landa, J., Altuna, B., Estarrona, A., Iruskieta, M., Arregi, X., Goenaga, X., & Arriola, J.M. (2024). CLARIAH-EUS: a Cross-border CLARIAH Node for the Basque Language and Culture. *In SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, CEUR-WS.org.* ISSN 1613-0073. *Autoreak alfabetikoki ordenatuta daude.

Agerri *et al.* (2024)

Agerri, R., Barnes, J., Bengoetxea, J., Calvo, B., and Fernandez de Landa, J., García-Ferrero, I., Toporkov, O., & Zubiaga, I. (2024). HiTZ@ Disargue: Few-shot Learning and Argumentation to Detect and Fight Misinformation in Social Media. *In SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, CEUR-WS.org.* ISSN 1613-0073. *Autoreak alfabetikoki ordenatuta daude.

1.3 Tesiaren egitura

Kapitulu honetan egindako sarreraren ostean, 2. kapituluan tesi honetako lana ulertu ahal izateko oinarrizko aurrekariak zein lan-lerro desberdinekin erlazionatutako literatura aurkeztuko dira. Amaieran, 6. kapituluan tesi lan honetatik ateratako ondorio nagusiak, ekarpenak eta etorkizuneko lana aurkeztuko ditugu.

Bestalde, tesiaren mamia edo edukia bera 3, 4, eta 5 kapituluetan banatuta dago. Bloke honetan tesiaren kapitulu nagusiak aurkeztuko dira, kapitulu bakoitzean ikerketa lerro zehatz bat lantzen da:

- **3. kapituluan**, **ezaugarri demografikoen identifikazioa** landu da. Zehazki, sare sozialetako erabiltzaileen ezaugarri demografiko zein komunitateen identifikazioa nola egin daitekeen aztertu da (L1).

- **4. kapituluan**, **jarreren detekzioa** landu da. Horretarako, hizkuntzarekiko independentea den jarrera detekzioa landu da, ataza erabiltzaile mailara eramanez (L2). Horretarako, erabiltzaile mailako datu jasoketa (L2.1) eta interakzioetan oinarritutako ezaugarritzea (L2.2) landuz.

- **5. kapituluan**, **joera politikoaren identifikazioa** landu da. Horretarako, erabiltzaile mailako joera politikoaren identifikazio dinamikoa aztertu da (L3). Interakzioetan oinarrituta, hurbilpen bitar zein alderdi politiko anitzetakoak konparatu dira (L3.1), gerora inplikazio maila desberdinak ere aztertuz (L3.2). Amaitzeko, baldintza horietan, interakzio eta testuetan oinarritutako hurbilpenak frogatu dira datuen hibridazioaren eragina aztertzeko (L3.3).

Tesia bi hizkuntzatan banatua dago: euskara eta ingelesa. 1, 2, eta 3 kapituluak euskaraz idatzita daude. 4, 5, eta 6 kapituluak, berriz, ingelesez. Gainera, tesiarekin erlazio zuzena daukaten (1.2.1. atala) artikulu guztiak ingelesez eskuragarri daude Eranskinetan.

2. KAPITULUA

Aurrekariak

Atal honetan, egoera eta ataza ezberdinetara moldatu daitekeen gizarte ikerkuntza konputazionala ulertu ahal izateko, beharrezko aurrekariak aurkeztuko ditugu. Atala bi zati ezberdinetan banatuko dugu. Alde batetik, erabiltzaileen ezaugarritzea burutzeko beharrezkoak diren oinarriak azalduko dira, datu jasoketatik hasi eta erabiltzaileen errepresentazio zehatzak burutzeko erabilgarriak diren hurbilpenak erakutsiz. Beste aldetik, erabiltzaileen ezaugarrien iragarpena demografia, jarrera edo joera politikoa bezalako ataza ezberdinetan aplikatu denez, ataza bakoitzarekin erlazionatutako lanak aztertuko dira, hauek testuinguruan jartzeko.

2.1 Oinarriak

Hedatu edo orokortu daitezkeen erabiltzaileen errepresentazioetatik ezaugarritze ezberdinak lortzeko asmoarekin aztertu eta erabilitako hurbilpen zein teknikak modu honetara antolatu ditugu: (i) Datu iturrien zehaztapena datuak lortzeko, (ii) datuen tratamendua ahalbidetzeko Ikasketa Automatikoaren aplikazioa eta, azkenik, errepresentazioa modu ezberdinak (iii) testuen zein (iv) interakzioen bidez.

2.1.1 Datu Iturriak

Sare sozialetako datu-multzo erraldoiak gizarte ikerketarako erabiltzeak hainbat onura ditu. Lehenik eta behin, datuen berehalakotasuna edukiko genuke, izan ere erabiltzaileen adierazpenen eguneraketak denbora errealean ematen direnez,

datu bilketa tradizionalen bidez ikertzeko zailak diren jokabideak aztertzeko aukera ematen dute (Golder and Macy 2014). Bigarrenik, sare sozialetako argitalpenak espontaneoak direnez, inkesten eta elkarrizketen testuinguruan agertzen diren erantzun alborapenak (Belli *et al.* 1999) ekidin eta fideltasun handiagoz jaso ditzakegu iritzi eta jokabideak. Hirugarrenik, inkestekin eta ohiko datu-bilketarekin alderatuta, datu hauek kostu baxukoak dira eta datuak biltzeko prozesua automatizatu daiteke (Cesare *et al.* 2017).

Hala ere, muga nabarmenak daude eskala handiko datu horien erabilerari dagokionez. Alde batetik, datu-korronte tradizionalen antzera, sare sozialen datuak ez dira beti ikertu nahi den populazioarekiko adierazgarriak, eta arazo hori konpontzen saiatu diren arren (Wang *et al.* 2015), alborapena identifikatu eta kuantifikatzea zaila da. Bestalde, ikerketarako sare sozialen datuen erabilerarekin lotutako pribatutasun eta kezka etikoak zabaltzen ari dira (Boyd and Crawford 2012; Vayena *et al.* 2015; Edelmann *et al.* 2020; Caton and Haas 2020), aldi berean ikerketarako lerro berriak irekiz. Horretarako, datu zientifikoak bidezko modu batez kudeatzeko printzipioak (FAIR) proposatu dira, aktibo digitalen aurkigarritasuna, irisgarritasuna, interoperabilitatea eta berrera-bilpena hobetzeko jarraibideak (Wilkinson *et al.* 2016).

Horiek horrela, baina, sare sozialak baliabide gutxiko hizkuntzen erabilera ikertzeko datu iturri aproposa dira, ikerketa anitz publikatu direlarik. Adibidez, galesera (*Cymraeg*) hiztunen eta Twitterren inguruko ikerketa batek erakusten du hizkuntza honetako hiztunak sare sozialetan ere aktibo daudela (Jones *et al.* 2013). Horrez gain, irlanderaz (*Gaeilge*) 80.000 txio baino gehiago erauzten eta aztertzen dituen beste lan bat dago eduki, sentimendu eta sarearen azterketa egiten duena (Mhichíl *et al.* 2018). Halaber, galesera, irlandera eta frisiera (*Frysk*) konbinatzen dituen beste ikerketa bat aurkitu dezakegu, txio ezberdinetan hashtag-en erabilera ikertzen duena (McMonagle *et al.* 2019). Aipatutako lan hauek erakusten dute Twitterrek baliabide gutxiko hizkuntzetarako ere baduela testu-datuak eskaintzeko ahalmena, hizkuntza eta kultura ugari aurkitu eta aztertzeko aukera emanez.

Gainera, sare sozialetatik jasotako testuak oso erabiliak dira hizkuntzaren prozesamenduan, produktu edo gai zehatzei buruzko iritziak erauzteko (Villena *et al.* 2013; Rosenthal *et al.* 2017), joera politikoak (Mohammad *et al.* 2016; Derczynski *et al.* 2017) eta gorroto hizkera identifikatzeko (Basile *et al.* 2019) edota oinarrizko zereginetarako, hala nola POS etiketatzea (Ritter *et al.* 2011), izendun entitateen identifikazioa (Baldwin *et al.* 2015), normalizazioa (Alegria *et al.* 2015) eta baita hizkuntzen identifikaziorako ere (Zubiaga *et al.* 2016).

Bestalde, hainbat dira Twitter sare sozialetik eratorritako erabiltzaileen arteko interakzioetan (jarraitu, birtxiokatu, erantzun, aipatu, atsegin....) oinarrituta buru-

tu diren ikerketak. Horien artean nabarmendu genitzake polarizazio politikoa aztertzeko (Conover *et al.* 2011b), afiliazio politikoa identifikatzeko (Pennacchiotti and Popescu 2011a), eta baita independentziaren aldeko mugimenduak aztertzeko (Zubiaga *et al.* 2017) erabilitakoak. Ikerketa hauetan erakusten da, erabiltzaileek egindako interakzioetan oinarrituta, badagoela komunitateak edo taldeak iragartzeko aukera. Beraz, lan zehatz honetan, metodologia antzekoak erabiliko ditugu, euskal komunitatearen baitan komunitateak nola ematen diren aurreikusteko, hau da, harremanak nola ematen diren ikusteko.

2.1.2 Ikasketa Automatikoa

Gizarte zientzien arloan datuak aztertzea eta ulertzea funtsezkoa da teoria berriak garatzeko eta gizartearen portaerak iragartzeko. Azken urteotan, Adimen Artifizialak garrantzi handia hartu du gizarte zientzietan, datu masiboak modu eraginkorrean prozesatu eta analizatzeko aukera ematen baitute. Honek, ikerlariei gizarte fenomeno konplexuak sakonago ulertzeko eta iragarpen zehatzagoak eta berehalakoak egiteko bide berriak ireki dizkie.

Adimen Artifiziala informatikaren adar bat da, eragiketa konputazional ugari biltzen dituena, erregela edo estatistikan oinarritutako metodo tradizionaletatik hasi eta Ikasketa Automatiko zein Ikasketa Sakoneko hurbilpenak ere barnebilduz (Russell and Norvig 2016; Kitchin 2014). Erregeletan oinarritutako sistemek aurrez definitutako arau multzo bat erabiltzen dute iragarpenak egiteko. Hau da, adituek eskuz idatzitako arauak dira, eta sistema hauek arau horiei jarraituz funtzionatzen dute. Ikasketa automatikoaren kasuan, aldiz, helburua datumultzoetan patroiak aurkitu eta ereduak sortzea da, etorkizuneko datuak aurreikusteko edo sailkatzeko. Honek aukera ematen du datu masiboak modu eraginkorrean erabiltzeko, gizakien esku-hartze gutxirekin edo batere gabe.

Ikasketa automatikoko sistemek sarrera datuetatik zuzenean "ikasten" dute inolako erregelarik programatu gabe. Izan ere, sistemak bere kabuz ikasten du datu-multzoetatik zuzenean patroiak identifikatuz. Funtzio matematikoetan oinarritutako algoritmoak erabiltzen dira datu horietatik ereduak ikasi edo entrenatzeko. Eredu hauek ahalmena daukate ezkutuko irudikapenak aurkitzen dituzten funtzio matematikoak estimatzeko, gerora datu berriei buruzko iragarpenak egin ahal izateko (Jordan and Mitchell 2015). Ikasketa automatikoa bi kategoriatan banatu dezakegu erabilitako datuen egituraketa kontutan hartuta: (i) gainbegiratutako ikasketa, non eredu bat entrenatzen den etiketatutako datuetan oinarrituta; (ii) gainbegiratu gabeko ikasketa, non datuetatik zuzenean eredu bat eraikitzen den etiketarik gabe. Horrela modalitateek helburu zehatz ezberdina daukaten arren,

ezaugarrien aurreikuspena zein ezaugarrien errepresentazioa hurrenez horren, helburu orokor bera konpartitzen dute: datu-multzo berrietara aplikatu daitezkeen ereduen sorrera.

Ikasketa gainbegiratuan, sarrera datu bat emanda aukeratutako ezaugarriak iragartzeko ahalmena duten ereduak sortzea da asmoa (Kotsiantis *et al.* 2007; Singh *et al.* 2016). Horretarako, aurrez etiketatutako datu-multzoak bat behar dira, zeinetan datuak aukeratutako ezaugarrien arabera etiketatuta dauden. Adibidez, testu zatiak sarrera datu bezala edukita aldeko edo kontrako polaritatea etiketatzea edo erabiltzaile baten interakzioak edukita sarrera bezala honen lerrokatze politikoa eskuz etiketatzea. Datu-multzoak eskuz etiketatzen dira normalean, horretarako etiketatu behar diren ezaugarrietan adituak diren anotatzaileak baliatuz. Eredua prestatzeko, etiketatutako datu-multzoa entrenamendu eta ebaluazio multzoetan banatzen da. Entrenamendu datuak erabiliz, aurrez zehaztutako algoritmoa eredu matematiko bat sortzen saiatzen da, zeinak sarrerako datuak eta etiketak erlazionatuko dituen. Algoritmoen artean erregresio logistikoa zein euskarri bektoredun makinak daude beste askoren artean (Pedregosa *et al.* 2011). Ereduak hainbat iterazio egiten ditu, sarrera bakoitzarentzat aurreikuspen bat eginez eta benetako etiketarekin alderatuz, errorea ahalik eta txikiena izan arte. Eredua behin entrenatuta, ebaluazio multzoan probatzen da, datu berrietan ereduaren portaera nolakoa izango den frogatzeko. Ebaluaketarako, iragarpen datuak datu-multzoko etiketa errealekin alderatzen dira, sortutako ereduaren kalitatea neurtuz. Emaitzak onargarriak badira, eredua erabiltzeko prest dago, datu berrietan aurrez finkatutako ezaugarriak aurreikusteko gai den tresna gisa.

Bestalde, gainbegiratu gabeko ikaskuntza, datu-multzo handiak aztertu eta horien egitura edo patroi ezkutuak aurkitzeko tekniken multzoa da. Hau da, algoritmo hauek ez dira etiketatutako datuetan oinarritzen, baizik eta datuen arteko antzekotasunak, desberdintasunak eta egiturak identifikatzeko asmoz erabiltzen dira. Helburua datuen arteko erlazioak ulertzea zein bertatik informazioa egituratzea da, datuetan begi hutsez hauteman ezin diren patroi ezkutuak aurkitzeko. Gainbegiratu gabeko ikaskuntzaren adibide bat klusterizazioa da, non datuak antzekotasunen arabera taldeetan banatzen diren. Klusterizazio teknikak, hala nola *k-means* (Forgy 1965) edo *hierarkikoa* (Defays 1977), datuen arteko distantziak edo antzekotasunak kalkulatzen dituzte, eta horien oinarrian talde edo komunitateak eratzen dituzte. Beste adibide bat dimentsio-murrizketa da, PCA (Osagai Nagusien Analisia), t-SNE (Van der Maaten and Hinton 2008) edo UMAP (McInnes *et al.* 2018) bezalako teknikak baliatuta, non datu-multzo handi baten dimentsioak murrizten diren, baina bere ezaugarri garrantzitsuenak mantentzen diren. Teknika hauek erabilgarriak dira errealitateko arazo konplexuak hobeto ulertzeko, non

datuen atzean dauden egiturak eta patroiak aurkitzea zaila den. Horrela egoera zehatz baten analisi kualitatiboa ahalbidetzeaz gain, errepresentazioak egiteko ere baliatu daitezke.

Bestalde, entzute handia izaten ari den ikasketa sakona ikasketa automatikoaren azpimultzo bat da, neurona sare sakonak erabiltzen dituena (LeCun *et al.* 2015; Goodfellow *et al.* 2016; Alpaydin 2021). Neurona sareek hainbat geruza dituzte, eta datuak geruza horien bidez igarotzen direnean, eredu konplexuak identifikatzeko gaitasuna dute. Eredu hauek datu kopuru handiak behar dituzte entrenamendurako, eta horretarako, konputazio ahalmen handia eta optimizazio algoritmoak erabiltzen dira, emaitza zehatz eta fidagarriak lortzeko. Hau bereziki erabilgarria da irudiak, ahotsa edo testua bezalako datu konplexuen analisi eta ulermenerako.

Aurreko atalean aipatu bezala, tesi lan honetan zehar erabiliko diren sarrera datuak testu zein interakzio datuetan oinarrituta egongo dira. Horregatik, datu mota hauekin kalitatezko errepresentazioak erauzteko moduak aztertuko dira. Horretarako, erabiltzaileen testu eta interakzioetatik abiatuta, hauen errepresentazioak hobekien burutzeko gainbegiratu gabeko teknikak bilatuko dira. Errepresentazio hauekin gainbegiratutako ikasketa algoritmoak elikatzea izango da asmoa, aukeratzen diren ezaugarri sozialak aurreikusteko kapazak diren sailkatzaileak entrenatzeko asmoarekin. Horrela, orokortu daitezkeen errepresentazio moduak topatu nahi dira, metodo berdina jarraituta hainbat ezaugarri sozial aurreikusteko gai izateko.

2.1.3 Testuen errepresentazioak

Sare sozialetan publikatzen diren testu edukietatik sailkatzailea elikatuko duten sarrera datuak lortzeko, hizkuntzaren prozesamenduko metodoak baliatuko dira. Metodo hauetako batzuk hitzen edo hitzen-sekuentzien errepresentazioak egitean oinarritzen dira, hitzetatik zenbakizko errepresentazioetara joz, gerora sailkatzaileak zenbakizko datuekin elikatzeko asmoarekin. Jarraian, tesi lan honetan erabili diren hitz edo testu sekuentziak kodetzeko erabilitako metodoak aurkezten dira: (i) errepresentazio estatistikoak; (ii) hitz-bektoreak eta (iii) testuinguruaren araberako errepresentazioak.

Errepresentazio estatistikoak

Hitz-zaku (ingelesezko bag-of-words edo BoW) ereduak hizkuntzaren prozesamenduan testuetan agertzen diren hitzak errepresentatzeko oinarrizko tekniketako bat da. Dokumentu edo esaldi bateko testua, barnean dituen hitzen zaku bezala

adierazten dira, hau da, zein hitz eta zenbat agerpen eduki dituzten jasotzen da. Agerpenak eta kontaketak mantendu arren, hitzen arteko ordena eta haien arteko harremanak baztertzen dira. Dokumentuak, aurrez zehaztutako hiztegi batean azaltzen diren hitzen kontaketekin ordezkatuko dira, hauekin errepresentazio sinpleak sortuz. Hala ere, terminoen maiztasun orokorrak jasotzen duten metodo hauek gehienetan ez dira testuaren adierazpide egokia. Hitz hutsalak, adibidez euskarazko "*eta*" edo "*da*", ia dokumentu guztietan maiztasun altuena duten hitzak dira, eta ondorioz errepresentazioetan pisu altua daukate, baino dokumentuaren inguruko informazio gutxi adierazten dute.

TF-IDF (ingelesezko Term Frequency-Inverse Document Frequency) algoritmoa (Salton and Yu 1973), edo termino maiztasuna alderantzizko dokumentu maiztasuna, terminoen maiztasunak neurtzen ditu baina dokumentu multzoan ematen diren agerpenak kontutan hartuta. Honi esker, dokumentu guztietan zehar agerpen maiztasun handiak dituzten hitzei garrantzia txikitzea die, kasu hauetan maiztasun handia ez baita esanguratsua. Horrekin batera, termino batek maiztasun orokor txikia badu, baina dokumentu zehatz batean askotan agertzen bada, hitz horri garrantzia handituko zaio. Maiztasunak eta terminoen garrantzi erlatiboa jasotzeko, algoritmoa bi zatiz osatuta dago, batetik hitzen maiztasuna (TF) eta bestetik alderantzizko dokumentu maiztasuna (IDF). Alde batetik, terminoen maiztasunak (TF) dokumentu batean termino zehatz batek duen maiztasuna adierazten du. Normalean termino zehatz bat dokumentuan dauzkan agerpenak, dokumentuko termino guztien kopuruarekin zatituta kalkulatzen da. Beste aldetik, alderantzizko dokumentu maiztasunak (IDF) termino baten urritasuna neurtzen du corpusean. Dokumentu kopuru osoaren logaritmo gisa kalkulatzen da terminoa duten dokumentu kopuruarekin zatituta. Era honetan, aurrez zehaztutako hiztegi bateko hitz edo terminoen tfidf maiztasunak kalkulatzen dira dokumentu guztietan zehar. Dokumentu edo testu zati bakoitzerako, tamaina zehatz bateko errepresentazioak edukiko dira, tamaina hau hiztegiaren tamaina berdina izanik. Ordea, hitz-zakuaren teknikarekin baino adierazpen zehatzagoak lortu arren, hautatutako hiztegiaren araberako eta honen tamainarekiko menpekoa izango da metodo hau.

Oinarrizko metodo hauetan, hitzen maiztasunak hartzen da kontuan, dokumentu edo testu zati bakoitzaren errepresentazioa dimentsio oso altuko ezaugarrien bitartez jasotzen delarik. Hala ere, errepresentazioak eskasak eta sakabanaketa handikoak izan daitezke, batez ere hiztegiak handiak direnean, horrek konputazio erronka eta memoria errekerimendu handiak sortuz. Gainera, egoera edo domeinu ezberdinetan parametro ezberdinak erabili behar dira, ezaugarri kopurua edota termino ohikoenen zein urrienen filtratzea bezalakoak. Amaitzeko, hitzen agerkidetzak azaldu eta zenbatzen diren arren, ez dira hitzen arteko erlazio seman-

tikoak jasotzen, hizkuntzaren fenomeno garrantzitsuenetako bat kanpoan utziz.

Hitz-bektoreak

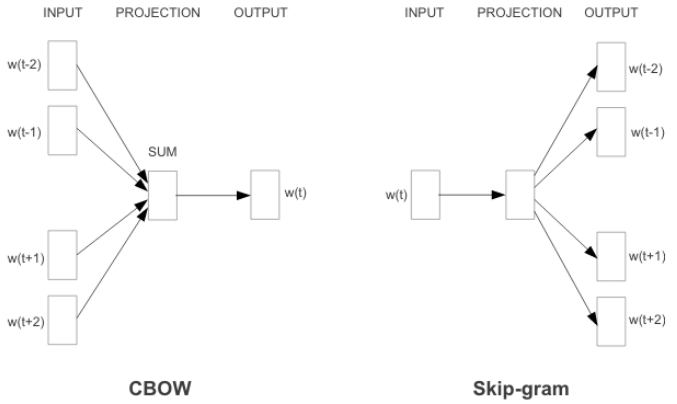
Hitzen bektore errepresentazio jarraituak, hitz-bektore gisa ere ezagutuak, hitzen informazio semantiko eta sintaktikoa jasotzen duten hitzen zenbakizko irudikapenak dira. Hitz-bektore gisa ezagututako errepresentazio hauek Word2Vec (Mikolov *et al.* 2013b), GloVe (Pennington *et al.* 2014) edo FastText (Bojanowski *et al.* 2017) bezalako tekniken bidez sortzen dira. Horrela, hitzak matrize eskas eta sakabanatuen bidez errepresentatu ordez, espazio jarraitu batean kokatzen dira dimentsio baxuko bektoreak baliatuta. Honi esker, hitzak entitate diskretu gisa tratatu ordez, entitate jarraitu bezala tratatuko dira, hitzen arteko erlazio semantikoak kontuan hartuz.

Izan ere, bektore hauek hitzen arteko erlazioak harrapatzeko diseinatuta daude, antzekotasuna eta analogia adibidez. Horrela, "*Gasteiz*", "*Donostia*" edo "*Londres*" hitz-bektoreen balioak gertu egongo dira elkarrengandik, "*zuzendari*", "*sagarra*" zein "*autobusa*" hitz-bektoreen balioak elkarrengandik ezberdinak izango diren bitartean. Horri esker, esanahi antzekoa duten hitzek, hitz-bektore antzekoa izango dute, hitzen errepresentazio jarraitu bat lortuz.

Word2vec (Mikolov *et al.* 2013b: a) ereduak hitzen errepresentazio jarraitu edo hitz-bektoreak modu azkar eta eraginkorrean ikastea ahalbidetzen du. Honek, autogainbegiraketa bitartez, testu multzo erraldoien gainean kalitatezko hitz-bektoreak entrenatzea posible egiten du. Horretarako, testu multzoan azaltzen den hitz zehatz bat eta bere inguruan dauden hitzen arteko probabilitatearen logaritmoak aurresaten ikasten dute. Ikasketa helburua modu ezberdinetan planteatzen da, horretarako *CBOW* eta *skip-gram* (2.1. irudian) arkitekturak proposatuta.

CBOW ereduak helburu-hitza bere inguruko hitzetatik aurreikusten du, inguruko hitzak erabiltzen ditu erdian dagoen helburu-hitza iragartzeko. *Skip-gram* ereduak helburu-hitz batetik inguruko hitzak aurreikusten ditu, hitz bakar bat erabiltzen du bere aurretik eta ostean dauden hitzak iragartzeko. Ereduak konparatzean, *CBOW* eredua azkarrago entrenatu eta maiztasun handiagoa duten hitzak hobeto irudikatu ditzake, *Skip-gram* ereduak datu-multzo txikiekin errendimendu hobea erakutsi eta maiztasun gutxiagoko hitzak hobeto errepresentatzen dituen bitartean (Mikolov *et al.* 2013a).

FastText (Bojanowski *et al.* 2017) ereduak aurreko ereduen gainean hobekuntza bat proposatzen du, hitzen errepresantazioarekin batera karaktere n-gramaz osatutako azpi-hitz unitateak txertatuz. Horretarako, hitz bakoitzaren karaktere n-grama kate bat bezala tratatzen da, corpuseko hitzak zatikatuz. Hitzak eta beren



2.1 Irudia – CBOW (inguruko hitzetatik, helburu-hitza aurreikusi) eta Skip-gram (helburu-hitzetik, inguruko hitzak aurreikusi).

karaktere n-gramak erabiliz, *skip-gram* eredu bat entrenatzen da errepresentazioak ikasteko. Hitz-bektoreak, karaktere n-gramen eta hitz beraren errepresentazioaren batura gisa adierazten dira, azpi-hitzen informazioa baliatuz hitzen errepresentazioak aberasteko. Honi esker, hitz-bektore hauetan informazio morfologikoaren txertaketa ahalbidetzen da aurrizki, artizki zein atzizkien karaktere-segidak kontutan hartzen baitira. Euskara bezalako morfologia aberatseko hizkuntzek, batez ere, hitzen (edo azpi-hitzen) errepresentazio hauetatik etekina atera beharko lukete. Izan ere, hitz beraren forma ezberdinen artean (Gasteiz, Gasteiz*tar*, Gasteiz*ko*, Gasteiz*en*...) informazioa partekatzea ahalbidetzen da. Horrez gain, hiztegiz kanpoko hitzak (edo hitz-bektoreak ikasi diren momentuan hiztegitik kanpo geratu diren hitzak) hainbat azpi-hitz edota hitz konbinatuta ordezkatu daitezke.

Testuingurudun errepresentazioak

Arestian aipatutako planteamenduek, hitz-bektore estatikoak proposatzen dituzte (Mikolov *et al.* 2013b; Bojanowski *et al.* 2017), hots, hitz jakin baterako bektoreetan oinarritutako errepresentazioak eskaintzen dira, hitzaren errepresentazioa gertatzen den testuingurutik independentea dena. Horrek esan nahi du ezin dela polisemia irudikatu. Beraz, "*banku*" hitza kontuan hartzen badugu, hitz-bektore estatikoek errepresentazio bakarra sortuko dute, nahiz eta hitz horrek zentzu desberdinak izan, hots, "*finantza erakunde*", "*eserleku*", etab.

Arazo horri aurre egiteko, testuingurudun hitz-bektoreak proposatzen dira,

Flair (Akbik *et al.* 2018), ELMO (Peters *et al.* 2018) edo Transformer (Vaswani *et al.* 2017) arkitekturan oinarritutako BERT (Devlin *et al.* 2019) eta RoBERTa (Liu *et al.* 2019) adibidez. Era honetan, hitz-bektore errepresentazioak sortzen dira, baina, hitz zehatzaren momentuko testuingurua ere barneratuz. Horrela, hitz bakoitzaren errepresentazioa beti finko edo estatiko mantendu ordez, inguruan dituen hitzen arabera moldatzen joango da, errepresentazio adierazgarriagoak eta zehatzagoak lortuz.

Flair (Akbik *et al.* 2018) hitz-bektoreak, testuinguru eta karaktere mailako informazioan oinarrituriko hitzen errepresentazioak dira. Horretarako, bi norabideko LSTM (Long Short-Term Memory) ereduak (Graves *et al.* 2013) baliatzen dira, ikasketa prozesuan zehar hizkuntza-eredu bat sortzen duelarik. Horrela, ikasketa prozesuan, aurreko karaktereen arabera hurrengo karakterearen iragarpena eginez trebatzen da, karaktere mailako hizkuntza-eredua sortuz. Karaktere mailako ereduek hitzen barneko egitura aztertzen dute, horrek akats ortografikoak edo hitz berrien esanahia ulertzea errazagoa egiten du. Gainera, bi-norabideko trataerari esker, hitz zehatz baten aurreko eta atzeko testuingurua zein honen posizioa jasotzeko ahalmena edukiko du. Hitz baten errepresentazioa ezagutzeko orduan, hitz-bektorearen balioa aldatu egingo da inguruko hitzen arabera, testuingurudun hitz-bektoreak lortuz. Izan ere, bi-norabideko trataerari esker, hitzen arteko harreman korapilatsuak zein polisemia bezalako kontzeptuak harrapatzeko aukera ematen da.

BERT (Devlin *et al.* 2019) hizkuntza-ereduak *Transformer*rak (Vaswani *et al.* 2017) erabiltzen ditu, arreta mekanismoaren bidez, testu bateko hitzen (edo azpihitzen) arteko erlazioak ikasteko. Horretarako, pilatutako hainbat Transformer kodetzaile geruza erabiltzen dira, testu zatien ezagutza maila ezberdinak jaso eta testu zati osoak prozesatzeko ahalmena emanez. Hizkuntza-eredua sortzeko egiturarik gabeko testuan oinarrituta bi zereginetan trebatzen da, maskaratutako hizkuntza eredu (Masked Language Model, MLM) eta hurrengo esaldiaren iragarpen (Next Sentence Prediction, NSP) atazetan. Alde batetik, maskaratutako hizkuntza eredu atazan, testuko hitzak ausaz maskaratzen (ezkutatu) dira eta eredua hitz horiek asmatzen saiatzen da. Horrela, ereduak emandako testuinguru osotik ikasten du eta hitzen arteko harremanak hobeto errepresentatzen ditu. Beste aldetik, hurrengo esaldiaren iragarpen atazan, bi esaldi jarraituren artean hurrenkera egokia asmatzea da helburua.

Gainera, errepresentazio dinamiko hauei esker, aukera dago aurretik sortuta dagoen hizkuntza eredu bat doitzeko, berri bat hutsetik entrenatu beharrean. Horrela, kalitate altuko hizkuntza eredu orokor batetik abiatuta, domeinu zehatz batera moldatu daitekeen hizkuntza eredu bat lortu daiteke. Izan ere, Flair edo

BERT, hasiera batean etiketarik gabeko testu kopuru handi batean trebatuak dira, testuinguru zabala barneratzen duten hitzen errepresentazio orokorrak ikasteko. Aurrez entrenatutako eredu horiek, jarraian, ataza edo zeregin zehatzetara doitu daitezke, entrenamenduarekin jarraituz. Doikuntza prozesua hainbat atazetan aplikatu daiteke, adibidez, sentimenduen detekzioa, entitate erauzketa edo galdera-erantzunetarako. Horrez gain, ezaugarrietan oinarritutako hurbilpenak ahalbidetzen dituzte, esaldi edo hitz mailako errepresentazioak lortzearekin batera ataza desberdinetara erraz egokitu eta abantaila konputazionalak lortze aldera (Devlin *et al.* 2019).

Hizkuntza eredu eleanitzak aurrentrenatzen direnean, hizkuntza ezberdinetako datu kopuruak oso ezberdinak dira, ingelesezko corpusak hizkuntza gehienek baino askoz tamaina handiagoa duelarik. Egileak arazo hau arintzen saiatzen dira, baliabide gutxiago dituzten hizkuntzen adibideen gainlaginketa burutuz (Devlin *et al.* 2019; Conneau *et al.* 2020). Estrategia hau batzuetan ez da nahikoa, bereziki baliabide urriko hizkuntzentzat, baina baita baliabide *ertaineko* hizkuntzentzat ere. Horren aurrean, hizkuntza eredu eleanitzetan ematen den hizkuntzen azpi errepresentazioari aurre egiteko, hainbat hurbilpenek hizkuntza zehatzetarako bereziki entrenatutako eredu elebakarrak ere proposatu dituzte (Agerri *et al.* 2020; Armengol-Estapé *et al.* 2021).

2.1.4 Interakzioen errepresentazioak

Sare sozialak gure eguneroko bizitzaren zati garrantzitsu bihurtu direla aipatu dugu, elkarri eragiteko eta elkar komunikatzeko moduak aldatuz. Interakzio hauek sare sozialetako erabiltzaileen artean parte hartzeko, konektatzeko eta komunikatzeko prozesu dinamikoari erreferentzia egiten diete. Hainbat ekintza barnebildu ditzakete, hala nola, edukia partekatzea, iruzkinak egitea, atsegitea, jarraitzea, mezuak bidaltzea eta eztabaidetan parte hartzea.

Interakzio horiek loturak sustatu eta harremanak eraikitzearekin batera, ideiak, informazioa eta emozioak trukatzeko aukera ere ematen dute. Horregatik, interakzio hauek oso baliagarriak dira erabiltzaile mailako ikerketak burutzeko. Izan ere, interakzioak baliatuta, erabiltzaileen errepresentazioak sortzen dira hauen ezaugarriak aurreikusteko asmoz. Interakzioetan oinarritutako errepresentazio horiek ideologia politiko zein sentimendu analisiak burutzeko atazetan erabili dira bereziki (Conover *et al.* 2011a: b; Barberá and Rivero 2015; Barberá 2015; Magdy *et al.* 2016; Pennacchiotti and Popescu 2011b).

Interakzioekin erabiltzaileen errepresentazioak lortzeko erabilitako hurbilpen ezagunenen artean indarrak-zuzendutako (*force-directed*) algoritmoa (Fruchter-

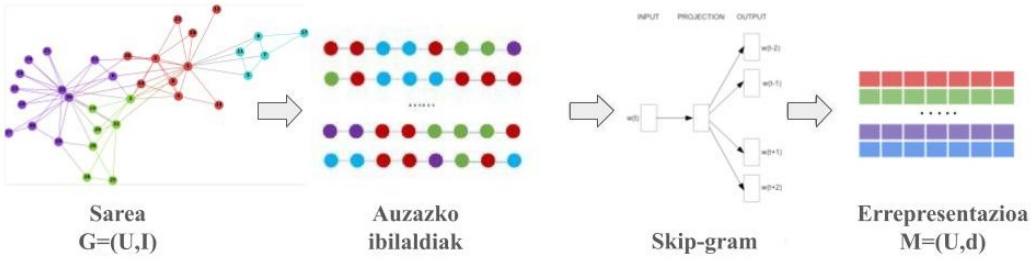
man and Reingold 1991) dugu. Indarrak-zuzendutako algoritmoa gainbegiratu gabeko erabiltzaileen errepresentazioak sortu eta aukeratutako sailkapen ataza burutzeko erabilia da Conover *et al.* 2011a; Darwish *et al.* 2020).

Algoritmo horretan oinarritutako ForceAtlas2 (Jacomy *et al.* 2014) inplementazioa oso erabilia da ere erabiltzaileen ezkutuko egiturak errepresentatzeko. Izan ere, sare sozialetako konpartitutako edukietan oinarrituta, gorroto gehien transmititzen duten diskurtsoen gaiak identifikatzeko (Garimella *et al.* 2018) zein konspirazio teorien inguruko zurrumurruen analisia egiteko (Bruns *et al.* 2020) baliatu da. Horrez gain, politika arloan, manipulazioaren karakterizazioan (Ferrara *et al.* 2020), erabiltzaileek elkarrizketa politikoetan jokatzen duten rolak identifikatzeko (Recuero *et al.* 2019) eta, baita, polarizazio zein elkarrizketa eza aztertzeko (Bruns and Highfield 2015) erabilia izan da.

ForceAtlas2 (Jacomy *et al.* 2014) hainbat nodo eta loturez osatutako sarea bi dimentsioko espazio batean bihurtzen duen algoritmo bat da, sareari gizakiontzat ulergarri edo interpretagarria den forma bat emanez. Algoritmo honek sare batean ageri diren nodo guztiak ezarritako erlazioen arabera ordenatzen ditu, indarrak-zuzendutako algoritmoari esker (Fruchterman and Reingold 1991). Hurbiltze eta urruntze prozesu baten ondorioz, erlaziorik gabeko nodoak elkar uxatzen dira, erlazionatutako nodoak elkar erakartzen diren bitartean. Metodo honek interakzio-matrize handiak bi dimentsioko errepresentazio bihurtzen ditu, datuen sakabanaketa eta memoria-erabilera nabarmen murrizten dutenak, informazio zati bat galtzen den arren.

Hurbilpen neuronalak ere erabiliak dira erabiltzaile mailako errepresentazioak ikasteko. Hurbilpen mota hauen artean ezagunenetariko eta arrakastatsuenen artean DeepWalk (Perozzi *et al.* 2014) eta node2vec (Grover and Leskovec 2016) ditugu (Jusup *et al.* 2022; Ma *et al.* 2021). Kasu hauetan, dimentsio baxuko nodo errepresentazioak sortzeko neurona sare artifizialetan oinarritutako geruza ezkutu bat baliatzen dute (ikusi 2.2. irudia). Alde batetik, DeepWalk hurbilpena mobilitatea eta erlazio sozialak (Yang *et al.* 2019), eragin sozialak (Qiu *et al.* 2018) eta hauekin lotutako albiste faltsuen detekzioa (Wu and Liu 2018) bezaklako atazetarako erabilia izan da. Bestalde, antzeko lan batzuk node2vec algoritmoan oinarritzen dira erabiltzaileen ezaugarritze bitartez tratu txarrak (Mishra *et al.* 2018), gorrotozko hizkera (Del Tredici *et al.* 2019) edo *troll* politikoen rolak (Atanasov *et al.* 2019) identifikatze aldera.

DeepWalk (Perozzi *et al.* 2014) algoritmoak erabiltzaileen errepresentazioak ikasten ditu sareko nodo konektatuen artean ausazko ibilaldi uniformeak simula-



2.2 Irudia – Sare errepresentazio metodo neuronalak. Interakzio (*I*) sare bat (*G*) osatzen duten erabiltzaileekin (*U*), dimentsio (*d*) zehatzetako errepresentazio dentsoak (*M*) sortzeko arkitektura.

tuz. Lehen pauso bezala, instantzia edo nodo bakoitzeko (u), luzeera jakin bateko (t) ibilaldi kopuru (γ) bat sortuko da sareko bestelako nodoetatik pasatzen dena, auzazko ibilaldi deituak. Ausazko ibilaldiko, nodo jakin batean hasi eta pauso bakoitzean berekin konektatuta dagoen nodo auzokide batera mugitzen den nodoen sekuentzia bat sortuko da. Ibilaldi bakoitzak hasierako instantziaren testuingurua ordezkatuko du, konektatuta dauden baino auzaz aukeratu diren nodoez osatuta egonik. Bigarren pausoan, sare neuronal bat ikasiko da instantzia baten errepresentazioa lortzeko. Ikasketarako, word2vec sistemako Skip-gram metodoa (Mikolov *et al.* 2013a) erabiltzen da, hitz-bektoreetan bezalaze, instantziatik bere testuingurua iragarriz. Ikasketa prozesu horretan, nodo baten probabilitatea aurreikusten da inguruko auzokideak kontuan hartuta. Ondorioz, testuinguru berdinetan azaltzen diren nodoek antzekotasunak edukiko dituzte, alta, antzeko errepresentazioak edukiko dituzte entrenatutako sarearen pisu matrizean. Ikasitako nodo errepresentazioak bestelako atazetarako baliagarriak izan daitezke, hala nola, nodoen sailkapena edota loturen iragarpenerako.

Node2vec (Grover and Leskovec 2016) algoritmoa DeepWalk antzekoa da, baina sarearen egitura zehazteko bi parametro gehitzen ditu ausazko ibilaldiak gertatzen diren bitartean hauek kontrolatzeko. Parametro horiek itzulera (*return:* p) eta sarrera-irteera (*in-out:* q) parametroak dira. Itzulera parametroak (p) ausazko ibilaldietan bisitatutako puntuetara itzultzeko probabilitatea kontrolatzen du, balio altuagoetan nodo bat berriro bisitatzeko probabilitatea gutxitzen da. Sarrera-irteera parametroak (q) sarearen zati urrunetara iristeko probabilitatea kontrolatzen du, balio altuagoekin urruneko nodoak barneratzeko aukera handiagoa suposatuz. Parametro hauei esker sarearen zein ezaugarri barneratu daitekeen kontrolatu daiteke. Alde batetik, homofilia edo komunitate bereko nodoak azpimarratuz,

eta bestetik, baliokidetasun estruktural edo sarean betetzen diren rol zehatzak nabarmenduz. Gerora, berriz ere skip-gram (Mikolov *et al.* 2013a) algoritmoa eta DeepWalken bestelako parametroak erabilita, instantzia edo nodo batetik abiatuta beren ingurukoak iragarri behar ditu sare neuronal bat baliatuz.

Azkenik, grafoetan oinarritutako neurona sareak (Graph Neural Networks, GNN) darabilten hurbilpenek, hala nola "*Graph Convolutional Networks*" (Kipf and Welling 2017) eta "*Graph Attention Networks*" (Velickovic *et al.* 2018), aldez aurretik lortutako ezaugarrien irudikapenak behar dituzte amaierako ereduak entrenatzeko. Hori dela eta, eredu neuronal hauek etiketatutako datuak eta ezaugarri osagarriak behar dituzte errepresentazioak ikasteko prozesuan. Ondorioz, eredu hauek ezin dira eraginkortasunez erabili gainbegiratu gabeko erabiltzaileen errepresentazioak eraikitzeko. Gainera, algoritmo hauek memoria errendimendu arazoak dituzte, auzokidetasun-matrizeak erabiltzaile askorekin sortzean beharrezkoa den memoria kantitatea kuadratikoki handitzen baita. Horregatik, hurbilpen hauek gure azterketatik kanpo geratzea erabaki dugu.

2.2 Erlazionatutako lana

Tesiaren helburu nagusia —orokortu daitekeen erabiltzaile mailako ezaugarritzea— ikerketa-lerro guztietan zehar agertu den arren, ataza ezberdinetan aplikatua izan da. Ondorioz, azpi-atal hau ataza zehatzen arabera antolatuko dugu, aurkeztutako hiru ikerketa-lerroetan lantzen diren atazen inguruko literatura aurkeztuz: (i) ezaugarri demografikoen identifikazioa (§2.2.1 - L1), (ii) jarreren detekzioa (§2.2.2 - L2) eta (iii) joera politikoaren identifikazioa (§2.2.3 - L3).

2.2.1 Ezaugarri demografikoen identifikazioa: adina

Sare sozialak gizarte ikerkuntza egiteko datu-iturri oparo bat direla ikusi da, datuen berehalakotasun, espontaneotasun eta kostu baxuagatik. Hala ere, adina, sexua edo maila sozioekonomikoa bezalako datu demografikoak ez daude sare hauetan beti eskuragarri. Datu demografiko hauek oinarrizko elementuak dira gizarte ikerkuntzan, populazioen osaerari, ezaugarriei eta beharrizanei buruzko informazio baliotsua eskaintzen baitute. Horregatik, beharrezkoa da egituratu gabeko datuetatik ezaugarri demografikoak nola iradoki daitezkeen ikertzea.

Gabezi hauei aurre egiteko, sare sozialetarako berariaz egokitutako Adimen Artifizialeko teknikak garatu dira sexua, adina edo kokapen geografikoa bezalako ezaugarri demografikoak iragartzeko (Cesare *et al.* 2017; Morgan-Lopez *et al.*

2017), adina eta generoa izanik ohikoenak (Cesare *et al.* 2017). Esaterako, adinarekin erlazionatutako informazioak ezagutza baliotsua eskaintzen du populazio baten egungo zein etorkizuneko egoerei buruz, populazio horrek duen ugalketa-ahalmen, biziraupen-tasa eta ekosisteman duen eginkizuna argitzen laguntzeko besteak beste. Adina edo bizitza-etapa automatikoki detektatzeko helburuarekin hainbat lan argitaratu dira, bereziki interesgarriak izanik Twitterreko erabiltzaileen inguruan burututakoak, sare sozial honek eskaintzen duen datuen ugaritasun eta eskuragarritasunagatik.

Adinaren identifikazio automatikoa burutzeko, ikasketa automatikoan oinarritutako hurbilpen gainbegiratuak baliatu dira, bereziki testuan oinarritutako datuak erabiliz (Rao *et al.* 2010; Al Zamal *et al.* 2012; Nguyen *et al.* 2013; Marquardt *et al.* 2014; Morgan-Lopez *et al.* 2017; Zaghouani and Charfi 2018). Adinaren identifikazio automatikoa egitera begira, erabiltzaile kopuru esanguratsu baten adina eskuz etiketatu dute. Erabiltzaile hauengandik lortutako datuekin, eskuz etiketatutako adina iradokitzeko sailkatzaile edo sistemak entrenatu dira, egoera eta hizkuntza ezberdinetara egokituta egonik. Sailkatzaile guzti horiek ikasketa automatikoan oinarritutako metodo gainbegiratuak dira, guztiak etiketatutako datu-multzo bat erabiltzen dutelarik entrenamendu eta ebaluaketarako.

Adinaren identifikazio automatikoa helburu duten lanek, datu-multzo propioak sortzen dituzte 300-3.000 erabiltzaile artean eskuz etiketatuz (2.1. taula) Batez ere testuan oinarritutako hurbilpenak dira denak, nederlandera, ingelesa edo gaztelania bezalako hizkuntzentzat garatutako sistemak izanik. Alde batetik, eskuz etiketatutako datu-multzo hauen tamaina ikerketaren arabera aldatuz doa, baina guztiek 300 eta 3000 erabiltzaile artean dauzkate etiketatuta. Bestalde, sailkapena burutzeko asmoarekin, adin tarte bitarrak (Rao *et al.* 2010; Al Zamal *et al.* 2012) edo hirutarrak (Nguyen *et al.* 2013; Morgan-Lopez *et al.* 2017) erabili dituzte batez ere.

Erreferentzia	Erabiltzaileak	Adin-tarteak	Hizkuntza
Rao *et al.* (2010)	1000	2	en
Al Zamal *et al.* (2012)	400	2	en
Marquardt *et al.* (2014)	306	5	en, es
Nguyen *et al.* (2013)	3110	3	nl
Morgan-Lopez *et al.* (2017)	3184	3	en

2.1 Taula – Twitterren adina detektatzeko erreferentziazko datu multzoak. Hizkuntzak: ingelesa (en), gaztelania (es) eta nederlandera (nl)

Adinaren identifikazio egoki bat ahalbidetzeko, datu iturri anitz eta erabiltzaileak errepresentatzeko era ezberdinak jorratu dira. Horrela, aitzindaria izan den ikerketetako batean, testua soilik baliatuta, hitzetan oianrritutako n-gramak zein bariazio lexikoetan oinarritutako ezaugarri soziolinguistikoak erabili dira adinaren identifikaziorako (Rao *et al.* 2010). Horrez gain, txioen testuetan azaldutako ezaugarri ezberdinak, hitz erabilienak zein n-gramak besteak beste, erabiltzaileen arteko erlazioekin konbinatuz ere burutu egin dute adinaren detekzio ataza (Al Zamal *et al.* 2012). Modu antzekoan, testu edukitik eratorritako n-gramak eta erabiltzaileen metadatuak erabiliak izan dira adina iragartzeko (Nguyen *et al.* 2013). Bestalde, sentimenduetan oinarritutako hitzen ezaugarriak ere erabiliak izan dira adinaren identifikaziorako, besteak beste, LWIC (Pennebaker *et al.* 2001) teknika baliatuta hitzetatik aurrez zehaztutako emozioak erauzi eta adina iradokiz (Morgan-Lopez *et al.* 2017). Bide antzekotik, LWIC zein bestelako sentimendu ezaugarriak eta emotikonoak ere erabili dituzte erabiltzaileen adin tartea hainbat aukeren artean identifikatzeko (Marquardt *et al.* 2014).

Aurretik aipatutako lanetan sailkatzaile gisa erregresio logistikoa (*Logistic Regression*) eta euskarri bektoredun makinak (*Support Vector Machine*) erabiltzen direla ikusi daiteke 2.2. taulan. Errendimendu onena erakutsi duten sistemak, adina bizitza-etapa bitar (Rao *et al.* 2010; Al Zamal *et al.* 2012) edo hirutar (Nguyen *et al.* 2013; Morgan-Lopez *et al.* 2017) gisa modelatzen dutenak dira. Emaitza baxuena 5 klase edo adin-tarte aurreikusi behar dituen sailkatzaileak lortu ditu, horretarako ezaugarri psikologikoak baliatuta (Marquardt *et al.* 2014). Arlo honetako emaitzarik onena % 86-ko asmatze tasaraino ailegatzen da (Nguyen *et al.* 2013), kasu honetan iradoki beharreko informazioa 3 klasetan banatu egiten da eta etiketatutako datu-multzoa handienetakoa da.

Erreferentzia	Sailkatzailea	Klaseak	Asmatze tasa
Rao *et al.* (2010)	SVM	2	% 74,11
Al Zamal *et al.* (2012)	SVM	2	% 80,50
Marquardt *et al.* (2014)	SVM	5	% 48,31
Nguyen *et al.* (2013)	LogReg.	3	% 86,32
Morgan-Lopez *et al.* (2017)	LogReg.	3	% 74,00

2.2 Taula – Twitterren adina detektatzeko erreferentziazko sistemak. Sailkatzaileak: *Support Vector Machine* (SVM) eta *Logistic Regression* (LogReg.)

Nabarmendu beharreko ondorioen artean, sarrera datuen nolakotasunak sistema berearen efizientzia baldintzatzen duela ikusi da, sailkatzailea, iradoki beha-

rreko klase kopurua edo datu-multzoaren tamaina baino erabakiorragoa dela erakutsiz. Horrela, testua eta erabiltzaileen jarraitzaile zein informazioa datu sarrera bezala erabiltzen dituzten hurbilpenek emaitza altuenak lortzen dituztela (Nguyen *et al.* 2013; Al Zamal *et al.* 2012). Bestalde, ezaugarri psikologikoak erabili dituzten hurbilpenek emaitza baxuenak lortu dituztela ikusten da (Morgan-Lopez *et al.* 2017; Marquardt *et al.* 2014). Hala ere, aztertutako hurbilpenen arteko konparaketa zuzena ezin denez egin, ezberdintasunean baino, antzeko puntuetan fokoa jartzea beharrezkotzat jo da.

Izan ere, hurbilpenak ezberdinak izan arren, guztiek testuan oinarritutako antzeko ezaugarritzeak baliatu dituztela ikusi da, hala nola, hitzen n-grama (Rao *et al.* 2010; Al Zamal *et al.* 2012; Nguyen *et al.* 2013) eta ezaugarri soziolinguistikoetan (Morgan-Lopez *et al.* 2017; Marquardt *et al.* 2014; Rao *et al.* 2010; Al Zamal *et al.* 2012) oinarritutakoak. Ondorio orokor bezala, adinaren detekziorako ezaugarririk garrantzitsuenak idazkeran oinarritzen direla ikusi da (Rao *et al.* 2010; Al Zamal *et al.* 2012; Nguyen *et al.* 2013; Morgan-Lopez *et al.* 2017). Horrela, gazteek hitz kolokial gehigo (Nguyen *et al.* 2013) erabiltzeaz gain, luzaketa alfabetiko gehiago (zorionak / *zorionaaaak*), hitz larrien erabilera hedatuagoa (Zer? / *ZEEER??*), laburdurak (Zer moduz? / *zmz*) eta argot hitz (bikain / *txatxi* ; zortea / *folla*) gehiago erabiltzen dituzte (Nguyen *et al.* 2014; Morgan-Lopez *et al.* 2017). Honek esan nahi du gazteen idazkera helduena baino gehiago aldendu egiten dela estilo estandarretik (Rao *et al.* 2010; Al Zamal *et al.* 2012; Nguyen *et al.* 2013; Morgan-Lopez *et al.* 2017).

2.2.2 Jarreren detekzioa

Jarreren detekzioa adierazpen batek gai jakin bati buruz agertzen duen ikuspuntua edo jarrera identifikatzean datza (ikusi 2.3. taula). Sare sozialek duten aukera komunikatiboen baitan erabiltzaileek beren iritziak publikatu eta partekatzen dituzte, jarrerak ikertzeko baliabide baliotsu bat sortuz (Mohammad *et al.* 2016; Hardalov *et al.* 2021). Horrek esan nahi du, jarreren detekzioari buruzko ikerketak interesgarriak direla iritzi publikoa hobeto ulertzeko, adibidez, txerto, klima-aldaketa edo migrazioa bezalako gaiei buruzko jarrerak ezagutzeko. Horrez gain, jarrera detekzioa tarteko zeregin garrantzitsutzat jotzen da gertakarien egiaztatze (Augenstein 2021) edo albiste faltsuak detektatzeko (Shu *et al.* 2017) sistemetan.

SemEval 2016 (Mohammad *et al.* 2016) ataza ezagunean jarreren detekzioa burutzeko Twitter sare sozialetako datuak baliatzea proposatzen dute, hainbat gai aztertuz. Ataza era honetan formulatzen dute, txio batetako testua eta gai bat zehaztuta, sistema automatikoak jarrera *aldekoa*, *neutrala* edo *kontrakoa* den au-

Etiketa	Edukia
alde	*Gogoko dut egunero bizikletan ibiltzea* *Txirrindulariak osasuntsuagoak eta aktiboagoak dira*
neutral	*Frantziako Tourra uztailean da* *Bizikleta saldu nahi dut*
kontra	*Errepidean gehien gorrotatzen dudana txirrindulariak dira* *Bizikletaz doazenak arroputz batzuk dira*

2.3 Taula – Jarreren detekzioa azaltzeko adibideak. Kasu honetan **txirrindularitza** gaia kontutan hartuta. Gai zehatz honekiko *aldeko*, *neutrala* edo *kontrako* jarrerak zeintzuk diren ikusi daitezke.

rreikusi beharra dauka. Jarreraren identifikazioa gaiarekiko menpekoa izango da, hau da, aurrez zehaztutako gaiarekiko jarrera adieraziko du soilik, polaritatea edo sentimenduak zeintzuk diren albo batera utziz. Era honetan, txio baten autoreak gai zehatz batekiko daukan jarrera inferitu ahalko da. Horretarako, 4.870 txio 5 gai ezberdinetarako (ateismoa, aldaketa klimatikoa, feminismoa, Hillary Clinton eta abortuaren legalizazioa) anotatzen dira bakoitzaren testuan oinarrituta etiketa aukeratuz. Honi esker, gai ezberdinetan oinarritutako jarreren detekzioa ahalbidetzen den arren, hizkuntza bakarrean zentratutako hurbilpena da, hots, ingelesean zentratutakoa.

Sare sozialetan aurkitu daitezkeen hizkuntza anitzak baliatuta, badira ere eleaniztasuna barneratzen duten beste ikerketa batzuk. Hizkuntza ezberdinak baliatzen dituzten hurbilpenen artean Katalana zein Gaztelera (Taulé *et al.* 2018) eta baita Frantsesa eta Italiera (Lai *et al.* 2020a) barnebiltzen dituztenak daude. Hala ere, aipatutako lan hauek gai ezberdinen inguruko jarreren detekzioa egiteaz gain, datu-multzo murritz eta orekatu gabeak dira. Hutsune horien aurrean, Zotova *et al.* (2021) lanean Catalonia Independence Corpus (CIC) proposamena luzatu zuten, hizkuntza-arteko jarreren detekzioa gai berdinaren inguruan aztertzeko asmoekin, kasu honetan Kataluniako erreferendumaren[1] inguruko jarrerak. Hori egiteko, erabiltzaile mailako datu jasoketa eta anotazioa metodo erdi-automatikoa proposatzen dute, txio mailako etiketak erabiltzailetara proiektatuta datu kopurua

[1]Kataluniako erreferenduma, Kataluniaren independentziarako 2017ko erreferenduma izan zen. Herritarrak honako galdera honi baiezkoa ala ezezkoa ematera deituta zeuden: «*Nahi duzu Katalunia errepublika erako estatu independentea izatea?*». Kataluniako Gobernuak antolatu zuen baina Espainako Gobernuaren debekuarekin. Informazio osagarrirako: `https://en.wikipedia.org/wiki/2017_Catalan_independence_referendum`

handituz. Gainera, datu bilketa gai berdinekoa eta momentu berdinekoa izanda bi hizkuntzetarako, hizkuntza arteko konparaketa ahalbidetu daiteke.

Orokorrean, jarreren detekzio ataza helburu duten hurbilpen gehienak testu soilean zentratzen dira. Gainera, hurbilpen gehienek Twitter erabiltzen dute datu-multzoen sorrerarako, testua txioetatik erauziz. Horrela, ataza hau zentralizatzea oinarri izanda, hainbat gai ezberdinetan eta 11 hizkuntza ezberdinetako datu-multzoak jaso eta batu dira (Küçük and Can 2020). Bestalde, domeinu eta hizkuntza arteko jarreren detekzioak esperimentazioa eskaintzen du 15 hizkuntza ezberdinentzat 16 datu-multzo ezberdin eskainiz (Hardalov *et al.* 2021: 2022). Hala ere, proposamen gehienek datu-multzoak sortzeko Twitter erabiltzen duten arren, testuan oinarritutako hurbilpenak dira, sare sozial honetan dauden bestelako datu motak alboratuz.

Testuaz gain, sare sozialetatik erauzi daitezkeen bestelako datuak baliagarriak dira ere erabiltzaile mailako jarrerak edo joerak adierazteko. Era honetan, gainbegiratu gabeko jarreren detekzioa burutzeko *force-directed* (Fruchterman and Reingold 1991) eta *UMAP* (McInnes *et al.* 2018) bezalako hurbilpenak baliatzen dira. Algoritmo hauek, interakzioekin sortu daitezkeen albokotasun matrize erraldoiak dimentsio baxuko aldagai bihurtzen dituzte, datuen sakabanaketa murriztuta errepresentazio dentso eta aberatsak sortuz. Darwish *et al.* (2020) proposamenean, aipatutako algoritmoan oinarritzen dira Twitterreko erabiltzaileen jarrera etiketa bidezko propagazioaren bidez egiteko, klusterren arabera erabiltzaileak ezaugarrituz. Aipatutako metodologia erabilia izan da interakzioetan oinarritutako ezaugarriak erauzita erabiltzaileek zein gairi buruz aritzen diren identifikatu eta beren jarrera identifikatzeko (Stefanov *et al.* 2020) eta baita Turkiako polarizazio politikoa automatikoki identifikatzeko ere (Rashed *et al.* 2021). Aipatutako hurbilpen hauek, interakzioetan zentratzen dira soilik, hau da, erabiltzaile zehatzek txio, hashtag edo beste erabiltzaile batekin edukitako interakzioetan. Horrez gain, lan horietarako datu-multzo propioak sortu eta erabili dituzten arren, ez dituzte publikoki partekatzen, erreplikazio eta esperimentazioa zailduz.

Abagune honetan, testu mailan etiketatutako datu-multzoak eta interakzio datuak batzen dituen hurbilpen bakarra aurkitu da, SardiStance deitua (Cignarella *et al.* 2020). Hurbilpen honetan, italierazko 3.242 txio baliatuta, sardinen mugimenduarekiko[2] jarrera aldekoa, neutrala edo kontrakoa identifikatzea da asmoa

[2]Sardinen mugimendua (edo sardinak Salviniren aurka) Italian 2019an sortutako mugimendu bat izan zen. Mugimendu honek protesta baketsuak antolatzen zituen Italiako eskuin muturreko buru zen Mateo Salviniren kontra. Sardina izena ekitaldi jendetsuak antolatzeko ideiatik sortu zen, parte hartzaileak sardinak lata batean bezala plazetan bilduz. Informazio osagarrirako: `https://en.wikipedia.org/wiki/Sardines_movement`

(ikusi 2.4. taula). Datu-multzo honek baina, txioen testua eta dagozkien jarrera etiketa edukitzeaz gain, txioaren eta bere egilearen hainbat metadatu ere eskuragarri ditu. Era honetan, txioak berak jasotako *birtxio*, *atsegin* eta *erantzun* ekintzak zein txioen autoreen *jarraitzaile*ak zeintzuk diren barnebiltzen dira datu multzoan. Testua eta bestelako metadatuak erabilita iritzia oinarri duten atazetan hobekuntza lortu daitekeela argi utzi da literaturako hainbat lanetan (Rajadesingan and Liu 2014; Magdy *et al.* 2016).

Etiketa	**Edukia**
alde	#Sardine Grazie Le piazze piene La gente intorno a me Grazie *#Sardinak Eskerrik asko Plaza betea Jendea nire inguruan Eskerrik asko*
neutral	Dai provate a contarvi #6000sardine *Tira, saiatu #6000sardina zenbatzen*
kontra	Penose le #sardine oggi a #Milano! Strumento di una #bieca politica *Gaur #Milanen #sardinak penagarriak izan dira! #politika ilun baten tresna*

2.4 Taula – SardiStance (Cignarella *et al.* 2020) datu-multzoaren adibidea. **Sardinen mugimendua**rekiko *alde*, *neutral* edo *kontra* jarrerak zeintzuk diren ikusi daitezke.

SardiStance datu-multzoa erabilita testua eta bestelako metadatuak baliatzen dituzten hurbilpenek, testu soila erabiltzen dutenak baino emaitza aski hobeak lortzen dituztela argi erakutsi da (Cignarella *et al.* 2020). Gainera, errendimendu onena lortzen duten sistemen artean, Transformerretan oinarritutako hizkuntza eredu aurre-entrenatuak erabiltzen dituzte. Honela, datu-multzo honetan sailkapen emaitza onenak dituen sistemak Transformer arkitekturan oinarritutako hainbat testu-sailkatzaileen emaitzak konbinatzen ditu erabiltzaileen *jarraitzaile*etatik eratorritako distantzia ezaugarriekin (Espinosa *et al.* 2020). Beste hurbilpen batean testu datuak (emotikonoak, karaktere bereziak zein hitz errepresentazioak) konbinatzen dituzte Multidimensional Scaling (MDS) baliatuta elkarrekintzetatik eratorritako ezaugarriekin (Ferraccioli *et al.* 2020). Erabiltzaileen errepresentazioak sortzeko node2vec eta deepwalk erabiliak dira ere, gerora testuan oinarritutako tfidf edota hitz-bektore errepresentazioekin konbinatzen dutelarik (Alkhalifa and Zubiaga 2020). Interakzio eta testuan oinarritutako sistema hauek, baina, hizkuntza eta egoera konkretu baterako daude prestatuta. Hortaz, bestelako hizkuntza eta egoeretan aplikatu daitezkeen jakiteko datu-multzo gehiago beharko genituzke.

Honen aurrean, testu eta interakzioetan oinarritutako jarreren detekzioa hizkuntza arteko egoeretan aztertu ahal izateko VaxxStance datu-multzoa proposatu genuen (Agerri *et al.* 2021). Datu-multzo honetan, sare sozialetan txertoen jarrerak aldekoak, neutralak edo kontrakoak diren identifikatzeko ataza proposatzen da (ikusi 2.5. taula). Horretarako, eta aurretik aipatutako hurbilpenek egin bezala, gai horrekin erlazionatutako txioak jaso dira euskara zein gaztelaniaz. Jasotako txioen testuaz gain, egileen *jarraitzaile* eta *birtxio* datuak gehitu dira. Era honetan, gai berdinaren inguruko jarreren detekzioa burutu ahalko da testu zein interakzioak baliatuta, baina hizkuntza arteko eta hizkuntza ezberdinetako analisia burutuz aldi berean.

Etiketa	**Edukia**
alde	#Koronabirusa geldiaraziko duen txertoaren zain gaude. #covid19 Covid-19aren txertoa, lorpen zientifiko handiena - Zientzia Kaiera
neutral	Zein fasetan ote doa txertoa? #DeskubritzenDirenPegatinak Bihar ipiniko diete txertoaren bigarren dosia egoitzakoei.
kontra	Ez det txerto toxiko hau jarriko ezta erotuta ere! nik eut txerto hoi jarriko, negazionista naiz?

2.5 Taula – VaxxStance (Agerri *et al.* 2021) datu-multzoaren adibidea. **Txerto**en gaiarekiko *alde*, *neutral* edo *kontra* jarrerak zeintzuk diren ikusi daitezke.

VaxxStance atazan emaitza onenak lortzen dituen hurbilpenak ezaugarrien ingeniaritza burutzen du, horretarako estiloan oinarritutako txio zein erabiltzaile mailako ezaugarriak, lexikoiak, mendekotasun sintaktikoen analisia eta baita interakzioetan oinarritutako hainbat ezaugarri sortuz eta konbinatuz (Lai *et al.* 2021). Hurbilpen horrek, hizkuntza bakoitzerako ezaugarri zehatz batzuk eskuz aukeratu eta doitu egiten ditu, bakoitzarentzat metodo ezberdin bat proposatuz eta hizkuntza artean konfluentzia gutxi edukiz. Horren aurrean, hizkuntzatik independentea den hurbilpen bat proposatzen dugu, erabiltzaileen interakzioak baliatzen dituena jarreren detekzioa hainbat testuinguru ezberdinetan aplikatuz (Fernandez de Landa and Agerri 2022). Hurbilpen honek, hizkuntza eta testuinguru desberdinetara egokitzeko gaitasuna edukitzeaz gain, testu mailako errepresentazioekin konbinatu daiteke errendimendua hobetze aldera.

2.2.3 Joera politikoaren identifikazioa

Aurreko atalean ikusi den moduan, pertsona askok sareko bitartekoak erabiltzen dituzte askotariko gaiei buruzko iritziak emateko. Gehien eztabaidatzen diren gaien artean politikarekin erlazionatutako gaiak daudela ezin da ukatu. Blogetan, foroetan eta albiste-guneetan modu irekian eztabaidatzen diren gaien eta erabiltzaileen joera politikoa ezagutzea interesekoa da, gai hauek iritzi edota politika publikoetan izan dezaketen eragina aurreikusi eta ulertzeko. Honela joera politikoaren identifikazioa erabili da hautetsien joera politikoa aurreikusteko AEBetako alderdi biko (Akoglu 2014) edo Brasil bezalako alderdi anitzeko (Vaz de Melo 2015) sistema politikoetan, parlamentutako bozak erabiliz. Hautetsien joera politikoa euren boto publikoetan oinarrituta ondoriozta daitekeen arren, azterketa horien hedapena parlamentu zehatzetara mugatzen da eta ezin da beste populazio batzuetara hedatu sare sozialetako datuekin egin daitekeen bezala.

Joera politikoa aurreikusteko sare sozialetako datuak baliatzen dituzten hurbilpen ezagunenen artean interakzioak baliatzen dituzten hurbilpenak azaltzen dira. Metodologia horiek erabiltzaileen arteko harremanetan oinarritzen dira, batez ere erabiltzaileak jarraitu (*follow*) edota edukia konpartitzea (*retweet* edo *birtxio*) bezalako interakzio ekintzetan. Besteak beste, Twitterren baitako joera politikoaren inguruko ikerketa aitzindarietako batean, ezker-eskubi lerrokatze politikoa aztertzen da AEBeko testuinguruan, birtxioak elkarrekintza polarizatuenak direla erakutsiz (Conover *et al.* 2011b). Erabiltzaileak jarraitu (Barberá and Rivero 2015; Barberá 2015) eta baita jarraipen zein edukia konpartitzeko ekintzak ere (Garimella and Weber 2017) baliatu dira erabiltzaileen ezker-eskubi joera politikoa identifikatzeko. Era berean, birtxioak Twitter erabiltzaileen inklinazio kontserbadore eta liberalak aurreikusteko ere erabili izan dira, hauekin inklinazio horiek kuantifikatu (Wong *et al.* 2013) eta polarizazio azterketak egitez (Barberá *et al.* 2015). Birtxioak ere, erabiliak izan dira hainbat gaien inguruan aldeko edo aurkako joerak aztertzeko (Darwish *et al.* 2020).

Joera politikoa ere sare sozialetako testuetan oinarrituta aztertua izan da. Honela, testuetatik eratorritako ezaugarri linguistikoak erabiliak izan dira joera politikoa aztertzeko kontserbakor-liberal ardatzean (Preotiuc-Pietro *et al.* 2017). Sentimenduen analisia ere erabilia izan da ezker-eskubi joera politikoa aztertzeko (Plà and Hurtado 2014). Horrez gain, testuak erabiltzen dituzten metodoak erabili dira erabiltzaileek alderdi demokrata edo errepublikarrarekiko duten lerrokatzea aztertzeko, besteeak beste, topiko-bektoretan oinarritutako hurbilpena (Kulshrestha *et al.* 2017) edo n-grametan entrenatutako autokodetzaileak (Yan *et al.* 2019). Testuan oinarritutako hitz-bektoreak Turkiako polarizazio politikoa aztertzeko ere

erabiliak izan dira (Rashed *et al.* 2021). Bestalde, PoliticEs 2022 atazan (García-Díaz *et al.* 2022), testuetatik ezker-eskubi joera politikoen identifikazioa aztertu da erabiltzaileen ezaugarritzean oinarriturik, Transformers-ak (Vaswani *et al.* 2017) oinarri dituzten metodoek emaitzarik onenak lortuz. Erabiltzaileen ezaugarritzea helburu duen antzeko lan batean, alderdi politikoetan oinarritutako hurbilpena erabiltzen da Italiako joera politikoak ezagutzeko (Fagni and Cresci 2022), emaitzarik onenak word2vec metodoarekin (Mikolov *et al.* 2013a) lortuz.

Bestalde, hainbat ikerketek ikuspegi konbinatua erabili dute, testu eta interakzioetan oinarritutako datuak baliatuz. Ikerketa aitzindari batek testuan eta interakzioetan oinarritutako datuak konbinatu eta alderatu zituen Twitterreko erabiltzaileen ezker-eskuinaren joera jasotzeko, birtxioak ezaugarri hau identifikatzeko datu-mota adierazgarriena dela erakutsiz (Conover *et al.* 2011a). Beste ikerketa berritzaile batek testua eta interakzioak erabili zituen, hala nola lagunak edo jarraitzaileak, erabiltzaileak alderdi demokrata edo errepublikanoarekin lerrokatzen diren ondorioztatzeko (Pennacchiotti and Popescu 2011b). Alderdi berdinen lerrokatzea aztertzeko, jarraitzaile, birtxio eta hashtagetako testua erabili da beste ikerketa batean (Hua *et al.* 2020). AEBetako esparruan jarraituz, beste ikerketa batek erabiltzaileen joera liberal edo kontserbadorea aztertu du, emaitza onenak lortuz testu eduki eta birtxio zein jarraitzaileetatik eratorritako sare baten konbinaketarekin (Lahoti *et al.* 2017). Azkenik, interakzio eta testuetan oinarritutako ezaugarriak erabili dira Kataluniako, Euskal Herriko eta Eskoziako independentziaren aldeko edo aurkako joerak identifikatzeko (Zubiaga *et al.* 2019).

Laburbilduz, joera politikoei buruzko ikerketa gehienak bi alderdi edo joera nagusien arteko sailkapen bitarrera mugatu dira (Conover *et al.* 2011b: a; Barberá and Rivero 2015; Barberá 2015; Garimella and Weber 2017; Barberá *et al.* 2015; Pennacchiotti and Popescu 2011b; Hua *et al.* 2020; Xiao *et al.* 2020) eta sailkapen anitzago bat proposatu duten hurbilpenak agertoki edo eskualde bakar batera mugatzen dira Boutet *et al.* 2012; Makazhanov and Rafiei 2013; Rashed *et al.* 2021). Horrek, ordea, metodo horien aplikagarritasuna eta ikerketa horietatik erauzitako ezagutza mugatzen ditu. Izan ere, gizarte-testuinguru bakoitzak bere errealitate politikoa du, bi aukera ideologiko baino gehiagotan islatzen dena (Lisi 2018) eta parte-hartze politikoan inplikazio maila desberdinak dituena (Almond and Verba 2015).

3. KAPITULUA

Ezaugarri demografikoen identifikazio automatikoa

Sare sozialek mugarik gabeko komunikazioa ahalbidetzeaz gain, ikerketa sozial eta linguistikorako tresna ere badira, datu kopuru handiak lortzeko iturri garrantzitsu bat direlako. Honek aukera ematen du talde zehatzen inguruko hausnarketa burutzeko, erabiltzaileek modu publikoan sortzen dituzten datuak baliatuta. Horrela, lan honen asmoa erabiltzaile euskaldun gazteen errealitatea aztertzea izango da, Adimen Artifiziala ikerketa sozialean aplikatuz. Euskara teknologia berrietara nola moldatzen den jakiteko eta gazteen euskararekiko atxikimendua zein den ezagutzeko asmo bikoitzarekin, sare sozialetan euskararen inguruko ikerketa burutu da, zehazki Twitterren. Horretarako, 8.000 erabiltzaileren euskarazko 6 milioi publikazio lortu eta publiko egin dira. Gerora, erabiltzaile bakoitza bizitza-etaparen (gazte edo heldu) arabera sailkatuko dugu, horretarako erabiltzaileek publikatutako testuak baliatuta. Horretarako, bi datu-multzo berri etiketatu direlarik, bata idazkera estiloan oinarrituta (*heldugazte* 1.000 publikaziorekin) eta bestea erabiltzaileen adinean oinarritua (*heldugazte-age* 80.000 publikazioarekin). Azkenik, erabiltzaileek partekatzen duten edukia kontutan hartuta, hauen arteko harremanak zeintzuk diren azaleratu eta azpi-komunitateak nola eratzen diren aztertu da. Aipatutako ezaugarri demografiko eta sozialak automatikoki iradokitzeko sistema adimentsuak garatu eta aplikatu dira, ikasketa automatikoan oinarritutako Hizkuntzaren Prozesamendua eta ezaugarrien errepresentazioa erabiliz. Guzti honekin, gizarte ikerkuntza Adimen Artifizialeko teknika bitartez burutzea posible dela erakutsi da, egituratu gabeko informazioa kudeatuta interpretagarria den ezagutza sortuz eta ikerketa egiteko modu berriei bide emanez.

3.1 Motibazioa eta Ekarpenak

Euskara baliabide urriko hizkuntza da, Euskal Herriko biztanleen % 28,4ak hitz egiten du eta % 44,8k ulertzen du (Eusko Jaurlaritza *et al.* 2016). Ofizialtasunari esker EAEko administrazio publikoan, hezkuntza sisteman eta zenbait hedabidetan presentzia duen hizkuntza da. Horrela, EiTBn (Euskal Irrati Telebista), euskal irrati-telebista publikoan, eduki guztiak euskara hutsean emititzen diren irrati eta telebista kateak aurki daitezke. Gainera, badaude beste hedabide independente batzuk, hala nola Berria (egunkaria), Argia (astekaria) eta HamaikaTB (telebista katea), zeinetan gaurkotasunari buruz euskara hutsean aritzen den. Hala ere, ohiko telebista eta albistegietan euskararen presentzia nahiko txikia izaten jarraitzen du, batez ere gaztelaniaz daudenekin alderatuta.

Testuinguru horretan, geroz eta gehiago erabiltzen diren Twitter bezalako sare sozialek garrantzi berezia dute euskara bezalako baliabide urriko hizkuntza batentzat, hiztun komunitateak aurkitzea erraza delako hizkuntza gutxituentzat ere (Jones *et al.* 2013; Mhichíl *et al.* 2018; McMonagle *et al.* 2019). Horrela, Twitterren euskaldunen komunitate sendo eta aktibo bat aurki daiteke, ikerketarako baliagarria izan daitekeen idatzitako testu eduki ugaria sortzen duena. Bestalde, gazteen kolektiboa bereziki aktiboa da sare sozialetan, bestelako esparruetan ez daukaten protagonismoa bertan bereganatzen baitute (Fernandez de Landa 2017). Gainera, sareetako erabiltzaileek edukia sortu eta partekatzen dute, beraiei buruzko informazioa modu inplizitu zein esplizituan konpartituz. Erabiltzaileek modu espontaneoan sortutako datu kantitate erraldoi horiek baliagarriak dira gizarte ikerketak egiteko, tradizionalki erabiltzen diren metodoen osagarri izanik (Baldwin *et al.* 2015; Nguyen *et al.* 2016; Rosenthal *et al.* 2017). Modu desegituratuan sortzen den informazio guzti hori ezagutza bihurtu daitekeela ikusi da Adimen Artifiziala baliatuta (Lazer *et al.* 2009; Edelmann *et al.* 2020). Horren haritik,sare sozialetatik eratorritako testu eta interakzio datuak baliatu nahi dira, ikasketa automatiko eta sakona erabilita, datu kopuru erraldoietan oinarritutako eskala handiko ikerketa soziala burutzeko. Esaterako, hainbat lanek Twitter erabili dute zurrumurruen hedapena (Derczynski *et al.* 2017), jarrera politikoaren detekzioa (Mohammad *et al.* 2016) edo gorroto hizkera antzematea (Basile *et al.* 2019) aztertzeko.

Geure kasuan, helburu orokor moduan, euskal erabiltzaile gazteek euskarazko edukia nola partekatzen duten aztertzea izango da. Era honetan, gazteen errealitate ezezagunera hurbilpen bat lortzeaz gain, XXI. mendeko erronketara euskara nola egokitzen ari den ezagutzeko aukera dugu. Horretarako, lehenengo azpihelburua erabiltzaile euskaldunak identifikatzea izango da eta beren euskarazko

publikazio pertsonal (txio) zein partekatutakoak (birtxioak) jasotzea. Bigarren azpi-helburua, lortutako erabiltzaile euskaldunetan gazteak eta helduak identifikatzea izango da, idazkera estilotik abiatuta, ezaugarri demografikoak automatikoki iradokitzeko sistema adimentsuak garatuz. Horretarako bizitza-etapa identifikatzeko gai diren sailkatzaile ezberdinak garatu dira, erabiltzaileen testua oinarri hartuta. Hirugarren eta azken azpi-helburua gazteen komunitateak zeintzuk diren identifikatu eta aztertzea da, euskal erabiltzaileek edukiak partekatzeko ekintzak baliatuta. Bestalde, lan honen bigarren mailako helburu bezala, gizarte ikerkuntzarako Adimen Artifiziala nola aplikatu daitekeen aztertu da, datu-bilketatik hasi eta emaitzen analisiraino pausu guztiak jorratuz.

Atal honen helburu nagusia Adimen Artifiziala gizarte ikerkuntzan nola aplikatu daitekeen ikertzea izan da, hurbilpen berriak proposatuz. Horrez gain, kapitulu honek ekarpen zehatzak egiten ditu datu bilketa zein etiketatzean eta baita edukien egituratze edo sailkapenean ere. Hortaz, atal honen ekarpen nagusiak bost hauek dira: (1) Lehenik eta behin, euskaraz idatzitako 6M publikazioz osatutako *Heldugazte-osoa*[1] corpus erraldoia jaso eta eskuragarri jarri da baliabide urriko hizkuntza honentzat, euskararen ikerketa ahalbidetzeko on-line inguruneetan. (2) Soziolinguistikan oinarrituta, euskarazko testu-sekuentziak erregistro formal edo informalekoak diren identifikatzeko *Heldugazte*[2] datu-multzoaren sorrera. (3) Gazte eta helduen edukiaren identifikaziorako *Heldugazte-age* [3] datu-multzoaren sorrera, gazte/heldu mailan erdi-automatikoki etiketatutako 80K publikazio dituena. (4) Euskarazko testu-sekuentzien sailkapenerako metodo ezberdinak garatu eta frogatu dira, besteak beste Transformerretan oinarritutako hizkuntza-eredu elebakar zein eleanitzen (Devlin *et al.* 2019; Agerri *et al.* 2020) aplikazioak aztertuz. Gainera, garatutako metodoak corpus erraldoian aplikatu dira, egoera errealen aurrean haien portaera kualitatiboki aztertu eta baloratuz. (5) Gainbegiratu gabeko ikasketa teknika neuronalak aplikatu dira komunitate detekziorako, konpartitutako edukiek adierazten dituzten erlazioak oinarri bezala hartuta erabiltzaileen preferentziak erauziz.

Atal honetan aurkezten diren datu-multzo zein testuak sailkatzeko metodologiak baliagarriak dira hizkuntzaren prozesamenduko beste hainbat atazetarako, eta aldi berean, ikerketa sozial eta demografiko berritzaileak burutzeko ere.

[1] *Heldugazte-osoa* corpusa:
`http://ixa2.si.ehu.es/heldugazte-corpus/heldugazte.osoa.tar.gz`

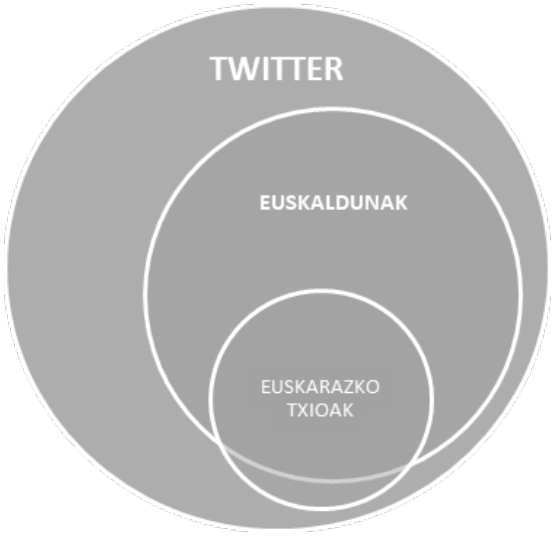
[2] *Heldugazte* datu-multzoa (informal-formal):
`https://github.com/ixa-ehu/heldugazte-corpus`

[3] *Heldugazte-age* datu-multzoa (gazte-heldu):
`https://github.com/joseba-fdl/heldugazte-age-corpus`

3.2 Euskal komunitatearen identifikazioa

Sare sozialen erabilera zabalduak (Eustat 2022; Eurostat 2023b), bereziki gazte kolektiboaren baitan (Eurostat, 2023a; Fernandez de Landa, 2017), aukera ezin hobea luzatzen du bertatik ikerketarako informazio baliotsua lortzeko. Zehazki Twitterrek datu-multzo handiak eskuratzea ahalbidetzen du, baita baliabide gutxiago dituzten hizkuntzetarako ere (Jones *et al.* 2013; Mhichíl *et al.* 2018; McMonagle *et al.* 2019). Era horretan, beharrezkoak diren datuak biltzeko gure azterketaren gaiarekin bat datorren Twitterreko erabiltzaileen komunitatea definitu behar da, ikerketaren unibertsoa izango dena.

Komunitate zehatz baten azterketa egitera bidean beharrezkoa da komunitatea zehaztea, lehenik eta behin aztertu beharreko unibertsoa mugatuz. Hau da, ikertu behar diren subjektuak identifikatzea izan behar da eman beharreko lehen pausua, kasu honetan unibertsoa euskaraz egiten duten erabiltzaileak izanik. Twitterreko sare erraldoian euskarazko erabiltzaileak topatzeko, euskaraz txio kopuru minimo bat argitaratzen duten erabiltzaileen identifikazioa egin beharko da, euskal komunitatea komunitate eleanitza baita. Horrela, euskarazko txioak argitaratu dituzten subjektuetatik euskal erabiltzaileak hautemateko, euskarazko txio kopuru minimo bat duten erabiltzaileak hautatu beharko dira, 3.1. irudian ikus daitekeen moduan.



3.1 Irudia – Unibertsoaren identifikazioa.

Euskaraz aritzen diren erabiltzaileak identifikatzeko *Umap*[4] plataforma erabili dugu, euskaraz publikatutako edukiari jarraipena egiten diona. Zehatzago esanda, azkeneko hilabetean aktibo eta txioen % 20 gutxienez euskaraz argitaratzen dituzten erabiltzaileen zerrenda jasotzen du *Umap*ek. Iturburu hau erabili dugu 8.189 erabiltzaileren hasierako lagina lortzeko.

Behin erauzi nahi den unibertsoa definituta, Twitter sare sozialetik informazioa erauztera jo dugu. Datuen erauzketa burutu ahal izateko Twitterreko APIa erabili da, honetarako Pythoneko *tweepy* liburutegia hautatuz. Datu-bilketa burutu ahal izateko metodo ezberdinak daude, informazioaren bilaketa modu ezberdinetan oinarritzen direnak. Aintzat hartutako bilaketa teknika ezberdinen artean hiru dira nagusi direnak (Rao *et al.* 2010; Al Zamal *et al.* 2012; Nguyen *et al.* 2013; Morgan-Lopez *et al.* 2017; Marquardt *et al.* 2014):

- *Streaming bidezko erauzketa*: Denbora errealean Twitterreko korrontean sortutako publikazioak jasotzeko balizko metodoa dugu hau. Metodo honi termino edo erabiltzaileen zerrenda bat ematen zaio sarrera datu bezala. Horrela, termino edo erabiltzaile horiekin erlazionatutako publikazioak itzuliko dizkigu metodoak. Metodoari deia egiten zaion momentutik aurrerako datuak jasoko dira bakarrik.

- *Termino bidezko erauzketa*: Aukeratutako terminoak dituzten Twitterreko publikazioak jasotzeko metodoa dugu hau. Metodoari aukeratutako terminoak sarrera datu bezala eman eta 15 minuturo gehienez jota termino horiek dituzten 45.000 txio inguru lortzeko aukera dago. Lortutako txio bakoitzean termino hori azalduko da eta lortutako emaitzak termino horrekiko menpekoak izango dira. Hala ere, metodo honekin soilik azkeneko asteko txioak lor daitezke, txio zaharrak lortzeko aukera galduz eta erauzketa mugatuz.

- *Erabiltzaileen erauzketa*: Aipatutako metodoetako bilaketa tekniketan ez bezala, bilaketa termino konkretuetara mugatu gabe, metodo hau soilik erabiltzaileetan oinarritzen da. Horrela, erabiltzaile zehatz baten publikazioak zein konpartitutako edukiak lortzeko aukera dago metodo honi esker. Hala ere, Twitterreko APIak erabiltzaile bakoitzeko 3.200 txioko muga du, erabiltzaile bakoitzeko gehienez txio kopuru hori lortu ahal izango da. Muga honetaz gain ere, denbora muga bat gehitu behar zaio APIari, 15 minuturo soilik 15 erabiltzaile erauzi ahal dira, datu-bilketa denboran zehar asko luzatuz.

[4] `https://umap.eus/`

Gure ikerketaren asmoa erabiltzaileen ezaugarritzea izanda, eta erabiltzaileen zerrenda luze bat daukagula kontutan hartuta, *erabiltzaileen erauzketa* egitea aukeratu da. Honek esan nahi du zerrendako erabiltzaile bakoitzaren txioak erauziak izango direla, erabiltzaile bakoitzetik azkeneko 3.200 txioak lortuz gehienez jota. Era honetan erabiltzaile bakoitza ordezkatzen duen testu eta interakzio (birtxio) kopuru esanguratsu bat lortuko da.

Txioak lortzeko Twitterreko APIa *tweepy* paketearen bidez erabili da, *timeline extraction* (*erabiltzaileen erauzketa*) modua aukeratuz. Horrela, gure lagineko 8189 erabiltzaileetako bakoitzaren azkeneko 3.200 txio eskuragarriak lortu dira. Datu-bilketa 2018ko maiatzaren 30etik 31ra bitartean egin zen, API akatsengatik erabiltzaile batzuk baztertu ondoren, 7.980 erabiltzaileren 10 milioi txio baino gehiago bilduz oso kostu material txikiarekin. Jarraian, txioak hizkuntzaren arabera sailkatu dira Twitter APIak eskaintzen dituen metadatuak erabiliz, euskaraz idatzitakoak identifikatu eta aukeratuz. Honela 5.198.043 txio pertsonal lortu dira, horietako 3.171.485 (% 61) euskaraz idatzitakoak. Bestalde, 5.473.031 birtxio edo birpublikazio lortu dira ere, horietako 2.891.136 (% 53) euskaraz izanik. Ondorioz, *Heldugazte-osoa*[5] izendatu dugun 6 milioi txioko corpusa osatu da, txio pertsonaletan eta birtxiotan banatzen duguna. *Heldugazte-osoa* corpusaren estatistika nagusiak 3.1 taulan aurkitu daitezke.

	Txio pertsonalak	Birtxioak
Txio kopurua	3.171.785	2.891.136
Unitate lexikoak (terminoak)	1.434.050	813.833
Token kopurua (hitzak)	37.350.268	39.329.204

3.1 Taula – *Heldugazte-osoa* corpusaren ezaugarriak.

3.3 Adin Tartearen Sailkapena: Gazte edo Heldu

Twitter sare sozialeko euskal erabiltzaile guztien artetik gazteak identifikatzea izango da hurrengo pausoa. Horrela, sare sozialetako erabiltzaileen adina automatikoki iragartzeko sistema bat garatu da, ikasketa automatikoan oinarritutako hurbilpenak egokienak izanik (Cesare *et al.* 2017; Morgan-Lopez *et al.* 2017). Honek esan nahi du, erabiltzaile bakoitzari dagozkion entrenamendu datuak eti-

[5]Heldugazte-osoa corpusa publikoki eskuragarri dago: `http://ixa2.si.ehu.es/heldugazte-corpus/heldugazte.osoa.tar.gz`

ketatu beharko direla adinaren arabera, gerora sailkatzailea sortu eta ebaluatzeko. Era berean, iragarri nahi den ezaugarri edo klase moduan zenbakizko adin-tarteak (Rao *et al.* 2010; Al Zamal *et al.* 2012; Morgan-Lopez *et al.* 2017; Marquardt *et al.* 2014) erabili ordez, bizitza-etapak erabiliko dira, sailkapena zehatzagoa baita (Nguyen *et al.* 2013). Hortaz, iragarriko den adina bizitza-etaparen araberakoa izango da, denboran zeharreko esperientzia konpartituak adinaren zenbakia baino adierazgarriagoak baitira (Nguyen *et al.* 2016; Eckert 2017). Honela, interesatzen zaigun *gazte* kategoria malguagoa definitzen dugu, zenbakizko adin muga zehazteko beharrik gabe. Era berean, bizitza-etapa modu bitarrean ulertuko dugu, gazte ez den guztia heldu bezala kontsideratuta, adinaren araberako sailkapena zehatzagoa izan dadin (Rao *et al.* 2010; Al Zamal *et al.* 2012). Horregatik, lan honetan erabiltzaileak bi adin-tarteren arabera sailkatzea erabaki da, hots, gazte eta heldu artean.

Horrela, gazte-heldu sailkatzailea sortzeko, aurrez etiketatutako datuak beharrezkoak dira entrenamendu eta ebaluaziorako. Ataza zehatz hau euskal erabiltzaileek publikatutako euskal edukian zentratzen denez, beharrezkoa da informazio hori ematen duten datuak eskura izatea. Hau da, euskaraz idatzitako txioak euren egileen bizitza-etaparen arabera etiketatu beharko dira, ez baitago aurrez egindako lanik hizkuntza honetan. Hala ere, erabiltzaileen txioak bizitzako etaparen arabera etiketatzeko zailtasunak aurkitu dira, bi arrazoi nagusi direla eta: (i) erabiltzaileen adina ez da ia inoiz agertzen txioen metadatuetan (Cesare *et al.* 2017) eta (ii) bizitzako etaparen arabera txio indibidualak eskuz etiketatzea ez da ataza erraza gizakiontzat. Beheko (1-3) adibideek txio indibidualak bizitzako etaparen arabera eskuz etiketatzeko orduan daukaten zailtasuna erakusten dute.

(1) "Zarauzko triatloian izena ematea lortu gabe, motibazioa falta."
"I have not managed to sign up for the Zarautz triathlon, I am unmotivated."

(2) "A zer nolako eguraldi kaxkarra ez al du gelditu behar edo."
"What a bad weather, shouldn't stop or what."

(3) "5 mila euro, bideo kamera eta telefono mugikor bat eroan dituzte lapurrek."
"5,000 euros, a video camera and a cell phone were taken away by the burglars."

Zailtasun horren aurrean saihesbide metodologiko bat jorratu da, erabiltzaileen idazkera estiloa aintzat hartzen duena. Horretarako testuen idazkera estiloa erabiliko da, adina iradokitzeko ezaugarririk garrantzitsuenak idazkeran oinarritzen baitira (Rao *et al.* 2010; Al Zamal *et al.* 2012; Nguyen *et al.* 2013; Morgan-Lopez *et al.* 2017). Gainera, idazkera estiloa adinarekin lotu daiteke zuzenean,

gazteen idazkera helduena baino informalagoa izanik (Rao *et al.* 2010; Al Zamal *et al.* 2012; Nguyen *et al.* 2013; Morgan-Lopez *et al.* 2017; Rosenthal and McKeown 2011; Nguyen *et al.* 2016). Horrela, gazteek helduek baino hitz kolokialagoak erabiltzen dituzte (Nguyen *et al.* 2013), hizkien errepikapena hitz barruan (Rao *et al.* 2010; Rosenthal and McKeown 2011) zein hiztegiz kanpoko hitzen erabilera (Rosenthal and McKeown 2011; Morgan-Lopez *et al.* 2017) nabarmen gehiago ematen da erabiltzaile gazteen artean. Gainera, idazkera aldatu egiten da adinean aurrera egin ahala, gazteagoek estilo ezohikoagoa edo informalagoa daukatelarik (Nguyen *et al.* 2013: 2016; Cesare *et al.* 2017). Era horretan, gazteen idazteko modua idazkera formal batetik gehien aldentzen dena izango da, alde batetik, helduen idazkera estilo formalarekin erlazionatuz, eta bestetik, gazteena estilo informalarekin.

Aurrekoa finkatuta, euskal erabiltzaileen bizitza-etapa iragartzeko, bi pausutan oinarritzen den hurbilpena proposatzen dugu. Idazkera estiloa iragartzetik hasita, erabiltzaileen bizitza-etapa iragartzea izango da helburua. Lehenbizi, txio solteen idazkera estiloa inferituko da, hauek formalak edo informal bezala sailkatuz. Bigarrenik, kontutan hartuta gazteek idazkera informalagoa daukatela helduek baino, txio informalen kontzentrazio altuena eta baxuena duten erabiltzaileak gazte eta heldu bezala identifikatuko dira hurrenez hurren. Azkenik, erabiltzaile hauen txioak baliatuko dira bizitza-etapa iragartzeaz arduratuko den gazte-heldu sailkatzaile automatiko gainbegiratuak sortzeko. Honi esker erabiltzaile euskaldunak automatikoki bereizi ahalko ditugu gazte eta heldu artean.

Egindako proposamenarekin, txioen testu informal edo formalen sailkapenetik, erabiltzaile gazte edo helduen identifikaziora igaro da. Hautatutako metodologia testu sekuentziak sailkatzean oinarrituko da, horretarako hizkuntzaren prozesamenduko teknikak erabiliz. Gure ataza euskal hiztunetan oinarritzen denez, pausu bakoitzerako anotatutako datu-multzo eta sailkatzaile berriak sortu dira, orain arte horrelako baliabiderik ez baitzegoen. Horrela, hautatutako hiztun-komunitateko erabiltzaileak gazte edo heldu diren sailkatzeko honako pausu hauek jarraitu dira:

(1) Txio mailan (i) testuaren eskuzko etiketatzea idazteko estiloa (informal/formal) oinarritzat hartuta eta (ii) horiek erabili *informal-formal* sailkatzaile bat sortzeko, txioen testua informal edo formal moduan automatikoki sailkatzeko.

(2) Erabiltzaileen mailan (i) bizitza-etaparen (gazte/heldu) etiketatze erdi automatikoa *informal-formal* sailkatzailea erabilita; (ii) erabiltzaile gazte eta

helduen txioak erabilita *gazte-heldu* sailkatzailea sortu, txioen testua gazte edo helduena den automatikoki iragarriko duena.

(3) Sortutako sailkatzaileen aplikazioa *heldugazte-osoa* corpusean, txio mailako (informal-formal) eta erabiltzaile mailako (gazte-heldu) hurbilpenak konparatzeko.

3.3.1 Metodologia

Gure ataza testu sekuentzien sailkapen gainbegiratuan oinarrituko da, aurretik ere horrela adinaren detekzioa horrela aurkeztua izan baita (Rao *et al.* 2010; Al Zamal *et al.* 2012; Nguyen *et al.* 2013; Morgan-Lopez *et al.* 2017). Horrek esan nahi du, sailkatzaileek anotatutako datuetatik ikasiko dutela eta testu sekuentzia bat edukita kapazak direla aurretik etiketatu den atributua aurreikusten. Geure kasuan testu sailkapena bi modu desberdinetan egin beharko da, kasu batean idazkera estilo formal edo informala aurreikusiz, eta beste kasuan, testuaren egilea gaztea edo heldua den iragarri beharko du. Hala ere, atributu ezberdinak (idazkera estiloa edo adin tartea) sailkatu behar diren arren, testu sekuentzien sailkapen ataza da oinarria baina bakoitzak bere entrenamendu datu propioak izango ditu.

Jarraian, testu sekuentziak idazkera estiloaren eta bizitza-etaparen arabera sailkatzeko erabilitako metodoak aurkezten dira: (i) perplexitate distantzian oinarritutako metodo estatistikoa (Gamallo *et al.* 2017); ikasketa automatikoa eta hitzbektoreak baliatzen dituzten (ii) FastText (Bojanowski *et al.* 2017) eta (iii) IXA pipes (Agerri *et al.* 2014) metodoak; ikasketa sakona eta testuingurudun hitzbektoreak erabiltzen dituzten (iv) Flair (Akbik *et al.* 2018) eta (v) BERT ereduan (Devlin *et al.* 2019) oinarritutako hizkuntza-eredu eleaniztun (Devlin *et al.* 2019) zein elebakarrak (Agerri *et al.* 2020).

Estatistikoa: Perplexitatea

Testu sekuentziak sailkatzeko, *hizkuntz distantzia* kontzeptua erabiliko da, hizkuntza aldaera bat beste batetik zein ezberdina den adierazten lagunduko duena (Gamallo *et al.* 2017). Hizkuntza aldaeren arteko distantzia neurtzeko, *perplexity* delako balio estatistikoa erabiliko da. Perplexitatea oso erabilia den ebaluazio metrika da hizkuntza-ereduen kalitatea neurtzeko (Chen and Goodman 1999). Horrez gain, zeregin zehatzetarako ere erabili izan da, hots, txio formal eta kolokialen artean sailkatzeko (González Bermúdez 2015) edo antzekoak diren hizkuntzen artean hizkuntza identifikaziorako (Gamallo *et al.* 2017). Perplexitate neu-

rriak, eredu probabilistiko batek lagin bat iragartzeko daukan ahalmena neurtzen du. Hizkuntza-ereduei aplikatzean, honek testu sekuentzia bat iragartzeko daukan probabilitate negatiboa adierazten du. Funtsean, perplexitate balio geroz eta txikiagoek adierazten dute ereduak ziurtasun handiagoa duela bere iragarpenetan.

Hizkuntza distantzia kontzeptura aplikatuta, hizkuntza aldaeren arteko distantzia linguistikoa kalkulatuko da perplexitatea baliatuta, karaktereetan oinarritutako n-gramen bitartez. Karaktere n-gramek informazio lexikoa eta morfologikoa kodetzeko gaitasuna dute. Horrez gain, n-grama luzeek (5 karaktere edo gehiago) erlazio sintaktikoak eta sintagmatikoak ere kodetzen dituzte, hitz baten amaiera eta hurrengoaren hasiera jasotzeko gai direlako sekuentzia berdinean (Gamallo *et al.* 2017). Hizkuntza aldaeren arteko distantzia kalkulatzeko, karaktereetan oinarritutako n-gramak erabili dira, 7-gramak erabiliz (Gamallo *et al.* 2017). Honela, aurrez zehaztutako hizkuntza-eredu bat erabilita, testu sekuentzia zehatzen perplexitatea kalkulatuko da ereduarekiko. Perplexitate neurri horrek testu sekuentzia eredutik zenbat aldentzen den adieraziko du, balio altuagoek hizkuntza distantzia altuagoa adieraziko dutelarik.

Hitz-bektore estatikoak: FastText

Hitz errepresentazioak edo hitz-bektoreak oso erabiliak dira hizkuntzaren prozesamenduan. *Word2vec* (Mikolov *et al.* 2013b) edo *GloVe* (Pennington *et al.* 2014) bezalako teknikei esker, esanahi semantiko antzekoa duten hitzek, hitz-bektore antzekoa izango dute, hitzen errepresentazio jarraitu bat lortuz. *FastText* (Bojanowski *et al.* 2017) ereduan hitz bakoitza bere kabuz eta karaktere n-grama kate bat bezala tratatzen da, informazio morfologikoaren txertaketa ahalbidetuz. Euskara bezalako morfologia aberatseko hizkuntzek, batez ere, hitzen (edo azpi-hitzen) errepresentazio hauetatik etekina atera beharko lukete. Gainera, FastText-ek hizkuntza askotarako aurrez prestatutako ereduak eskaintzen ditu, euskara barne (Grave *et al.* 2018). Eskuragarri dagoen euskarazko eredua *Common Crawl* zein *Wikipedia*ko datuekin entrenatu da, CBOW eta posizio-pisuak erabiliz, 300 dimentsio, 5 karaktereko luzeera duten n-grama, 5 tamainako leihoa eta 10 negatibo erabilita (Grave *et al.* 2018).

Ezaugarri testualak: IXA pipes

IXA pipes (Agerri *et al.* 2014) metodoa, ikasketa automatikoan oinarritua dago, sailkatzaile bezala pertzeptroiaren algoritmoa (Collins 2002) erabiliz. Sistema honek entrenamenduko datuetatik eratorritako informazio lokala konbinatzen du

etiketatu gabeko testuetatik induzitutako ezaugarrien clusterrekin. Era honetan, entrenamendu datuetako hitzak, hiru errepresentazio modu ezberdinen konbinaketarekin egingo da: Brown (Brown *et al.* 1992) klusterrak, Clark (Clark 2003) klusterrak eta word2vec (Mikolov *et al.* 2013b) klusterrak. Hitz bakoitza, aipatutako tekniketatik eratorritako klusterren konbinaketarekin ordezkatzen da. Horretarako, sekuentziako hitzak clusterretako lexiko bakoitzean dauden hitzekin mapatzen da. Hitzen errepresentazioak burutzeko aipatutako hiru teknika ezberdinetatik (Brown, Clark eta word2vec) eratorritako clusterrak pilatu eta konbinatzen dira. Clusterren ezaugarri horiek, hitz bakoitzari talde batekiko kidetasuna ematen diote, entrenamenduan ikusi gabeko hitzak, ikusitakoekin erlazionatzen dira cluster berdinean azalduz gero. Horrela, eskuz etiketatu beharreko datu kopuru handiekiko dependentzia arindu egiten da, etiketatutako datu-multzo txikiekin ere sailkapen egoki bat egitea ahalbidetuz (Agerri *et al.* 2014). Metodo honek emaitza onak lortu ditu hainbat atazatan, hala nola izendun entitateen identifikazio (Agerri and Rigau 2016) zein iritzi erauzketan (Agerri and Rigau 2019) hainbat hizkuntzetarako, euskara barne.

Testuingurudun hitz-bektoreak: Flair

Flair-ek testuingurua barneratzen duten hitz-bektoreei eta ikasketa sakoneko sistemari egiten die erreferentzia (Akbik *et al.* 2018). Flair hitz-bektoreen ezaugarri nabarmena, hitzak testuinguru zehatzaren arabera errepresentatzean datza. Hitz bakoitza modu isolatuan tratatzen duten hitz-bektore estatikoek ez bezala (word2vec, GloVe, FastText...), Flair-ek inguruko hitzak hartzen ditu kontuan, errepresentazioa inguruko hitzek baldintzatua izanik. Testuingurudun hitz-bektore hauek eskuragarri daude hainbat hizkuntzetarako, euskara barne (Akbik *et al.* 2018). Horrez gain, Flair ere sailkapenerako sistema bezala erabilia izan daiteke, horretarako *pipeline* berezia eskaintzen baitu. Testu sekuentzien sailkapenerako, karaktereetan oinarritutako hitz-errepresentazioak BiLSTM-CRF (Huang *et al.* 2015) arkitekturan oinarritutako sistema batetik pasatzen dira. Gainera, sailkapena atazetarako ere, mota ezberdinetako hitz-bektoreak pilatzeko aukera ematen du (FastText esaterako), hitz-bektore ezberdinek ematen duten informazioaren konbinaketa ahalbidetuz.

Flair hitz-bektoreak arrakastaz aplikatu dira sekuentzia etiketatze zereginetan, erreferentziazko hainbat datu-multzo publikotan emaitzarik onenak lortuz (Akbik *et al.* 2018). Zehazki, emaitza onak lortzen ditu ingelesezko izendun entitateen identifikazioan zein kategoria morfosintaktikoaren iradokizunean (Akbik *et al.* 2018). Horrez gain, ataza horietan ere emaitza onak lortu ditu euskararako (Agerri

et al. 2020; Fernandez de Landa and Agerri 2021a), ezaugarri testualetan (Agerri *et al.* 2014) eta hitz-bektore estatikoetan (Bojanowski *et al.* 2017) oinarritutako hurbilpenak gaindituz, baina, Transformerretan oinarritutako hizkuntza-ereduak (Agerri *et al.* 2020) baino xumeago arituz.

Transformerretan oinarritutako hizkuntza-ereduak: mBERT eta BERTeus

Hizkuntzaren prozesamenduko beste zeregin askotan bezala, testu sailkapeneko atazetan ere errendimendurik onena erakusten duten sistemak Transformerretan (Vaswani *et al.* 2017) oinarritutako hizkuntza-ereduak dira (Devlin *et al.* 2019; Liu *et al.* 2019). Hizkuntza-eredu hauek hitzen errepresentazio aberatsak sortzeko ahalmena daukate, BERT (Devlin *et al.* 2019) bezelako ereduek sekuentziako hitz guztiak hartzen baitituzte kontuan, eta horrela testuinguruaren ulermen sakonagoa garatzen dute. Gure ataza zehatza euskarazko testu sekuentziak sailkatzean oinarritzen denez, euskara barnebiltzen duten BERT ereduak erabiliko ditugu. Bi izango dira erabili eta alderatuko ditugun ereduak: (a) mBERT eredu eleaniztuna (Devlin *et al.* 2019) eta (b) BERTeus (Agerri *et al.* 2020) euskarazko eredu elebakarra.

mBERT: Eredu hau BERT ereduaren bertsio eleaniztuna da, Wikipediako 104 hizkuntza handienekin aurrez-entrenatua dagoena. Eredu eleaniztun hauek oso ondo funtzionatzen dute baliabide handiko hizkuntzekin erlazionatutako zereginetan, esate baterako, ingelesa edo gaztelania. Hala ere, baliabide urriko hizkuntzak ez daude behar bezala ordezkatuta hizkuntza eredu erraldoi hauetan (Agerri *et al.* 2020). Besteak beste, entrenamendurako corpusean hizkuntza txikiek datu gutxiago daukate ingelera edo gaztelera bezalako hizkuntzekin konparatuta (Devlin *et al.* 2019; Conneau *et al.* 2020). Horrez gain, badirudi eredu eleaniztunek emaitza hobeak lortzen dituztela antzeko egitura duten hizkuntzekin sortuak direnean (Karthikeyan *et al.* 2020). Hortaz, euskara bezalako hizkuntza gutxitu eta isolatu batek, hizkuntza-eredu eleanitzetan azpi ordezkatua izateko arriskua dauka.

BERTeus: Eredu hau BERT arkitekturan oinarritutako euskarazko hizkuntza eredu elebakarra da. Euskal hizkuntza barnebiltzen duen eredu bat baino, euskal hizkuntzarako prestatutako eredu propio bat dugu hau. Agerri *et al.* (2020) lanean erakusten dutenez, euskarazko BERT eredu elebakarra entrenatzeak emaitza hobeak lortzen ditu bertsio eleaniztunarekin alderatuta. Eredu eleanitzetan hizkuntza gutxituek daukaten azpi-errepresentazioarekin haustea da eredu honen

asmoa. Hortaz, euskarazko testua sailkatzea helburu duen gure atazarako ere, eredu honekin esperimentuak egingo ditugu.

3.3.2 Txio mailako hurbilpena: informal-formal

Euskarazko testu sekuentziak informalak edo formalak diren identifikatzea izango da atal honen asmoa. Horretarako, datu-multzo zein sailkatzaile berriak sortuko dira. Lehenbizi, *heldugazte-osoa* corpusetik ausaz aukeratutako euskarazko adierazpenak idazkera estiloaren arabera eskuz etiketatuko dira, formal edo informal bezala etiketatutako txio sorta bat lortuz. Bigarrenik, txioak automatikoki etiketatuko dituen sailkatzailea sortuko da, metodo estatistiko eta ikasketa automatikoan oinarritutako sistemak eta hitzen errepresentazio teknika anitzak probatuz. Era horretan, euskarazko testu sekuentzien idazkera estiloa automatikoki iragartzea lortuko da.

Heldugazte datu-multzoa: txio informal edo formalak

Txioen euskara erregistro formal edo informala iragartzeko asmoz, txio pertsonalak eskuz etiketatuko dira. Horretarako txioen idazketa moduan oinarritu gara, hiztegitik kanpoko hitzak edo esamolde kolokialak dituzten txioak informal gisa sailkatzean oinarritzen dena. Metodologia hau txio formalak eta kolokialak sailkatzeko aurreko lanetan oinarrituta dago (González Bermúdez 2015). Jarraian lan honetan sailkatzen ditugun txio mota formal eta informalen adibide bana aurkezten dugu. Bi kasuetan idazteko estiloaren desberdintasunak ikus daitezke. Horrela, txio informaletan euskalki edota forma kolokialak agertzen dira (*ein, examin, bau, det*), txio formaletan aldiz, euskal gramatika estandarreko formak azaltzen dira.

(1) **Txio informala**:
"inoizz ezdet ein mateko examin bau au baino okerro."
"This is the worst exam I have ever done."

(2) **Txio formala**:
"Killian Jornet fenomenoa da Zegama-Aizkorri irabazi du beste behin. Non dago mendizale gazte honen muga?"
"Killian Jornet is a phenomenon he won the Zegama-Aizkorri again. Where is this mountaineer's limit?"

Kontutan hartuta datuen iturria erabiltzaile bakoitzaren txio pertsonalak direla, *Heldugazte-osoa* corpusaren zati txiki bat etiketatu da, sailkatzaile ezberdinak garatu eta ebaluatzeko asmoarekin. Txioak idazteko estiloaren arabera sailkatzeko asmoa edukita testuaren garbiketa bat egin da, karaktere alfanumerikoak dituzten hitzak soilik mantentzeko. Horrela, emotikonoak, hashtag-ak, erabiltzaileen izenak (@) eta URL estekak kendu dira. Gainera, 5 token baino gehiago dituzten txioak bakarrik kontuan hartu dira. Bukatzeko, ausaz 1.000 txio pertsonal aukeratu dira eta anotatzaile batek idazteko moduaren arabera eskuz anotatu ditu. Era honetan, txio batek *formal* etiketa eramango du, hizkuntza estandarrean idatzia izan bada, edo *informal* etiketa, txioa modu kolokialean idatzia izan denean.

Era honetan eskuz anotatutako *Heldugazte*[6] datu-multzoa sortu da, idazkera estiloan oinarriturik euskarazko txio formal eta informalek osatua. Datu-multzoaren egiturari erreparatuz (3.2. taula), ikusi daiteke 1.000 txioz osatzen den etiketatutako corpus honek txio formal eta informalen kopuru antzekoa daukala, corpus orekatu bat izanik. Txioen batez besteko luzera 10 tokenekoa da, txio laburrenak 5 token dauzka eta luzeenak 34 token. Ikasketa automatikoan oinarritutako sailkapenak burutu eta ebaluatzeko, datu-multzoa bi zatitan auzaz banatu dugu, % 65 utziz entrenamendurako eta % 35 ebaluaziorako.

	train	test	totala
Formal	312	180	492
Informal	338	170	508
totala	650	350	1.000

3.2 Taula – Heldugazte (informal-formal) datu-multzoaren ezaugarriak.

[6] *Heldugazte* datu-multzoa (informal-formal) publikoki eskuragarri dago: `https://github.com/ixa-ehu/heldugazte-corpus`

Esperimentuen ezarpenak

Testu sekuentzia baten idazkera estiloa zein izango den iragartzea izango da ataza honen helburua, hots, testuaren erregistroa formala edo informala den inferitu beharko da. Horretarako testu sailkapen ataza bezala definituko da, txioaren idazkera estiloa modu bitarrean sailkatuz, formal edo informala. Esperimentuak egiteko, *Heldugazte* entrenamendu-multzoa entrenamendurako erabili dugu eta test-multzoa ebaluatzeko. Sarrera datuetako testuan gutxieneko aurreprozesaketa egiten dugu; URLak, hashtag-ak eta erabiltzaile-izenak kentzen ditugu, etiketatxio bikoteak utziz, aurreko atalean (3.3.2) azaldutako adibideetan erakusten den bezala. Horrela, 3.3.1. atalean azaldutako hainbat sistema erabili dira sailkatzaileak sortzeko: (i) perplexitatean oinarritutako hizkuntz distantziaren metodoa (Gamallo *et al.* 2017), (ii) *IXA pipes* metodoa (Agerri *et al.* 2014), (iii) SVM eredua *FastText* hitz-bektoreekin (Mikolov *et al.* 2018) eta (iv) *Flair* eredua (Akbik *et al.* 2018).

(i) Perplexitatea: Testuak formal edo informal gisa sailkatze aldera, hizkuntza eredu formal baten eta testu bakoitzaren arteko perplexitate distantzian oinarrituko da sistema hau. Lehenbizi, karaktere 7-grametan oinarritutako hizkuntza eredu bat sortuko dugu Egunkaria eta Berria egunkarien testuez osatutako corpus bat erabilita (Gamallo *et al.* 2017). Bigarrenik, datu-multzoko testu edo instantzia bakoitzaren perplexitate distantzia kalkulatuko da sortu den hizkuntza eredu formalarekiko. Perplexitatearen balioa zenbat eta handiagoa izan, orduan testua hizkuntza eredu formaletik geroz eta gehiago aldentzearen seinale izango da. Hortaz, testu bat formaltzat jotzeko perplexitateak hartu beharko lukeen balio minimoa aukeratu beharko da. Horretarako, balio zehatz bat aurrez zehaztu beharko da, informal eta formaltasunaren muga zehaztuko duena. Hau da, atalase bat finkatu behar da, non balio horretatik behera testuak formalak kontsideratuko diren eta balio horretatik gora testuak informaltzat joko diren. Atalase edo muga-balioa zein izango den finkatzeko entrenamendu-multzoa erabiliko da, 0tik 10era [0, 10] doazen balioak probatuz eta 0,01 iteratuz aldiro. Horrela, perplexitate distantzia 4,4 balioan zehaztu da, entrenamendu-multzoan emaitza onenak lortutako balioa izanik. Hizkuntza eredu formalarekiko sekuentzia bakoitzaren perplexitate balioa kalkulatu ostean, atalase baliotik gorako guztiak informaltzat (> 4,4) joko dira, balio txikiagoak formaltzat (< 4,4) jotzen diren bitartean.

(ii) SVM FastText: Esperimentu honetarako, instantzia edo esaldi bakoitzeko hitzak FastText ereduan duten hitzen errepresentazio bektorialarekin ordezkatu ditugu. Hitzen ordezkapenerako, aurrez entrenatutako FastText-en euskarazko hitz-bektoreak erabili dira (Grave *et al.* 2018). Instantzia bakoitza hainbat hitzez

osatuta dagoenez, esaldi mailako errepresentazio orokor bat egingo da, horretarako hitz-bektore guztien batezbesteko bektore bat sortuz (Kenter *et al.* 2016). Instantzia bakoitzaren batez-besteko errepresentazioak, *Support Vector Machine* (SVM) sailkatzaile bat entrenatzeko erabiliko dira. Horretarako, *Scikit-learn* implementazioa (Pedregosa *et al.* 2011) erabiliko da. Hiperparametroen aukeraketa entrenamendu-datuen gainean bost iteraziodun baliozkotze gurutzatua (5-fold CV) aplikatzen egin da, C = 1,1 zehaztuz.

(iii) IXA pipes: IXA pipes sailkatzailea entrenatzeko instantzia bakoitzeko hitzak ezaugarri testualetatik eratorritako errepresentazioen clusterrekin ordezkatuak izango dira. Esperimentuetarako aurrez entrenatutako dimentsio anitzeko clusterrak erabili dira, Elhuyar Web Corpusean (Leturia 2012) eta Tokikomen corpusean (tokiko albiste-guneak arakatuz lortutako 600M hitzeko corpusa) Brown, Clark eta W2V algoritmoen ezaugarrietatik eratorriak. Honela, konbinaketa ezberdinak frogatu ostean, honako aukeraketa egin da eredua entrenatzeko: Elhuyar Web Corpuserako, Brown 3.200 klase, Clark 600 klase eta word2vec 300 klase; eta Tokikomeko corpuserako, Clark 300 klase eta word2vec 500 klase. Ezaugarrien aukeraketa entrenamendu-datuen gainean bost iteraziodun baliozkotze gurutzatua (5-fold CV) aplikatzen egin da.

(iv) Flair: Arkitektura hau erabili da dokumentuen sailkapenean oinarritutako sistemak entrenatzeko eta horretarako hitz-bektore errepresentazio ezberdinak erabili dira: euskararako Flair testuingurudun hitz-bektoreak, Flair karakterebektoreak eta aurretik erabilitako FastText-en euskarazko hitz-bektoreak. Euskararako Flair testuingurudun hitz-bektoreak entrenatzeko hainbat iturrietatik eratorritako euskarazko edukia erabili da, 249M inguru hitzez osatua. Hitzen errepresentazio egokienen aukeraketa burutzeko, entrenamendu-datuen gainean bost iteraziodun baliozkotze gurutzatua (5-fold CV) aplikatuz egin da.

Ebaluazio emaitzak

Atal honetan, txioak informalak edo formal diren iragartzeko sailkatzaile bakoitzak daukan ahalmena erakusten da. 3.3. taulak, aurreko atalean deskribatutako sistemek ebaluazio-multzoko testu sekuentzien idazkera estiloa iragartzean lortzen dituzten emaitzak erakusten ditu. Azpimarratu behar da eskuzko ingeniaritza minimoa izan dela sistemak garatu direnean, atributu sorkuntza (feature engineering) minimoa eginez. Era honetan, entrenamendu datuetara gehiegi egokitzen diren ezaugarriak gehitzea saihestu da, errendimendua galduz orokortzeko gaitasunaren truke. Horren bidez, orokortzeko gai eta sendoak diren sailkatzaileak garatu nahi izan dira. Horrela, nahiz eta etiketatutako datu-multzoa txikia izan,

domeinuz kanpoko (out-of-domain) atazetan ere inferentzia egokiagoak lortzeko aukera irekiz.

Emaitzei begira, ikusi daiteke asmatze tasa guztiak nahiko gertu daudela bata besteagandik eta 0,8tik gorakoak direla. Hala ere, Perplexitatean oinarritutako erdi-gainbegiratutako metodo estatistikoak emaitza okerrenak lortzen ditu. Bestelako metodo guztiak, gainbegiratuak, emaitza hobeak lortzen dituzte, ikasketa automatikoaren abantailak erakutsiz. Bestalde, ikusi daiteke Flair eta IXA pipes sistemek emaitza hobeak lortzen dituztela, 0,86 eta 0,88ko asmatze tasak lortuz hurrenez hurren. Bi sistema hauek gailentzeko arrazoia, hitzen errepresentazioa burutzeko orduan hainbat metodo ezberdinen konbinaketa erabiltzen dituztela izan daiteke. Era honetan, Flair sistemak Flair eta FastText errepresentazioak erabiltzen dituen biatrtean, IXA pipes sistemak Brown, Clark eta word2vec hitzen errepresentazioak lortzen ditu. Era honetan, esan beharra dago idazkera estilo formal eta informal artean ezberdintzeko hurbilpen arrakastatsuenak ikasketa automatikoan eta hitz-bektoreen errepresentazio anitzak konbinatzen dituztenak dira.

Sistema	Asmatzea	Etiketa	Erroreak	Doitasuna	Estaldura	F1 Balioa
Perplexitatea	0,825	Informal	26	0,805	0,847	0,825
		Formal	35	0,848	0,806	0,826
SVM FastText	0,832	Informal	24	0,843	0,823	0,836
		Formal	33	0,834	0,828	0,829
IXA pipes	0,886	Informal	20	0,882	0,881	**0,882**
		Formal	20	0,889	0,888	**0,889**
Flair	0,866	Informal	22	0,869	0,858	0,863
		Formal	24	0,868	0,877	0,872

3.3 Taula – Ebaluazio emaitzak Heldugazte (informal-formal) test-multzoan.

Emaitzak ikusita, Flair zein IXA pipes sailkatzaileak egokiak direla uste dugu txio pertsonalak klase formal eta informaletan sailkatzeko. Aplikazio praktikoari begira, IXA pipes eredua aukeratzea erabaki da, Flair sistemak denbora eta konputazio (GPU bat) eskakizun handiagoak behar baititu entrenamendu zein inferentziarako. Horrela, *heldugazte-osoa* corpusean azaltzen diren euskal erabiltzaileen milioika txioak errazago eta azkarrago etiketatu ahal izango dira.

3.3.3 Erabiltzaile mailako hurbilpena: gazte-heldu

Euskarazko testu sekuentziak gazteenak edo helduenak diren identifikatzea izango da atal honen asmoa. Helburu zehatz honetarako, datu-multzo zein sailkatzaile berriak sortuko dira. Lehenbizi, etiketatutako datuak erdi-automatikoki lortzeko metodologia berri bat proposatu dugu, erabiltzaileen idazkera estiloan oinarritzen dena. Erabiltzaileen publikazioak idazkera estiloaren arabera desberdintzen dituen informal-formal sailkatzailea (3.3.2. atala) baliatuta, erabiltzaile informalenak zein formalenak identifikatu eta gazte edo heldu bezala kontsideratuko dira hurrenez hurren. Erabiltzaileen bizitza-etapa aintzat hartuz, bakoitzaren publikazioetara etiketa proiekzioa egin da, heldu eta gazte etiketadun txio mailako datu-multzoa osatuz. Bigarrenik, etiketatutako datu-multzoa sortu eta gero, erabiltzaileen txioak automatikoki etiketatuko dituen sailkatzailea sortuko da, Transformerretan oinarritutako hizkuntza-eredu elebakar eta eleaniztunak probatuz. Era horretan, txio baten autorearen bizitza-etapa automatikoki iragartzeko ahalmena lor daiteke. Hau da, euskarazko txio bat gazte edo heldu batek idatzia izan den identifikatu ahalko da.

Heldugazte-age datu-multzoa: gazte edo helduen txioak

Txioak gazteenak edo helduenak diren ezberdinduko dituen *gazte-heldu* sailkatzaileak entrenatu eta ebaluatzeko, datu-multzo propioa sortuko dugu. Horretarako, metodo erdi-automatiko bat proposatzen dugu, idazkera estiloa aintzat hartzen duena:

(1) Lehenik eta behin, *Heldugazte-osoa* corpuseko erabiltzaileen 6M txioak automatikoki etiketatu dira idazkera estiloaren arabera informal-formal sailkatzailea erabilita (3.3.2. atalean).

(2) Bigarrenik, erabiltzaileak euren denbora-lerroko txio informalen proportzioaren arabera ordenatu ditugu. Mutur bateko erabiltzaileek txio informalak izango lituzkete batez ere eta beste muturreko erabiltzaileek txio formalak.

(3) Hirugarrenik, muturreko 100 erabiltzaileen (50 informalenak eta 50 formalenak) denbora-lerroen eskuzko ikuskapen bat egin da. Urrats honetan bereziki lagungarria izan da ikuskapena erabiltzaile mailan egitea, denbora-lerroak erabiltzailea ezaugarritzeko testuinguruko informazio gehiago eskaintzen duelako. Eskuzko azterketa honek egiaztatzen du erabiltzaile gazte

eta helduen etiketatze erdi-automatikoaren emaitzak onargarriak direla (3.4. taulako adibidea).

(4) Laugarrenik, sailkapenaren mutur informalenean dauden 500 erabiltzaileak erabiltzaile gazte kontsideratuko dira eta mutur formalenean dauden 500 erabiltzaileak heldu bezela kontsideratuko ditugu. Proposatutako metodo berri honi esker, gazte eta heldu gisa anotatutako 1.000 erabiltzaile lortu dira.

Erabiltzaile mailako anotazio erdi-automatikoa burutu ostean, txio mailako datuak lortzera igaroko gara berriz ere. Erabiltzaile mailatik txio mailara igaroz, erabiltzaileak datu kopuru berdinarekin ordezkatu nahi ditugu, datu-multzo orekatu bat lortuz. Horretarako, erabiltzaile bakoitzeko, ausazko 80 txio aukeratu dira, txio bakoitza erabiltzaileari egotzitako heldu edo gazte etiketarekin anotatuz. Era honetan datu-multzo esanguratsu eta heterogeneo bat sortu da, txio indibidualak gazte edo heldu bezala anotatuta dituena.

Etiketa	**Edukia** (txioa)
heldu	Taldeak mikel laboaren lanean oinarritu du bere hurrengo diskoa. *The band has based their next album on the work of Mikel Laboa.*
heldu	Gure herriko ateak zabalik dituzu. *The doors of our town are opened.*
gazte	Buaa q follaa eun guztia eon zea ikasi ordez jolasateeenn jajaja. *How lucky! You have been all day playing instead of studying hahaha.*
gazte	Batzutan ze gutxi aguantatze zaituten. *Sometimes I can't stand you.*

3.4 Taula – Heldugazte-age (gazte-heldu) datu-multzoko adibideak.

Emaitza bezala *Heldugazte-age* [7] datu-multzoa dugu. Datu-multzo handi eta orekatu honek 80 mila txio dauzka gazte edo heldu bezala etiketatuta (ikusi 3.5 taula). Datu-multzoa osatzen duten txioen adibide bat 3.4 taulan ikus daiteke. Datuak entrenamendu (*train*), garapen (*dev*) eta ebaluazio (*test*) multzoen arabera ausaz banatu dira esperimentuetarako. Horrela, klase bakoitzeko 24K txio daude

[7] Heldugazte-age datu-multzoa (gazte-heldu) publikoki eskuragarri dago: `https://github.com/joseba-fdl/heldugazte-age-corpus`

eskuragarri entrenamendurako eta 8K txio garapen zein ebaluaziorako, hurrenez hurren.

	train	dev	test	totala	*Erabiltzaileak*
gazte	24.000	8.000	8.000	40.000	500
heldu	24.000	8.000	8.000	40.000	500
totala	48.000	16.000	16.000	80.000	1.000

3.5 Taula – Heldugazte-age (gazte-heldu) datu-multzoaren ezaugarriak. Erabiltzaile mailan etiketatutako txioak bizitza-etaparen arabera.

Esperimentuen ezarpenak

Ataza zehatza testu sekuentzia baten egilearen bizitza-etapa (gazte/heldu) iradokitzean oinarrituko da, 3.3.2. atalean ez bezala, oraingo honetan ataza ez da idazkera estiloan oinarrituta egongo. Hala ere testu sailkapen ataza berdina da: sarrera datuak testu sekuentziak (txioak) izanda, gazte edo heldu etiketa iragartzea da asmoa. Esperimentuak egiteko, *Heldugazte-age* entrenamendu-multzoa entrenamendurako erabili dugu eta test-multzoa ebaluatzeko. Sarrerako txioetan gutxieneko aurreprozesaketa egiten dugu; URLak, hashtag-ak eta erabiltzaile-izenak kentzen ditugu, etiketa-txio bikoteak utziz, 3.4. taulan azaldutako adibideetan erakusten den bezala.

Aurreko atalean sortutako *Heldugazte-age* datu-multzoa, beraz, 3.3.1. atalean aurkeztutako hiru testu sailkatzaile ezberdin trebatzeko erabiliko da: (i) *IXA pipes* (Agerri *et al.* 2014), (ii) *mBERT* (Devlin *et al.* 2019) eta (iii) *BERTeus* (Agerri *et al.* 2020). IXA pipes (Agerri *et al.* 2014) aukeratu da oinarri-lerro bezala 3.3.2. atalean lortutako emaitza altuengatik, ezarpen berdinak erabiliz sistema entrenatzeko. Horrez gain, mBERT eta BERTeus-en errendimenduak alderatuko ditugu bizitzako etapa detektatzeko atazan, eredu eleaniztun eta elebakarren portaerak neurtzeko asmoa baitugu baliabide urriko hizkuntzen arloan. Bi ereduentzat oinarrizko birdoitze hiperparametro berberak erabili dira (Agerri *et al.* 2020).

Ebaluazio emaitzak

3.6 taulan aurreko atalean deskribatutako sistemak test-multzoaren gainean erabilita lortzen diren emaitzen berri ematen da. Lehenik eta behin, azpimarratu beharra dago aukeratutako metodo guztiek emaitza altuak lortzen dituztela, 0,95-etik gorako asmatze tasa zein F1 balioak erdietsiz. Gainera, sistemen arteko al-

deak ez dira horren handiak, nahiz eta BERTeus-ek emaitza onenak lortu dituen. Oinarri-lerro bezala hautatutako IXA pipes metodoak mBERT-ek bezain emaitza onak lortu ditu, metodo honen fidagarritasuna frogatuz beste behin ere. Bestalde, Transformerretan oinarritutako hizkuntza-ereduei erreparatuz, BERTeus eredu elebakarrak mBERT eredu eleaniztunak baino emaitza hobea lortu duela ikusi da. Honek erakusten du hizkuntza zehatzetan oinarritzen diren lanabesak garatzea beharrezkoa dela, batez ere baliabide urriko hizkuntzetarako (Agerri *et al.* 2020).

Sistema	Asmatzea	Doitasuna	Estaldura	F1 Balioa
IXA pipes	0,956	0,977	0,935	0,955
mBERT	0,955	0,972	0,936	0,954
BERTeus	**0,963**	0,968	0,958	**0,963**

3.6 Taula – Ebaluazio emaitzak Heldugazte-age (gazte-heldu) test-multzoan.

Sailkatzaileek lortutako emaitza altuen aurrean giza ebaluazio baten beharra ikusi dugu, lagin baten eskuzko anotazio baten bitartez egingo dena. Eskuzko azterketa honetarako, datu-multzoko test-sortatik ausaz aukeratutako 200 txio eskuz etiketatzea erabaki da. Bi giza anotatzailek 200 txioak etiketatu dituzte, 0,78 puntuko anotatzaileen arteko adostasuna eta 0,55ko Kappa balioa lortuz. Zenbaki hauek erakusten dute anotatzaileen arteko adostasuna moderatua izan dela, atazaren zailtasuna agerian utziz. Gainera, bi anotatzaileen asmatze tasa 0,795 eta 0,775 puntuetakoa izan da hurrenez hurren. Puntuazio hauek 3.6. taulan jasotako sistema automatikoen emaitzekin alderatzean, argi uzten dute txio mailako gazte edo heldu etiketak eskuz esleitzea oso lan zaila dela. Anotatzaileen arteko desadostasun eta asmatze tasa baxuek erakusten dute *Heldugazte-age* datu-multzoa lortzeko proposatutako metodoaren (3.3.3 atala) eraginkortasuna, gizakiek kostata egin dezaketena, gure metodoarekin zehaztasun eta erraztasun handiagoarekin eginez.

3.3.4 Aplikazioa

Orain arte, erabiltzaile gazteak identifikatzeko helburua duen atal honetan (3.3. atala), testu-sekuentziak sailkatzeko baliagarriak diren *informal-formal* (3.3.2. atala) eta *gazte-heldu* (3.3.3. atala) sailkatzaileak proposatu dira. Hala ere, sistema hauek tamaina txikiko (txioak: 240 karaktere baino gutxiago, bat edo bi esaldi) testu-sekuentziak sailkatzeko entrenatuta daude. Hau da, erabiltzaileen adierazpen zehatz edo txioak sailkatzeko pentsatuta daude, soilik erabiltzaile batek sor-

tzen duen edukiaren zati txiki bat sailkatzeko prestatua. Hortaz, esaldi edo txio mailatik, dokumentu edo erabiltzaile mailako sailkapen bat ematera igaro beharko gara. Hau da, erabiltzaile baten publikazio zehatzak sailkatuta, erabiltzaile mailako sailkapen orokorra lortzea da helburu. Txio mailako sailkapenetik, erabiltzaile mailara igarotzeko honako pausu hauek emango dira: (i) txioak banan banan sailkatuko dira informal-formal zein gazte-heldu sailkatzaileak erabilita; (ii) erabiltzailearen txio guztien etiketak kontutan hartuta erabiltzailea gazte edo heldu bezala etiketatuko da; (iii) lortutako erabiltzaile mailako emaitzak aztertuko dira metodo ezberdinekin lortutako sailkapenak konparatzeko asmoarekin.

Erabiltzaile bakoitzaren txioak sailkatzeko, aurrez garatutako informal-formal (3.3.2. atala) eta gazte-heldu (3.3.3. atala) testu-sekuentzia sailkatzaileak erabiliko dira. Hots, *heldugazte* datu-multzoarekin entrenatutako *IXA pipes* sailkatzailea (informal-formal) eta *heldugazte-age* datu-multzoarekin birdoitutako BERTeus sailkatzailea (gazte-heldu) erabiliko dira erabiltzaileen txioak banan banan automatikoki etiketatzeko. Horretarako, gutxienez euskaraz idatzitako 10 txio pertsonal dauzkaten erabiltzaileen denbora-lerroak erabili dira. Horrela, *heldugazte-osoa* corpusetik arakatu ditugun 7.980 erabiltzaileetatik, 7.087 erabiltzaile aukeratu dira.

Erabiltzaile bakoitzaren txioak banaka etiketatu ostean txioak informal/formal edo gazte/heldu bezala etiketatuta egongo dira. Txio mailatik erabiltzaile mailara etiketak proiektatzeko etiketen kontzentrazioan oinarrituko gara. Erabiltzaile mailako etiketatze automatikoa txioen etiketen kontzentrazioan oinarriturik egingo denez, kontzentrazioaren balioa ezarri beharko da atalase moduan. Era honetan, denbora-lerro zehatz batean txioen % 60a *informal* edo *gazte* bezala etiketatuta badago, erabiltzailea gaztetzat joko dugu. Bestalde, denbora-lerro baten txioen % 40a soilik *informal* edo *gazte* bezala etiketatuta badago, erabiltzailea heldutzat joko dugu. Horrela, atalase finko bat zehaztu ordez, bi tarte aukeratu dira, denbora-lerro bakoitzean *gazte* edo *informal* gisa etiketatutako txioen % 60 eta % 40ean kokatuak. Ziurgabetasun tarteko (% 40 - % 60 artean) kontzentrazioan geratzen diren denbora-lerroak, "indeterminatu" bezala sailkatuko dira, hau da, ez dugula nahikoa froga erabiltzailearen bizitza-etapa erabakitzeko. Beraz, erabiltzaile *indeterminatu* hauek, gazte/heldu etiketarik gabe geratuko dira. Horrela, "indeterminatua", kategoria sintetiko berria sortu dugu, zeregin bitar bat hirutar bihurtuz. *Indeterminatu* klasea gehitzeak zalantzazko kasu zailak gazte edo heldu gisa sailkatzeko konpromisoa saihesteko onura du. Honekin estaldura galtzen da doitasunaren mesedetan, lortutako emaitzak fidagarriagoak izanik.

3.7. taulak *informal-formal* eta *gazte-heldu* metodoak erabilita, *gazte*, *heldu* edo *indeterminatu* gisa sailkatutako erabiltzaile kopurua erakusten du. Ikusten

Sailkatzaileak	Heldu	Indeterminatu	Gazte
Informal-formal (3.3.2)	5.213	911	963
Gazte-heldu (3.3.3)	4.472	980	1.635

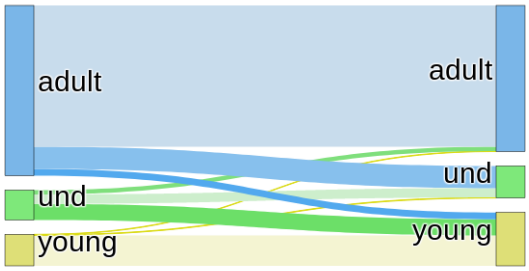
3.7 Taula – Erabiltzaileen sailkapena adin tarteen arabera.

denez, desberdintasun nagusia metodo bakoitzak lortutako erabiltzaile gazteen kopuruari dagokio. Hurrengo azpiatalean desberdintasunak sakonago aztertuko ditugu.

Metodoen arteko konparaketa

Informal-formal (3.3.2. atala) eta gazte-heldu (3.3.3. atala) hurbilpenek sailkatze automatikoa burutzerako orduan, emaitzetan ezberdintasun nabarmenak agertu dituzte. Hasteko, sailkatutako erabiltzaileen artean % 21.79ko ezberdintasuna erakutsi da. Gainera, aldaketei azaleko begirada bat emanez (3.2. irudia), aldakuntza esanguratsua ikus daiteke gazte gisa etiketatutako erabiltzaileen artean. Ikus daitekeenagatik, *informal-formal* sailkatzaileak joera dauka erabiltzaileak heldu bezala etiketatzeko, *gazte-heldu* sailkatzaileak baino nabarmen gehiago etiketatu baititu. Bereziki deigarria da *gazte-heldu* sailkatzaileak gazte gehiago sailkatu dituela, *informal-formal* sailkatzaileak baino. Azaleko azterketa hau burutu ostean, aldaketak non eta nola eman diren aztertuko da, konparaketa matrize eta adibideen analisi sakon bat eginez.

Aldakortasunak zehatzago aztertzeko, bi hurbilpenen iragarpenekin matrize bat irudikatu da, 3.8. taulak erakusten duen bezala. Matrizearen diagonalaren beheko partean ematen dira aldaketa gehienak, gazte-heldu sailkatzaileak gazte gehiago identifikatzeko daukan joera erakutsiz. Ikus daitezkeen bezala, hiru aldaera garrantzitsuenetako bitan (ind/gaz eta hel/gaz), gazte-heldu metodoak klase gaztea hautatzen du. Bestalde, aldakortasun handiena duen multzoan, heldu-gazte metodoak *indeterminatu* klasea aukeratzen du, informal-formal metodoak *heldu* iragartzen duen bitartean. Aipatutako kasu zehatz hauek sakonago aztertu dira, horretarako eskuzko analisi bat burutuz denbora-lerro batzuk ausaz aukeratuta. Honela, erabiltzaile berdinarentzat etiketa ezberdinak dauden kasuetan sailkatzaile egokia zein den erabakitzeko. Konparazio honen helburua sailkapen metodo ezberdinen arteko kategoria arteko aldakortasuna aztertzea da (helduetatik gazteetara, adibidez). Horrela, aldaketa multzo bakoitzerako kasuen % 10eko ausazko lagina aztertu da.



3.2 Irudia – *Informal-formal* (ezkerrean) eta *gazte-heldu* (eskubian) hurbilpenen Sanky diagrama. Heldu (adult), indeterminatu (und) eta gazte (young) klaseen arteko aldakortasunak.

		GZT-HLD			
		gaz	ind	hel	*tot*
INF-FOR	gaz	**933**	16	14	*963*
	ind	493*	**285**	133	*911*
	hel	209†	679‡	**4.325**	*5.213*
	tot	*1.635*	*980*	*4.472*	*7.087*

3.8 Taula – Informal-formal (INF-FOR) eta gazte-heldu (GZT-HLD) hurbilpenen iragarpenen arteko aldakortasuna gazte (gaz), indeterminatu (ind) eta heldu (hel) klaseetarako. Aldakortasun esanguratsuenak INF-FOR eta GZT-HLD artean: (‡) hel/ind; (*) ind/gaz; (†) hel/gaz.

Ezberdin sailkatutako erabiltzaileen eskuzko azterketak erakutsi du gazte-heldu sailkatzaileak emaitza hobeak lortzen dituela informal-formalekin alderatuta. Jarraian ereduen emaitzen eskuzko azterketaren 3 adibide ikusiko ditugu, bakoitza aldakuntza esanguratsu batetako erabiltzailea izanik. @erab1 eta @erab2-ri dagokienez, gazte-heldu metodoak erabiltzaileak *gazte* bezala sailkatu dituela erakusten du, informal-formal hurbilpenak *heldu* eta *indeterminatu* bezala sailkatu dituen bitartean. Euren txioak ikusita, badirudi erabiltzaileak gazteak direla idazkera estiloan oinarrituta, baina baita azterketei buruz hitz egiten dutelako ere, oro har gazteei lotutako jarduera. Bi adibide hauek erakusten dute informal-formal

metodoak ez daukala gazteak identifikatzeko ahalmen handia. @user3-ren kasua polemikoagoa da, zailagoa ematen baitu eskuragarri dagoen edukian oinarrituta erabiltzailearen bizi-etapa zehaztea, beraz, badirudi *indeterminatu* sailkapen-etiketa egokia dela. Laburbilduz, emaitzen analisi kualitatiboa burutu eta gero esan beharra dago gazte-heldu sailkatzailearen emaitzak nabarmen hobeak direla.

- @*erab1*-en etiketa aldaketa[†]: heldu (inf-for) / gazte (gzt-hld):

  (1a) Horrelakoekin gustua ta guzti hartzen zaio ikasteari.
  With this, you take pleasure in learning.

  (1b) Buenobueno ba ikasiko dut gehio jaja ta ikusikozu gaindituko dutt jaja.
  Weeeell weeeell, I'll learn more haha and you'll see if I can pass the exam haha.

- @*erab2*-en etiketa aldaketa[*]: indeterminatu (inf-for) / gazte (gzt-hld):

  (2a) Ze txupi txatxi no me da la nota.
  Awesome I don't get to pass...

  (2b) Ai naiz rayatzen pixkat asko con la mierda de la uni.
  Oh I'm going crazy a little bit with university shit.

- @*erab3*-en etiketa aldaketa[‡]: heldu (inf-for) / indeterminatu (gzt-hld):

  (3a) A zer nolako eguraldi kaxkarra ez al du gelditu behar edo.
  What a bad weather, shouldn't stop or what.

  (3b) Gu erakusteko prest, etorri daitezela lasai eskuzabalik hartuko ditugu eta.
  We are ready to show it, we will wait for them with open arms.

3.2. irudiak eta 3.8. taulak erakusten dute *informal-formal* eta *gazte-heldu* metodoen sailkapenen arteko ezberdintasunak. Azaleko azterketa kuantitatiboaren arabera, *gazte-heldu* metodoaren emaitza orekatuagoak ikusi daitezke. Ezberdin sailkatutako erabiltzaileak eskuz aztertu ostean, ikusi da *gazte-heldu* metodoak, *informal-formal* metodoak baino egokiago burutzen duela ataza. Ebaluaketa kualitatibo honekin, konfirmatzen da *gazte-heldu* metodoak (3.3.3) emandako sailkapen emaitzak *informal-formal* metodoak (3.3.2) emandakoak baino egokiagoak dira. Ebaluaketa kuantitatibo eta kualitatiboaren arabera, badirudi *gazte-heldu* metodoa (3.3.3) egokiena dela *Heldugazte-osoa* corpuseko erabiltzaileak automatikoki sailkatzeko *gazte/indeterminatu/heldu* artean

3.4 Elkarrekintza sarea: gainbegiratu gabeko aplikazioa

Atal honetan sareko euskal gazte erabiltzaileen artean gertatzen diren harremanak aztertuko dira. Abiapuntua aurreko atalean *gazte* gisa sailkatutako 1.635 erabiltzaileek egindako euskarazko birtxioak izango dira. Horrela, identifikatutako euskal erabiltzaile gazteek beren denbora-lerroan konpartitutako edukiak jasoko dira. Birtxioak aukeratu dira erabiltzaileen arteko interakzio ekintzak direlako eta aipamenak bezalako elkarrekintzek baino hobeto erakusten dutelako erabiltzaile arteko korrelazioa (Conover *et al.* 2011b). Azterketa egiteko, birtxio bakoitzetik ateratako bi ezaugarri erabiliko dira: (i) publikazioa konpartitzen duen erabiltzailea edo birtxiokatzailea (iturburu) eta (ii) publikazioa sortu duen erabiltzailea (helburu) edo birtxiotua. Aukeratutako ezaugarri hauek erabiltzaileen arteko harreman edo erlazio bat zehazten dute. Hau da, erabiltzaile zehatzen interakzio konkretuak kontutan hartuta, lagin guztiaren harremantze dinamikak azaleratzea izango da asmoa.

Zehazki, 1.635 erabiltzaile gazteren 418.903 birtxioetatik 24.837 nodo eta 148.304 konexio atera dira. Nodoak birtxioak egiten dituzten erabiltzaileei dagozkie (gure 1.635 erabiltzaileko lagina) baina baita hauek jasotzen dituzten erabiltzaile ezberdinei ere (gure laginekoak izan edo ez). Bestalde, konexioek adierazten dute iturburu-erabiltzaile batek beste helburu-erabiltzaile bat behin edo gehiagotan birtxiokatu duen.

Interakzioetan oinarritutako datu hauekin erabiltzaileen harremanak zeintzuk diren ikertuko ditugu, horretarako konparaketa eta komunitateen sakoneko azterketa eginez. Lehenik eta behin, gazteengandik interakzio gehien jasotzen dituzten erabiltzaileak identifikatu dira, gazteen erreferenteak azalaraziz. Bigarren urrats batean euskal erabiltzaileen komunitate inplizituak identifikatu dira, horretarako interakzioekin erabiltzaile eredu bat sortuz eta azpitaldeen arabera banatuz.

3.4.1 Euskal erabiltzaile gazteen erreferente euskaldunak

Azpiatal honetan euskal gazteek konpartitutako euskarazko edukiaren azterketa egingo da. Zehazki euskal gazteek konpartitutako euskarazko publikazioen egile errepikatuenak zeintzuk diren aztertuko da. Horrela, gehien konpartituak izan diren euskal erabiltzaileak identifikatuta, euskal gazteen artean arrakasta gehien daukaten egileak topatuko ditugu. Horretarako, 3.9. taulan erabiltzaile erreferentzialak zeintzuk diren ikus ditzakegu, bi era ezberdinetan antolatuta.

Alde batetik, gazteek konpartitutako euskarazko publikazioen erabiltzaileak daude (3.9a. taula), eduki arrakastatsuena zein erabiltzailerena den ezagutzeko. Sailkapen honetan erabiltzaile bakoitzak behin baino gehiagotan konpartitu ahalko du erabiltzaile zehatz bat, hau da, erabiltzaile bakoitzaren eduki ezberdinak jasotzen ditu. Bestetik, erabiltzaile zehatz bat zenbat erabiltzaile gaztek konpartitu duten adierazten duen taula dugu (3.9b. taula), erabiltzaile hedatuena zein izan den jakiteko. Horrela, erabiltzaile erreferentzial hauen ezaugarriak aztertuta, erabiltzaile gazteen artean gehien mugitzen diren edukien nolakotasuna iradoki ahalko da.

Erabiltzailea	Aldiak
@berria‡	8671
@argia‡	5646
@ernaigazte*	4553
@topatu_eus‡	4236
@enekogara	3274
@naiz_info‡	3262
@ZuriHidalgo	2568
@AskeGunea*	2561
@RealSociedadEUS	2531
@larbelaitz‡	2471
@ArnaldoOtegi*	2188
@iBROKI‡	2031
@LeakoHitza‡	1893
@athletic_eus	1818
@euskaltelebista‡	1744

(a) Zenbat aldiz konpartitua.

Erabiltzailea	Erabiltzaileak
@berria‡	998
@argia‡	844
@naiz_info‡	710
@larbelaitz‡	585
@topatu_eus‡	531
@ArnaldoOtegi*	518
@ernaigazte*	478
@enekogara	454
@HamaikaTb‡	442
@jpermach*	427
@axierL‡	413
@ielortza	407
@MaddalenIriarte‡*	404
@boligorria‡	398
@GureEskuDago*	394

(b) Zenbat erabiltzailek konpartitua.

3.9 Taula – Euskal erabiltzaile gazteen erreferenteak. Hedabideekin lotutako erabiltzaileak (‡). Politikarekin lotutako erabiltzaileak (*).

3.9. taulan ikusi daitekeen moduan, erabiltzaile arrakastatsuenen artean euskal hedabideekin lotutako kontuak daudela ikus dezakegu: @berria (Berria - euskarazko egunkaria), @argia (Argia - euskarazko aldizkaria), @naiz_info (Naiz - informazio orokorreko euskal webgunea), @topatu_eus (Topatu! - gazteei zuzendutako euskarazko hedabide digitala), @HamaikaTb (Hamaika telebista - euskarazko telebista katea), @euskaltelebista (ETB - Euskal telebista kate publikoa) and @LeakoHitza (Lea-Artibai eta Mutrikuko Hitza - Komunikazio proiektu lokala).

Horrez gain, hedabide hauetan kazetari lanetan ari diren norbanakoak ere aurkitu ditzakegu: @larbelaitz, @axierL, @boligorria (Argiako kazetariak); @MaddalenIriarte eta @iBROKI (kazetariak Euskal Telebistan). Beraz, erabiltzailerik garrantzitsuenak euskal hedabideei dagozkie, Twitterren izaera komunikatiboa agerian utziz.

Bestalde, politikarekin erlazionatutako erabiltzaileak ere ageri dira erreferenteen artean. Horrela, Ezker Abertzaleko norbanako (@ArnaldoOtegi, @jpermach) eta erakundeen (@ernaigazte) erreferentzialtasuna nabarmena da ere gazteen artean. Euskal Herriko mugimendu zibil eta politikoekin lotutako erakundeak ere azaltzen dira erreferenteen artean: @AskeGunea (Aske Gunea, - babesa agertzeko desobedientzia zibileko mugimendua) eta @GureEskuDago (Gure Esku Dago - erabakitze eskubidearen aldeko herri mugimendua). Ikusi den bezala, politikarekin erlazionatutako erabiltzaileek ere arrakasta daukate gazteen artean, ezker independentistarekin lotutako joera politikoa nabarmenduz.

Gazteen erreferenteen analisia burutu ostean, esan beharra dago horietako bi baino ez direla gazteei lotutako kontuak, @ernaigazte eta @topatu_eus gazteek osatutako erakundeei lotutako erabiltzaileak dira biak. Alde batetik, @ernaigazte Ernai gazte antolakundearekin erlazionatutako erabiltzailea da, ezker abertzaleko gazteak ordezkatzen dituena. Bestalde, @topatu_eus gazteei lotutako komunikabide digitaleko erabiltzailea da, aldi berean ezker abertzalearekin erlazionatutakoa. Gazteen gazte erreferente hauek ere, hedabideekin eta politikarekin estuki erlazionatuta daude, gazteen artean ere Twitterren ematen den joera errepikatuz. Erabiltzaile gazteen artean erreferente gazteen gabezia hori, Twitterren ezaugarriekin lotuta egon liteke, gehienbat gai politiko edota berrien inguruan aritzeko sarea baita.

3.4.2 Euskal erabiltzaile gazteen azpi-komunitateak

Erabiltzaileen harremanak sakonean aztertzeko interakzioak baliatuko ditugu, erabiltzaileak harremanen arabera kokatu eta antolatzeko. Antolaketa horretarako gainbegiratu gabeko metodoak erabiliko dira, interakzioak baliatuta erabiltzaileen errepresentazio dentsoak sortzeko. Errepresentazio horiek lortzeko hurbilpen ezagun eta eraginkorrenen artean, DeepWalk (Perozzi *et al.* 2014) eta node2vec (Grover and Leskovec 2016) dauzkagu 2.1.4. atalean ikusi bezala. Metodo hauek erabiltzaileak ordezkatzen dituzten dimentsio baxuko ezaugarriak sortzen dituzte, horretarako etiketatu gabeko datu kopuru handiak baliatuz. Node2vec (N2V) algoritmoak sarearen egitura kontrolatzeko aukera ematen duenez, erabiltzaileen errepresentazioak ikasteko erabiliko da.

Beraz, euskal erabiltzaileak errepresentatuko dituen eredua sortzeko, erabiltzaile gazteek konpartitutako edukiak erabili dira, hots, birtxioak. Eredua sortzeko birtxiokatzaile-birtxiotu pareak erabili dira sarrera datu bezala. Horrela, N2V gure datuetan aplikatzeko, hiperparametro balio lehenetsiak ezarri dira: *walks per node = 10* (ibilaldi kopurua), *walk length = 80* (ibilaldi luzera), *window or context size = 10* (testuinguru tamaina), eta optimizazioa *epoc* bakarrean exekutatzen da (Perozzi *et al.* 2014; Grover and Leskovec 2016). Bestalde, itzulera eta sarrera-irteera parametroak, azpi-komunitateen inguruko informazio zehatzagoa lortzeko finkatu ditugu $p = 1$ eta $q = 0,5$ balioak aukeratuz (Grover and Leskovec 2016). Argibide gehiagorako ikusi 2.1.4. atala.

Euskal erabiltzaileek konpartitutako edukia erabilita N2V eredu bat entrenatu da, interakzioetan oinarriturik, erabiltzaile bakoitza dimentsio anitzeko espazioko puntu batean kokatuz. N2V eredua sortu ondoren azpitaldetan banatu dugu, euskal erabiltzaile gazteen baitako azpi-komunitateak edo azpitaldeak nola eratzen diren aztertzeko. Horretarako, lortutako eredua lau cluster ezberdinetan zatitu da, azpitalde kopurua inertzia balioei erreparatuta aukeratuz. Cluster bidezko zatiketak, modularitatean oinarritutako algoritmoak (Blondel *et al.* 2008) ez bezala, atera beharreko komunitate kopuru zehatza hautatzeko aukera ematen du. Ereduaren bistaraketarako erabilitako 3.3. irudiak erakusten du N2V metodoak argi eta garbi bereizten diren komunitateak sortzen dituela, eta horrek, aldi berean, interpretagarriak egiten ditu, erabiltzaileen artean dauden harremanak ulertzea erraztuz.

Erabiltzaileak irudikatzen dituen N2V errepresentazioa lau komunitatetan banatu ondoren, azpitalde bakoitzaren ezaugarri nagusiak ondorioztatu dira. Prozesu honetarako, komunitate bakoitza ordezkatzen duten nodo garrantzitsuenetan oinarritu gara, hau da, gehien konpartituak izan diren erabiltzaileetan jarriko dugu fokua. Gerora, erreferentzialak diren erabiltzaile hauen nolakotasuna aztertuko da modu kualitatiboan, gaien arabera ordenatuz. Erabiltzaileak aztertuta, azpitalde bakoitzari ezaugarri orokor bat esleituko zaio, komunitatearen identitatea markatuko duena. Gai horiek desberdinak dira grafikoko azpitalde bakoitzean, komunitate bakoitzaren ezaugarriak edo desberdintasunak erakutsiz. Jarraian, grafikoan jasotako lau azpitaldeetako bakoitzaren ezaugarri nagusiak deskribatuko ditugu.

- *Albisteak* (% 29,96): Komunitate honetan, sailkapenaren buruan aurkitzen diren nodoak aurkitzen dira, Euskal Herriko komunikabide eta gaurkotasunarekin zuzenean erlazionatutako erabiltzaileek osatuta. Horrela Euskal Herriko komunikabideak (@berria, @argia, @HamaikaTb, @eitbAlbisteak, @euskaltelebista, @zuzeu, @euskadi_irratia, @Gaztezulo, @Sustatu, @eitbeus ...) eta euskal kazetariak (@MaddalenIriarte, @boligorria, @ur-

tziurkizu, @zaldieroa, @bzarrabeitia, @AneIrazabal ...) dira komunitate honetako nodo erreferenteak. Horrez gain, euskara edo Euskal Herriarekin erlazionatutako edukia sortzen duten erabiltzaile aktiboak ere badira (@ielortza, @kalaportu, @KikeAmonarriz, @maia_jon).

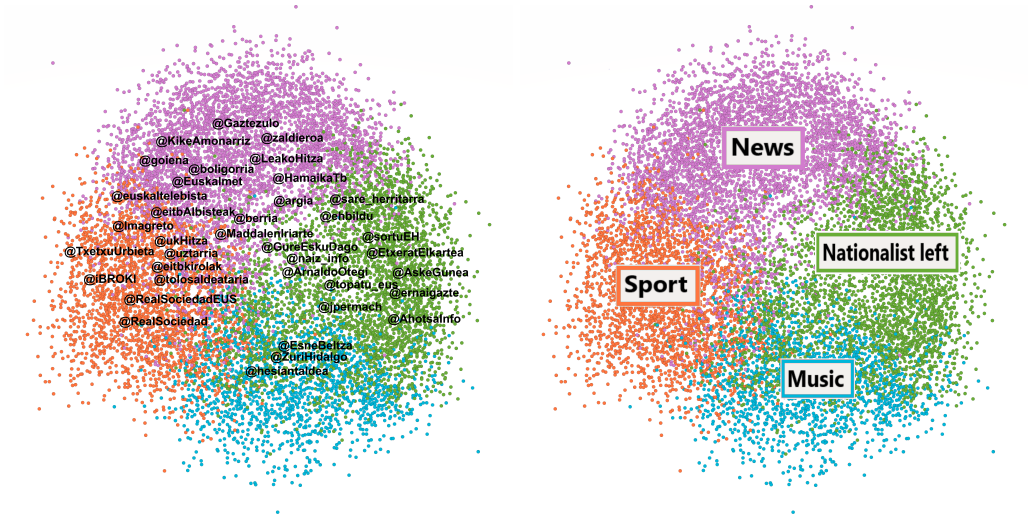
- *Ezker abertzalea* (% 26,98): Azpitalde zehatz honek Ezker Abertzale edo independentistarekin erlazionatutako erabiltzaileez osatuta dago. Erabiltzaileak erakunde politiko eta sozialekin (@ernaigazte, @GureEskuDago, @AskeGunea, @ehbildu, @sortuEH, @EtxeratElkartea ...) zein mugimendu politiko honetako pertsona erreferenteekin (@ArnaldoOtegi, @jpermach ...) erlazionatu ditzakegu. Azpitalde tematiko honetan ere komunikabideen presentzia edukiko genuke, aukera politiko honekin erlazionatutakoak (@naiz_info, @topatu_eus, @info7irratia, @AhotsaInfo ...).

- *Kirolak* (% 22,58): Kirol azpitaldean nodo garrantzitsuenak pertsona kazetariak (@iBROKI, @XabierEuzkitze, @Imagreto, @TxetxuUrbieta, @jontolest, @unaizubeldia ...) edo albistegiak (@eitbkirolak, @ukHitza, @3ErregeenMahaia ...) dira, bereziki kirol arloan espezializatuta daudenak. Azpitalde zehatz honetan ere, erabiltzaile konpartituenak egunkari eta telebista kateei egiten diete erreferentzia, beste behin ere komunikabideekin zuzenean erlazionatuta egonik. Komunitate honetako beste nodo garrantzitsu batzuk kirol taldeekin lotutakoak dira, hala nola futbol taldeak (@RealSociedad, @RealSociedadEUS, @SDEibar, @AthleticClub ...) edo jokalariak (@InigoMartinez, @mikelsanjo6, @ilarra4 ...), txirrindulari ezagunak (@AmetsTxurruka, @mikelastarloza, @Markelirizar ...) zein euskupilota enpresak (@ASPEpelota ...).

- *Musika* (% 20,49): Musika azpitaldean leku nabarmenetan agertzen dira euskaraz abesten duten musika taldeak edo abeslariak (@ZuriHidalgo, @vendettaska, @hesian-taldea, @EsneBeltza, @gatibu, @ZeEsatek ...) , nahiz eta musikarekin lotutako beste kontu batzuk ere oso aktibo daudela dirudien (@GustokoMusika, @euskalkantak5, @KantuBatGara ...).

Aztertutako azpitaldeek erakusten dute guztiek harreman zuzena dutela Euskal Herriarekin lotutako gai edo kontuekin. Hala, ikusten da euskara erabiltzen dela Euskal Herriko gaurkotasuna (albisteak) eta politikarekin (ezker abertzalea) erlazionatutako edukiak partekatzeko. Gainera, ikus daiteke aisialdiarekin erlazionatutako euskal musika eta kirol edukiak ere asko konpartitzen direla gazteen artean. Hau da, badirudi Twitterreko elkarrekintzen asmo nagusia politika eta

gizarte gaiei buruzko edukiak partekatzea dela, baina euskal komunitateari eta hizkuntzari arreta argia emanez.



(a) Nodo garrantzitsuak. **(b)** Gaiak

3.3 Irudia – Erabiltzaile gazteen sarea komunitateen arabera zatikatua.

3.3. irudiak erakusten du euskal erabiltzaile gazteen errepresentazioaren bistaratzea, erabiltzaileak puntuen bidez adierazita daude eta bakoitzaren kolorea azpi-atalei dagokie. Ikusi daiteke, gazteak oro har, gizarte gaiekin (politika eta albisteak) zein aisialdiarekin (musika eta kirola) zerikusia duten gaien inguruan erlazionatzen direla. Komunitateak detektatzeko aplikatutako metodologia dela eta, azpitaldeak modu koherentean mapatzeko gai izan gara, komunitate bakoitza gaien arabera antolatuz. Hau da, komunitate bakoitzak grafikoan duen posizioak eta komunitateen arteko hurbiltasunak erakusten dute gaiek haien artean zein erlazio duten. Modu honetan gizarte gaiekin (politika eta albisteak) lotutako komunitateak elkarren ondoan daudela ikus dezakegu, aisialdiarekin (musika eta kirola) lotutako komunitateekin gauza bera gertatzen den bitartean. Politikarekin (Ezker abertzalea) lotutako komunitatea albisteetatik eta musikatik gertu dago, sare sozialetatik mugitzen diren euskarazko albiste zein musika talde batzuen jarrera politikoa erakutsiz. Bestalde, kirolarekin erlazionatutako azpitaldea politikarekiko urrutien dagoena da, musika eta albisteekin gertatzen den moduan. Gainera, lau azpitaldeetatik hirutan (Albisteak, Ezker Abertzalea eta Kirolak) hedabideak eta kazetariak dira erabiltzaile erreferenteak, berriro ere frogatuz komunikabideak garrantzitsuak direla gazteen artean euskarazko edukiak zabaltzeko.

3.5 Ondorioak

Atal honetan, sare sozialetara konektatuta dauden euskal hiztun gazteen errealitatea ezagutzen lagunduko hurbilpen metodologiko bat proposatu da. Horretarako, ikasketa automatikoan oinarritutako teknika aurreratuak garatu, ebaluatu eta aplikatu dira eskala handiko datuetatik abiatuta, ezaugarri demografikoak iradoki eta komunitate azterketak burutzeko. Lehenik eta behin, ia 8.000 euskal erabiltzaileren 6 milioi publikazio baino gehiagoz osatutako corpus bat eskuratu eta publikatu da. Bigarrenik, Twitter sare sozialeko euskal erabiltzaileak sailkatu dira gazte eta heldu artean, Hizkuntzaren Prozesamendua oinarri duen metodologia berri bat proposatuta. Hirugarrenik eta azkenik, gazteen komunitateak zeintzuk diren ikusi da, konpartitzen duten edukietatik erlazionatzeko moduak erauziz. Lan honekin, frogatuta geratzen da gizarte-zientzia eta konputazio-zientzien arteko konbinaketa aberasgarria dela. Gainera, Hizkuntzaren Prozesamendurako eta komunitateak detektatzeko ikasketa sakoneko teknika aurreratuekin esperimentatu dugu.

Ikerketaren oinarri izango diren datuak lortzeko, sare sozialetatik informazioa lortzeko moduak aztertu eta aplikatu dira. Horrela, bildutako datu kantitatea ia 8K erabiltzaile euskaldunen 6 milioi publikazio baino gehiagok osatzen dute. Sare sozialak euskara bezalako hizkuntza gutxituentzat ere datu-iturri garrantzitsua direla frogatu da. Gainera, erakutsi dugu, Twitter datu-iturri baliotsua dela testu zein elkarrekintzak islatzen dituzten datuak jasotzeko, ia 40 milioi hitzez osatutako testu corpusa eta ia 3 milioi interakzio pare lortuz. Lortutako datu kopuruari esker eskala handiko azterketa sozial zein linguistikoa egitea ahalbidetu da.

Erabiltzaileen ezaugarri sozialak iragartzeko asmoarekin, metodologia berri bat aurkeztu dugu euskarazko testu kantitate erraldoiak prozesatuaz. Horretarako, ikasketa automatiko eta sakoneko hurbilpenak aplikatu ditugu Hizkuntzaren Prozesamendu bitartez testu sailkapena egin eta erabiltzaileen bizitza-etapa heldu eta gazte artean sailkatzeko. Erabiltzaileen adina iragartze aldera soziolinguistika erabili da, erabiltzaileen idazkera estiloa aintzat hartuz. Gainera, bi hurbilketa sekuentzialen konparaketa aurkezten dugu, ezaugarri linguistikoetan oinarritutakoa eta erabiltzaileen ezaugarrietan oinarritutakoa. Lehenengoari dagokionez, eskuz 1.000 txio etiketatu dira *formal* eta *informal* etiketarekin. Bigarrenerako, erabiltzaile mailako etiketatze erdi-automatikoa egin da, *gazte* edo *heldu* gisa zuzenean etiketatutako 80.000 txio lortuz eta aurreko datu-multzoaren kalitatea eta tamaina hobetuz. Testu sailkapenerako ikasketa automatikoko teknika modernoekin esperimentatu da, Transformer oinarritutako modeloak ere erabiliaz. Horri esker, bizitza-etaparen arabera etiketatutako 80K txioko datu-multzo berri bat sortu du-

gu, hots, *Heldugazte-age* datu-multzoa. Gainera, esperimentuetan lortutako emaitzek adierazi dute metodo berriak kalitate oneko etiketa-datuak sortzen dituela gazte/helduen sailkatzaileak trebatzeko. Horrela, etiketatze esfortzu txiki bat eta erabiltzaile-mailako etiketatze erdi-automatiko konbinatuta, datu-multzo esanguratsuak sortzeko metodo eraginkorra dela frogatu dugu.

Horrez gain, ikusi da ikasketa sakoneko metodo ez-gainbegiratuen aplikazioa eraginkorra dela erabiltzaileen errepresentazio dentsoak sortu eta bertatik komunitateak zeintzuk diren iragartzeko. Erabilitako sarrera datuek, hau da, edukia konpartitzeko ekintzek, informazio sozial eta politikoa inplizitua barneratzen dutela konfirmatu da. Gainera, erabilitako erabiltzaileen errepresentazio dentsoak, azpitaldeen araberako zatiketa ahalbidetzeaz gain, azpitalde hauen arteko erlazioa zein den erakusteko gai da ere. Gainera, azpitaldeetako erabiltzaile arrakastatsuenak baliatuta, komunitate ezberdinen identitatea modu intuitiboan definitzea lortu da. Horrela, erabiltzaileen harremantze-ekintzetan oinarritutako datuak bistaratu eta interpretatu daitezke teknika hauek erabiliz, egituratu gabeko informazioa ezagutza bihurtuz inolako anotazio esfortzurik gabe.

Azkenik, sortutako corpus eta datu-multzo guztiak publikoki eskuragarri jarri dira, euskara bezalako baliabide urriko hizkuntzen ikerketa bultzatu eta errazteko. Gainera, lan honetan aurkezten diren metodologia zein datuak Hizkuntzaren Prozesamendurako beste zeregin batzuetarako eta ikerketa soziala burutzeko baliagarriak izan daitezkeela uste dugu.

Era berean, ikerketatik bertatik ateratako ondorioetatik abiatuta, etorkizuneko lanetarako aplikagarriak izango diren hainbat aurkikuntza interesgarri lortu dira. Batetik, erabiltzaileen testua eta eskuzko anotazioak erabilita, erabiltzaileen ezaugarriak iradokitzeko ereduak garatu daitezkeela ikusi da. Bestetik, interakzioetan oinarritutako datuen baliagarritasuna hauetatik erabiltzaileen informazio sozial zein politiko potentziala erauzteko. Etorkizuneko lan moduan, Hizkuntzaren Prozesamendua eta interakzioetan oinarritutako datuak bestelako atazetan erabiltzeari ekingo diogu, hala nola, jarrera edo ideologia politikoekin lotutako atazak.

4. CHAPTER

Social Features for Language Independent Stance Detection

The large majority of the research performed on stance detection has been focused on developing more or less sophisticated text classification systems, even when many benchmarks are based on social network data such as Twitter (Mohammad *et al.* 2016; Taulé *et al.* 2018; Zotova *et al.* 2021). This chapter aims to take on the stance detection task by placing the emphasis not so much on the text itself but on the interaction data available on social networks. For a more in-depth exploration of this issue, we create a dataset to detect stance in Tweets referring to vaccines, a relevant and controversial topic during the Covid-19 pandemia. The dataset is proposed in a multilingual setting, providing data for Basque and Spanish languages. The objective is to explore crosslingual approaches which also complement textual information with social features obtained from the social network. Additionally, we propose a new method to leverage social information such as *friends* and *retweets* by generating *Relational Embeddings*, namely, dense vector representations of interaction pairs. We empirically demonstrate that our method can be applied to any language and target without any manual tuning. Experiments on seven publicly available datasets and four different languages show that combining our relational embeddings with textual methods helps to substantially improve performance, obtaining state-of-the-art results for six out of seven evaluation settings, outperforming strong baselines based on large pre-trained language models, or other popular interaction-based approaches namely DeepWalk or node2vec.

4.1 Motivation and Contributions

Stance detection consists of identifying the viewpoint or attitude expressed by a piece of text with respect to a given target (Mohammad *et al.* 2016). With the enormous popularity of social networks, users spontaneously share their opinions on social media, generating a valuable resource to investigate stance. This means that research on stance has a social impact, for example, to help addressing misinformation on vaccines, or to better understand public opinion about topics such as climate change or migration. Furthermore, stance detection is considered an important intermediate task for fact-checking (Augenstein 2021) or fake news detection (Shu *et al.* 2017).

The SemEval 2016 task on stance detection in Twitter (Mohammad *et al.* 2016) presented a dataset with tweets expressing FAVOR, AGAINST and NEUTRAL stances with respect to five different targets, a trend followed by many other researchers (Derczynski *et al.* 2017; Taulé *et al.* 2018; Zotova *et al.* 2021; Hardalov *et al.* 2022). However, despite many of them using Twitter-based datasets, the large majority address the task by considering only the textual content of tweets (Augenstein *et al.* 2016; Schiller *et al.* 2021; Hardalov *et al.* 2021; Li *et al.* 2021; Ghosh *et al.* 2019; Küçük and Can 2020; Sobhani *et al.* 2017; Glandt *et al.* 2021).

However, an interesting new dataset for Italian was released in 2020 as part of the SardiStance@Evalita 2020 shared task (Cignarella *et al.* 2020), which included not only the texts of the tweets labeled with stance, but also social network information relative to the authors of the tweets. This social network information includes retweets, user accounts profile, friends and followers, among others. In this context, we propose the VaxxStance dataset at IberLEF 2021 (Montes *et al.* 2021), with the aim of detecting stance in social media on vaccines in general. The dataset provides data in two languages, Basque and Spanish, and its objective is to promote crosslingual research on stance detection using both the text and the information provided by the Twitter social network. Thus, and unlike previous approaches, we provide, for a given topic, multilingual coetanous data of gold-standard quality in a corpus which allows to experiment using both social and textual features in multilingual and crosslingual settings.

Although these new datasets have facilitated the development of new techniques for stance detection considering also interaction data, most of them employ manually engineered features tailored to each specific data type (Espinosa *et al.* 2020; Lai *et al.* 2021; Alkhalifa and Zubiaga 2020), making it difficult to

generalize across languages and targets. Thus, further research is required to fully understand the potentiality of interaction data to perform stance detection and its relation with concepts such as political homophily, political polarization, echo chambers or demographic analysis (Conover *et al.* 2011b; Colleoni *et al.* 2014; Zubiaga *et al.* 2019).

To address this problem, we propose a new methodology to enhance the stance detection of social media content, named Relational Embeddigs. Our approach focuses on both innovative data collection and novel user representation algorithms, leveraging textual and interaction-based data. By integrating these components, we aim to improve the accuracy and robustness of stance detection systems across various languages and targets, thereby contributing to more reliable analysis of social media data. This chapter presents a detailed exploration of our methodology, highlighting its potential to significantly advance the field of computational social science by enabling deeper insights into public opinion dynamics and discourse analysis.

This chapter focuses on stance detection of tweets by placing the emphasis on the interaction data commonly available in social media. To this end, we make the following contributions: (i) a new public dataset for stance detection in two languages containing text and interaction data; (ii) a novel method to represent and exploit interaction data, such as *friends* and/or *retweets*, by generating Relational Embeddings based on one-to-one relations; (iii) comprehensive experiments on seven publicly available datasets and four different languages show that our relational embeddings behave robustly across different targets and languages without any specific manual engineering; (iv) combining our method with text-based classifiers helps to systematically improve their results, outperforming also ensembles of large pre-trained language models (Giorgioni *et al.* 2020); (v) we empirically demonstrate that our new Relational Embeddings clearly outperform popular graph-based approaches to encode interaction data, such as DeepWalk or node2vec; (vi) exhaustive ablation and error analyses show that the method used to obtain the *retweet* data and the size of the users community is crucial for state-of-the-art performance using our technique.

4.2 Dataset Generation: VaxxStance

Following the formulation of stance provided by Mohammad *et al.* (2016), the VaxxStance dataset (Agerri *et al.* 2021) consists of determining whether a given tweet expresses an *against*, *favor* or *neutral* (none) stance towards vaccines. Addi-

tionally, and inspired by the SardiStance 2020 shared task (Cignarella *et al.* 2020), the dataset includes two different types of data: Textual and Social (retweets, friends and user data), for two languages, Basque and Spanish. The dataset is publicly available in the task website[1].

4.2.1 Collection and Annotation

In a first attempt we tried to do the data collection and annotation for both languages following the same methodology. However, as it will be explained below, due to the idiosyncrasies of Basque it was necessary to devise an alternative, more viable, method for that language, especially to obtain the required textual data.

In any case, we did specify a number of criteria that both languages needed to comply with. First, the datasets a required to have a balanced distribution in the ratio users/tweets to avoid that a large number of tweets belonged to a very few users. Second, the tweets in the training set had to be written by different users from those contained in the test set. This is to avoid obtaining artificially high results due to the existence of user-based information in both the training and test sets. As such, the general idea is that both the textual and user-based (or social) knowledge would help each other in order to better classify stance. Finally, we use the annotation guidelines from the SemEval 2016 task (Mohammad *et al.* 2016).

Basque

Basque is spoken by roughly the 30% of the population in the Basque Country, and understood by around 50%. Due to the fact that Basque is a co-oficial language, it does have presence in the regional public administration, as well as in the education system and some news media, including a public television broadcaster. Still, the presence of Basque in mass media is extremely low, especially when compared to Spanish, the 4th most spoken language in the world.

The increasing popularity of Social Media such as Twitter allow researches to collect large amounts of textual or social data even among users of low resourced languages. This provides a valuable resource to study new NLP tasks such as stance detection not only for large and popular languages, but also for low resourced ones. Still, the collection process of enough tweets relevant to the VaxxStance task was rather challenging.

At first we experimented with a keyword extraction method using the following specific keywords: *"txertoa"* (vaccine) and *"txertaketa"* (vaccination),

[1] `https://vaxxstance.github.io/`

"negazionista" (negationist), *#pfizer*, *#moderna*, *#astrazeneca* and their respective inflections. However, it was surprising to find that the traffic of Basque tweets relative to those topics were relatively low.

We therefore decided to try an alternative, more brute-force, method. First, we collected all the available timelines of users that are identified to write mostly in Basque (around 10k users). The content of these timelines amount to around 8M tweets. Second, relevant tweets were selected following a simple keyword search using the same keywords listed for the previous attempt. Third, a first annotator manually labeled a set of around $1,400$ tweets. Finally, those same $1,400$ tweets, belonging to 210 users, were blindly annotated by a second annotator. The final composition of the textual part of the dataset can be seen in Table 4.1.

		Train	Test
Tweets		1,072	312
	Favor	327	85
	Neutral	524	135
	Against	219	92
Users		149	61

4.1 Table – Textual data in the Basque dataset.

We would like to note that the most difficult part in the process was finding enough users that explicitly expressed a stance AGAINST vaccines.

Spanish

Around $2,700$ tweets written in Spanish stating an opinion about *"vaccines"* were collected and annotated, as shown by Table 4.2. In order to avoid a potential bias derived from the current COVID-19 pandemic situation, the tweets were collected from the beginnings of Twitter until current time. They were also restricted to the peninsular variant of the Spanish language in order to avoid problems derived from the use of different terms in other variants such as Colombian, Peruvian, etc. To guide this process we used the Google tool *Google Trends*[2] which allowed us to locate temporal spaces where events related to vaccines had occurred, identifying the type of event and the date on which it happened. Some examples are the peaks in traffic for and against the vaccination against measles, which was a

[2] `https://trends.google.es/trends/?geo=ES`

consequence of some measles outbreaks that happened in Spain during 2019. By using keywords related to the event and restricting the dates obtained, we managed to introduce tweets related to events other than the COVID-19 vaccination process.

		Train	Test
Tweets		2,003	694
	Favor	937	359
	Neutral	591	195
	Against	475	140
Users		1,261	414

4.2 Table – Textual data in the Spanish dataset.

In addition to the tweets collected through the events identified in Google Trends, for the rest of the tweets collected we applied the following process. First, we used a set of keywords such as "*vaccine*", "*vaccination*", as well as terms related to diseases whose vaccines have generated some controversy in society and in anti-vaccine movements, e.g., "*chickenpox*", "*autism*", "*MMR*", etc. After a first manual analysis, we observed that the vast majority of the tweets collected did not express a stance. In order to solve this problem, we then extracted the hashtags most commonly used in these tweets and manually analysed those that were used to express a position in favour and/or against vaccines. Some examples of these hashtags are *#YoMeVacuno*, *#VaccinesWork*, *#COVID19*, *#vacuna*, *#yomevacuno*, *#VacunaCOVID19*, *#YoNoMeVacuno*, *#gripe*, *#Plandemia*, *#yosimevacuno*, etc.

By using these hashtags, we managed to increase the number of tweets to start with the manual labeling. The labelling was performed manually by two annotators, using a third annotator to resolve disagreements. For this we used the web platform created by Cignarella *et al.* (2020), to whom we would like to thank for their help using it.

Once the manual annotation was completed, the set of AGAINST tweets was much smaller than those expressing a FAVOR or NEUTRAL stance. To address this issue, we identified several accounts of users that may potentially be identified as supporters of anti-vaccine movements and manually collected tweets from these users expressing an AGAINST stance. This step was performed taking care in complying with the general criteria of not overlapping users between the training and evaluation set.

4.2.2 Social Media Information

The main objective is studying the usefulness of the context provided by social media information to classify stance in a crosslingual setting. With this objective in mind, we collected social information relative to the *friends* of the authors of the tweets as well as their *retweets*. The context provided by *friends* and *retweets* can be leveraged to generate relation graphs that in turn may be used to improve the classifiers.

Table 4.3 shows the social media data gathered with respect to the tweets in the train and test partitions for each of the languages. In addition to the retweets of the tweets included in the datasets, for Basque we also decided to collect the retweets made by the users, namely, by extracting the retweets from the users' timelines (TL). This strategy was applied in order to alleviate the small number of retweets obtained from the tweets in the train and test partitions.

		Train	Test
Basque	Friends	119,977	53,029
	Retweets	203	0
	Retweets (TL)	130,369	61,438
Spanish	Friends	1,708,396	438,586
	Retweets	6,832	2,148

4.3 Table – Social Media Information by language.

Finally, apart from social media information, the dataset also includes the meta information of each annotated tweet as well as the information related to each user.

4.2.3 Final dataset

Table 4.4 shows the composition of the VaxxStance dataset, including both textual and social information. Regarding the textual information, it can bee seen that the Spanish set is roughly double in size with respect to the Basque one, although the distribution of classes across the train and test set, as shown by Tables 4.1 and 4.2, is quite similar.

In regards to the social information, it is apparent that the Basque language community on Twitter is significantly smaller, comprising approximately 10% of the user base compared to Spanish. This difference reflects the comparatively smaller presence of Basque-speaking users on the platform. Correspondingly, the

		Train	Test
Basque	Tweets	1,072	312
	Users	149	61
	Friends	119,977	53,029
	Retweets	203	0
	Retweets (TL)	130,369	61,438
Spanish	Tweets	2,003	694
	Users	1,261	414
	Friends	1,708,396	438,586
	Retweets	6,832	2,148

4.4 Table – Composition of the VaxxStance dataset.

data concerning "*friends*" relationships also mirrors this ratio, with the number of connections being approximately 10% of those observed for Spanish users (see Table 4.4). Moreover, when examining the number of "*retweets*", the data for Basque is notably sparse, with only a limited number of instances obtained. This scarcity prompted the decision to include data from each user's timeline (*retweets TL*) to augment the available social information.

In summary, the VaxxStance dataset provides the first benchmark to investigate crosslingual approaches to stance detection based on both textual and social features. While the Basque set is slightly smaller than some previous approaches (Taulé *et al.* 2018; Cignarella *et al.* 2020; Zotova *et al.* 2021) it is still larger and more balanced than the data provided for any of the topics in the SemEval 2016 dataset, which is perhaps the most popular benchmark for stance detection (Mohammad *et al.* 2016).

4.3 Other Stance Detection Datasets

For comprehensive experimentation, we need stance detection datasets that include, in addition to the labeled textual data, interactions of the users that published the tweets, such as each user's *friends* and *retweets*. To the best of our knowledge, apart from our previously introduced VaxxStance, there is only one publicly available dataset that includes social information, named SardiStance (Cignarella *et al.* 2020) as seen in Section 2.2.2.

In order to include more data and languages, we tried to obtain user informa-

tion for SemEval 2016 (Mohammad *et al.* 2016) and other English datasets (Lai *et al.* 2020b; Conforti *et al.* 2020; Glandt *et al.* 2021), without much success. In the case of SemEval 2016, we managed to retrieve less than 30% of users. For other datasets, it was simply not possible to extract the tweets based on the IDs published. However, we did manage to extract the interactions (over 80%) for other dataset, namely, the Catalonia Independence Corpus (Zotova *et al.* 2021).

Thus, the final choice consists of seven datasets on three different topics (Independence of Catalonia, antivaxxers, Sardines movement) and four languages (Basque, Catalan, Italian and Spanish). The number of labeled tweets and their distribution between train and test sets can be seen in Table 4.5. This choice of data offers a varied relation user-tweets (very low in SardiStance, quite high in CIC), which would also allow to test the robustness of our method.

	Tweets			Interactions		
	Train	Test	Total	Users	RTs	Friends
CIC-ca	8,038	2,010	10,048	691	10M†	24M†
C-ca*	8,056	1,992	10,048	691		
C-es	8,036	2,011	10,047	334		
C-es*	8,016	2,031	10,047	334		
S	2,132	1,110	3,242	2,827	575K	3M
V-eu	1,072	312	1,384	210	190K	170K
V-es	2,003	694	2,697	1,675	9K 552K†	2.1M

4.5 Table – Stance detection Datasets: C (CIC), S (SardiStance), V (VaxxStance); * means no overlap of users across train and test; RT (retweets). † mark represents supplementary interaction-based data added by us.

4.3.1 Catalonia Independence Corpus

The Catalonia Independence Corpus (CIC) includes coetaneous tweets in both Spanish and Catalan (Zotova *et al.* 2020), is multilingual, quite large (10K tweets) and reasonably balanced. In the original CIC data, 92.50% of the users in the Catalan set occur also in the test set, whereas for Spanish the proportion is even higher, namely, 99.72%. In order to avoid any possible overfit to authors' style, a second version of the dataset (Zotova *et al.* 2021) distributes the tweets in such a way that their authors do not appear across the training, development and test sets

(CIC*). We extracted a total of 10M retweets and 24M friends of the users within the dataset, adding social data.

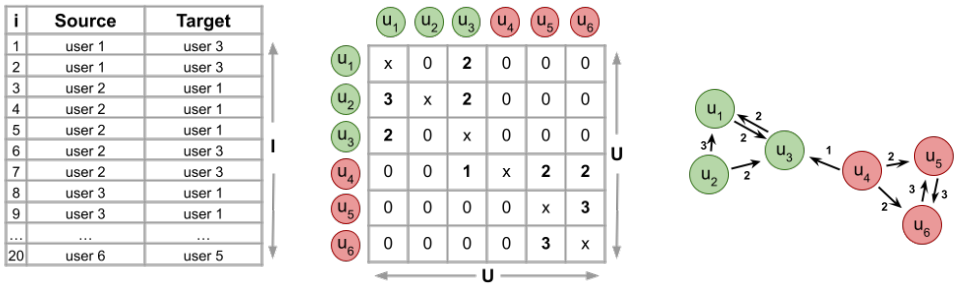
4.3.2 SardiStance dataset

This dataset contains tweets in Italian about the Sardines movement (Cignarella *et al.* 2020). In addition to the textual data, it also provides social and user information, such as the authors' friends and the retweets. We were unable to extract any supplementary data, because both tweet and user identifiers are encrypted.

4.4 Method

We proposed a new method to generate vector-based representations of interactions in social networks, such as *friends* and *retweets*. These new representations, which we refer to as *Relational Embeddings* (RE), are then leveraged to propose two techniques to perform stance detection: (i) building classifiers using just RE (§4.4.2) and, (ii) combining RE with various classifiers based on textual data (§4.4.3).

4.4.1 Relational Embeddings



(a) Interaction pairs **(b)** Frequency matrix **(c)** Interaction network

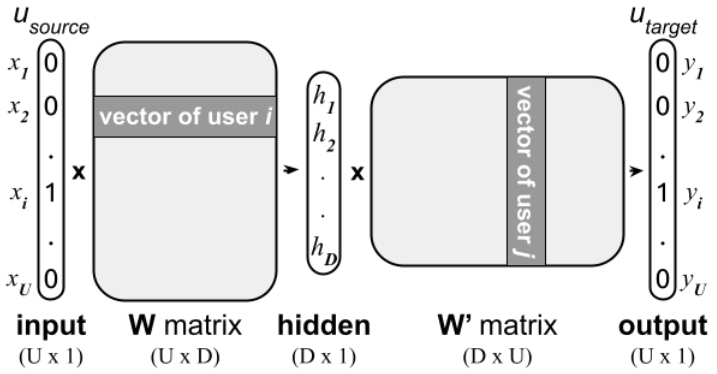
4.1 Figure – Different representations of the same data based on 20 interactions (I), generating a made-up directed network with 6 nodes (U) and 9 edges

In this work the type of interactions we will use are *retweets* and *friends*, which are seen as relations between two users, one generating the action (source) and the other receiving it (target). Thus, the actions of *retweeting* or *following* other users

are considered interaction pairs. Generally, these interactions should help to reveal users' preferences by capturing meaningful information from their performative actions.

The first step in our method consists of gathering the interactions from the users included in the labeled data, namely, the one-to-one *retweet* and *follow* actions between the users/authors of the tweets. It should be noted that a set of *retweet* and *follow* interactions can consist of independent one-to-one actions without direct relation between them. This is why in our model we consider each interaction pair as a single instance without any preprocessing or modification.

Using this interaction data, our model is then trained in an unsupervised manner to predict, in each instance, a target user from a given user. Note that the instances used as input are actual interactions pairs (Figure 4.1a). Thus, they do not correspond to sparse interaction frequency matrices (Figure 4.1b) or neighbours arisen from interaction networks (Figure 4.1c). Thus, the input of the model consists of real interactions, without generating artificial ones as random walks do.



4.2 Figure – One hidden layer Artificial Neural Network.

In order to obtain our relation representations, we use a single hidden-layer neural network (Figure 4.2). The network is used to train a dense interaction representation model using the *friends* and/or *retweet* based data. Each user is encoded as a one-hot vector of size U, where U is the number of users among interaction pairs (I) in a specific dataset. Given a one-hot vector U, the aim of the single hidden-layer feed-forward neural network consists of predicting the target user. The dimensions of the hidden layer (D) determine the size of interaction vectors representing the target user, which correspond to the number of learned features. During training, the weights W and W' are modified to minimize the

loss function due to back-propagation. According to Equation 4.1, the summation goes over all the interaction pairs (I) in the training corpus, computing the log probability of correctly predicting the target user (u_{target}) from the source user (u_{source}) for each interaction (i). The training process is done by sub-sampling the most frequent instances and with negative sampling (Mikolov *et al.* 2013a). Finally, the W matrix is used to represent each user based on their interactions, thus being the representation of the RE model. Therefore, user representations derived from interactions will be extracted from that matrix. Furthermore, users with similar interactions should have similar representations, turning many interaction pairs into dense interaction representations of D dimensions.

$$\frac{1}{I}\sum_{i=1}^{I}\log p(u_{target}|u_{source}) \tag{4.1}$$

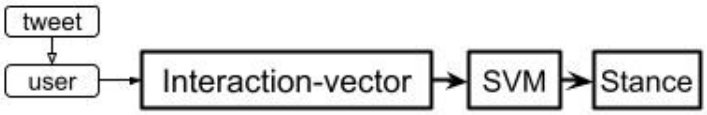
4.4.2 Interaction-based Classifier with Relational Embeddings

Our first system consists of a linear classifier taking as input only the Relational Embeddings described in the previous section. Building such a system will allow us to understand the performance of the generated RE models on their own.

Each of the tweets from a dataset will be represented by its author's (user) interaction vector, which represents the interactions of its author. By doing so, we effectively project the relations of the author into tweet level, generating a link between the relational data and the stance labels. In this step it is possible that some users may be repeated among the data, but their assigned stance label will be that of the corresponding tweet. It should be noted that, although possible, it is quite uncommon to have a user with tweets labelled differently across the data. Thus, each tweet is converted into an interaction vector, represented by the specific user's vector weights in the RE model. Those users not present in the model are represented as vectors of zeros. This is usually due to the inability of retrieving user interaction data, either because the user has disappeared from Twitter or because their profiles are kept private. As shown in Figure 4.3, the final interaction vectors for each tweet are used to train a SVM (RBF kernel) classifier without any additional input.

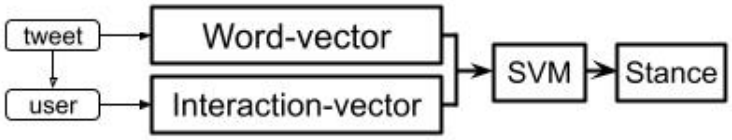
4.4.3 Combining Textual and Interaction Data

In order to combine textual and interaction data, we use both the texts conveyed by a given user and its associated social media interactions as data input. More



4.3 Figure – Relational Embeddings + SVM model architecture.

specifically, we propose two ensemble methods with the aim of combining textual and interaction-based features:
SVM-based combined model: In Figure 4.4, we obtain FastText (*FTEmb*) dense or TF-IDF sparse word vectors to represent each tweet. Subsequently, the interaction vector of each author is concatenated with the textual vector of the tweet. In cases where no relational information is available, a vector of zeros is added to the textual vector. Finally, the concatenated textual and interaction vectors are used to train an SVM model with an RBF kernel.

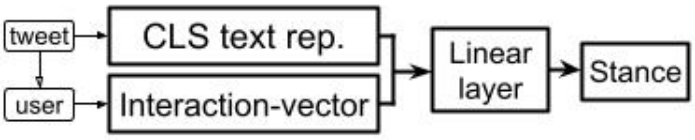


4.4 Figure – SVM based combined models architecture.

Transformer-based combined model: As shown in Figure 4.5, Transformer models and interaction embeddings are combined by concatenating the start-of-sequence representations of each tweet with the interaction vector of the user. When there is no user information related to interactions, a vector of zeros is concatenated to the start-of-sequence vector representation. Finally, a linear layer is added for classification (Devlin *et al.* 2019) on top of the concatenated vectors, while fine-tuning the model end-to-end.

4.5 Baselines

We experiment with node2vec and DeepWalk for direct comparison with other popular interaction-based methods and various text-based classifiers with text as the only input.



4.5 Figure – Transformer based combined model architecture.

Text-based Methods

In order to compare interaction-based with textual only approaches, we choose three commonly used text classification systems for stance detection (AlDayel and Magdy 2021; Küçük and Can 2020; Hardalov *et al.* 2021; Zotova *et al.* 2021) to establish a baseline to compare with our RE models:
Word Embeddings + SVM: We use FastText CommonCrawl models trained using the C-BOW architecture and 300 dimensions on a vocabulary of 2M words (Grave *et al.* 2018). For classification, each tweet is represented as the average of its word vectors (Kenter *et al.* 2016) which is then used to train a SVM (RBF kernel) classifier.
TFIDF + SVM: TF-IDF (Term Frequency Inverse Document Frequency) vectorization is applied in order to reduce word vector dimensionality by lowering the impact of words that occur too frequently in the selected corpus. TF-IDF vectorizing is applied over the text of the tweets, selecting most salient features and reducing sparsity. The obtained TF-IDF vectors are used to learn a SVM (RBF kernel) model.
XLM-RoBERTa: We use XLM-RoBERTa (Conneau *et al.* 2020) for text classification, a LM pre-trained for 100 languages on 2.5 TB of CommonCrawl text. This model has been widely tested for stance detection with state-of-the-art performance (AlDayel and Magdy 2021; Ghosh *et al.* 2019; Küçük and Can 2020; Espinosa *et al.* 2020; Zotova *et al.* 2021).

Interaction-based Methods

Apart from Relational Embeddings described in the previous section (§4.4.1), we also use graph-based approaches (Section 2.1.4) to extract user representations of the tweets' authors which are then used to train a SVM (RBF kernel) classifier, as described in Section 4.4.2.
DeepWalk (Perozzi *et al.* 2014): Given a node in the network, this algorithm learns feature representations to predict their context or neighbours. In this item-

to-context (Skip-gram) predicting task, the neighbours to predict may be artificially generated by simulating random walks among the connected nodes.
node2vec (Grover and Leskovec 2016): Similar to DeepWalk, but adds two parameters to control the structure of the network during the generated random walks. The Control parameters focus on probability of revisiting points and on the probability of visiting further points.

4.6 Experiments

Retweets are used to share specific content from other users' publications. The reiteration of these actions may demonstrate attachment to a user or its content, actively showing the specific preferences of the source user. Furthermore, although retweet actions are more likely to encode latent information related to community or polarization (Conover *et al.* 2011b; Zubiaga *et al.* 2019), we also wanted to include *friends* related data, which is a result of a *following* action. This passive action allows the source user to be aware of what is being said without sharing or promoting any content. Finally, we also combine both *retweets* and *friends* in a *mixed* representation to test whether merging passive and active interaction types in the same interaction space helps to embed social information. Therefore, for each dataset (CIC, SardiStance and VaxxStance) three different types of interaction-based embeddings were trained, based on the data source: (i) *retweet*, (ii) *friends* and, (iii) *mixed* embeddings.

The best interaction-based embeddings for each dataset was chosen by evaluating them with the classifiers built with interaction-based representations (RE, DW and N2V) via 5-fold cross validation on the training data. The results showed that, for RE, *retweet* was the best interaction for CIC and VaxxStance-eu, whereas the *mixed* embeddings were the best for SardiStance and VaxxStance-es. With respect to DW and N2V, *retweets* were best for CIC and VaxxStance, while *mixed* performed better for SardiStance.

The procedure for selecting the remaining hyperparameters, such as the dimensionality of the interaction embeddings, is described next. The dimensionality for the interaction-based embeddings were chosen by training a SVM classifier via grid search with 5-fold cross-validation. Based on results reported in previous work on reducing huge interaction matrices into low dimensional features for stance detection (Darwish *et al.* 2020; Stefanov *et al.* 2020), dimensions for interaction-based embeddings were selected between 10 or 20 dimensions. The best performing RE for CIC and SardiStance were of dimension 20, while for

VaxxStance they were of dimension 10. With respect to DW and N2V, the best dimensionality for CIC-es, CIC-es* and SardiStance was 10, whereas the best one for CIC-ca, CIC-ca* and VaxxStance correspond to 20. Moreover, for DW and N2V we set the usual default values for the hyperparameters for these algorithms: walks_per_node = 10, walk_length = 80, window or context_size = 10, and the optimization is run for a single epoch (Perozzi *et al.* 2014; Grover and Leskovec 2016). Specifically, for node2vec, we set p=1 and q=0.5 in order to enhance network community related information (Grover and Leskovec 2016). Grid search with 5-fold cross-validation was also used to optimize C and Gamma hyperparameters for every SVM system (RE, DW, N2V, FTEmb, TF-IDF and all the combinations). For XLM-RoBERTa, hyperparameter tuning was done by splitting the training set into a train and development sets (80/20). Results on the development set allowed to obtain the following hyperparameters: 128 maximum sequence length, 16 batch size, 2e-5 learning rate and 5 epochs.

Finally, as it is customary for this task, despite training and predictions being done for the 3 classes, evaluation is performed by calculating the averaged F1-score over the AGAINST and FAVOR classes (Mohammad *et al.* 2016).

4.6.1 Evaluation Results

Table 4.6 reports the results obtained for every system and dataset. They show that combining our Relational Embeddings with the text-based classifiers systematically helps to substantially improve results. This results in SOTA performance for every language and dataset except for Basque. This is partially due to the fact that the system based only on RE (RE+SVM) obtains high scores for most of the settings, often outperforming the XLM-RoBERTa baseline. Moreover, the experimental results demonstrate that the RE consistently outperform node2vec and DeepWalk across all datasets, underscoring the significance of considering interactions as pairs rather than generating arbitrary random walks.

For SardiStance, our result is slightly better than the state-of-the-art (Espinosa *et al.* 2020) on this dataset. However, unlike our approach, which is based on REs and which does not require any manual tuning, they included a large number of manually engineered features based on external resources for sentiment, psychological features, as well as social network features. In any case, the results of RE+SVM clearly improve over the best textual system published at the SardiStance shared task, which scored 68.53 in F1 score, based on an ensemble of Transformer models (Giorgioni *et al.* 2020).

Something similar occurs in the case of VaxxStance, where the state-of-the-

		C-ca	C-ca*	C-es	C-es*	S	V-es	V-eu	avg.
Interactions	RE	<u>82.2</u>	70.2	<u>85.2</u>	<u>84.4</u>	<u>71.7</u>	<u>85.5</u>	48.4	<u>75.4</u>
	N2V	69.7	61.8	65.9	49.3	65.7	76.2	29.0	59.7
	DW	69.1	68.0	66.5	64.0	66.4	79.5	25.7	62.7
Text	FTEmb	61.6	62.6	57.4	59.8	56.1	71.3	47.7	59.5
	TF-IDF	75.3	71.6	73.7	73.1	63.4	76.5	<u>54.4</u>	69.7
	XLM-R	77.6	<u>74.6</u>	74.2	73.9	57.2	82.5	41.2	68.7
Combined	FTEmb + RE	82.2	69.2	88.6	**87.7**	74.0	89.1	73.2	80.6
	TF-IDF + RE	**82.2**	**80.2**	**92.5**	86.6	**74.6**	**90.2**	**75.3**	**83.1**
	XLM-R + RE	78.8	75.9	76.8	77.2	60.2	81.2	51.8	71.7
	FTEmb + N2V	71.7	63.7	71.3	57.7	70.3	80.4	48.4	66.2
	TF-IDF + N2V	77.6	74.1	80.1	72.6	70.9	85.8	54.1	73.6
	XLM-R + N2V	77.2	73.7	72.9	74.6	56.3	80.6	39.2	67.8
	FTEmb + DW	72.9	67.3	71.6	72.9	67.7	82.7	48.5	69.1
	TF-IDF + DW	79.2	75.9	81.1	79.7	70.1	86.0	54.7	75.2
	XLM-R + DW	78.0	75.1	73.5	72.8	55.6	78.7	47.3	68.7
Previous SOTA		74.7	74.9	74.7	71.8	74.4	89.1	77.7	76.7

4.6 Table – Evaluation results F1-macro (Mohammad *et al.* 2016) using interaction-based systems (RE, N2V and DW), text-based systems (FTEmb, TF-IDF and XLM-RoBERTa) and their combinations. Previous SOTA: CIC (Zotova *et al.* 2020: 2021), SardiStance (Espinosa *et al.* 2020) and VaxxStance (Lai *et al.* 2021). Transformer results are the average of 5 randomly initialized runs. In **bold**: best overall result. <u>Underlined</u>: best results using either interaction or text data.

art (Lai *et al.* 2021) for textual classifiers is of 57.34 for Basque and 80.92 for Spanish, respectively. The results reported by Lai *et al.* (2021) were obtained by manually implementing more than 30 different features tailored to each specific dataset and language. The features were based on stylistic information, tweet and user data, various lexicons, dependency parsing, and network information. Furthermore, they crawled 1M tweets for each language to obtain a larger word embedding model to generate also word embedding-based features. Our method, without any specific tuning to the dataset, or so many additional resources, obtains new SOTA for VaxxStance-es and competitive results in Basque.

The best overall system is the one combining Relational Embeddings with TF-IDF (TF-IDF+RE+SVM), systematically outperforming any other system. Similarly, N2V and DW embeddings also achieve their optimal performance when combined with TF-IDF encoding. While this may seem a bit surprising, TF-IDF

and statistical classifiers have proven to be highly competitive for stance detection in Twitter (Mohammad *et al.* 2016; Ghosh *et al.* 2019; Küçük and Can 2020; Zotova *et al.* 2021), even better than text classifiers leveraging large pre-trained language models.

The results presented in Table 4.6 can also be interpreted as ablation tests. First, among the interaction-based classifiers, only those using our RE outperform most of the time the textual classifiers. Second, combining both interaction and textual representations helps to systematically improve over the results of each type of representation used in isolation.

4.7 Discussion

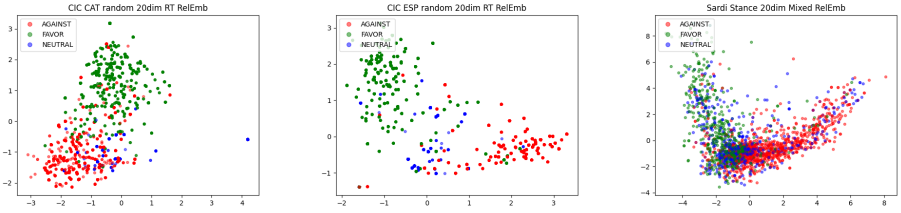
Our methodology for embedding sparse information is based on the use of actual interaction-pairs rather than on random walk methods employed by previous approaches such as N2V or DW. The experimental results reported show the superiority of our RE.

However, in order to have a better understanding of the RE and their performance and cross datasets, we plotted the users' representations and their stance. Thus, Figure 4.6 shows a 2D visualization obtained by applying a PCA dimensionality reduction to each RE from the training data.

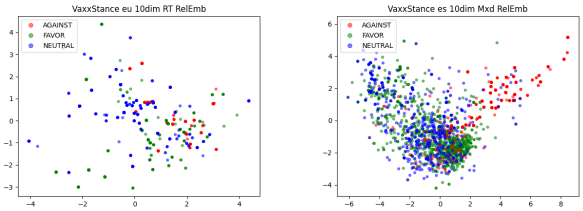
The visualizations show that there is a clear relation between the readability of the RE visualizations and the results obtained by the interaction-based systems using RE (RE+SVM). First, CIC embedding visualizations (Figures 4.6a and 4.6b) show very clearly defined communities of users who are in FAVOR or AGAINST of the independence of Catalonia. Furthermore, the RE obtained from the SardiStance interaction-based data also correlate to well-defined communities (Figure 4.6c) but with more overlap of stance labels than in the CIC representations, as is the case for VaxxStance-es.

However, the visualization of the VaxxStance-eu RE in Figure 4.6d shows that the target embedding representation does not manage to distinguish that clearly between FAVOR and AGAINST users. This might be due to the very small community of Basque Twitter users (see Table 4.5). In other words, Basque users in FAVOR or AGAINST vaccines naturally interact much more than users from the other two analyzed datasets. This might also explain the lower results obtained by the RE+SVM classifierin the VaxxStance-eu data.

The graphs also show that the specific nature of some targets makes them more suitable to generate our RE. Thus, topics that may reflect political homophily, such



(a) CIC-ca* **(b)** CIC-es* **(c)** SardiStance

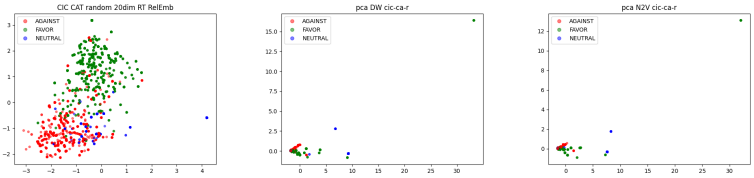


(d) VaxxStance-eu **(e)** VaxxStance-es

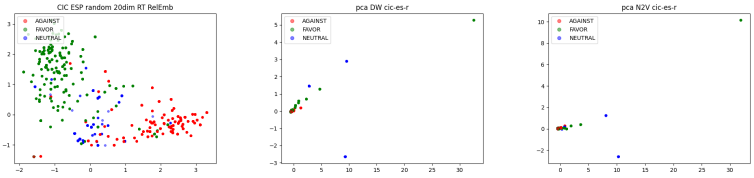
4.6 Figure – Relational embedding representation of training set users (PCA dimension reduction to 2).

as the independence of Catalonia, seem to generate clearer FAVOR and AGAINST communities than for the other two topics (vaccinations and the Sardines movement). However, in the case of VaxxStance, the small size of the community of Basque Twitter users may add difficulties to catch those orientations in such a small dataset.

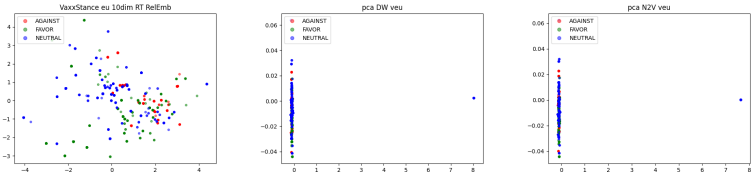
Figure 4.7 provides a direct comparison between the RE, DW and N2V user representations for each of the training datasets. The muddled DW and N2V visualizations with respect to those obtain with RE may explain the lower results obtained by DW and N2V.



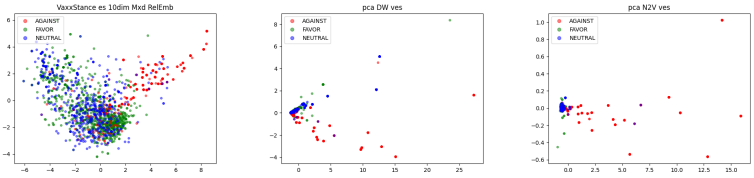
(a) CIC-ca* RE **(b)** CIC-ca* DW **(c)** CIC-ca* N2V



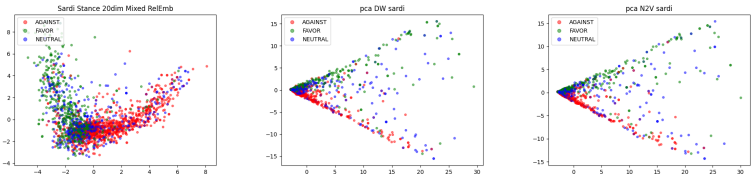
(d) CIC-es* RE **(e)** CIC-es* DW **(f)** CIC-es* N2V



(g) VaxxStance-eu RE **(h)** VaxxStance-eu DW **(i)** VaxxStance-eu N2V



(j) VaxxStance-es RE **(k)** VaxxStance-es DW **(l)** VaxxStance-es N2V



(m) SardiStance RE **(n)** SardiStance DW **(o)** SardiStance N2V

4.7 Figure – Comparison between Relational Embedding (RE), Deep Walk (DW) and node2vec (N2V) representations of training set users (PCA dimension reduction to 2).

4.8 Conclusion

This chapter aims to address the task of stance detection by employing not only the text itself but also the social information data available on social networks derived from *retweets* and *friends* interactions. With that objective, we provide a new dataset for stance detection towards vaccines across two different languages: Basque and Spanish. As a novelty for stance detection in these languages, systems can use textual and social information to train their systems in multilingual and crosslingual settings. In addition, we propose Relational Embeddings, a new method to represent interaction data such as *retweets* and *friends*. REs help to reduce the sparsity of interaction data by behaving like dense graphs, being able to embed information related to stance from different data sources without any manual engineering.

While this technique is language independent and fast to train and to apply, REs behave robustly across different datasets, targets and languages, substantially and consistently improving results by combining them with text-based classifiers. The results also show that using RE only we can outperform most text classification baselines. Furthermore, a direct comparison with previous approaches such as DeepWalk and node2vec show the superiority of our approach. The results and analysis performed shows that we need to pay more attention to social network data, aiming to address the shortcomings discussed by further researching different strategies to leverage such interaction data. Finally, despite our research is focused on advancing stance detection, we believe that the versatility of our REs extends far beyond this specific domain. We posit that these embeddings hold immense potential for effective application across a large number of Computational Social Science and Natural Language Processing tasks, particularly those intricately linked to social media, especially those related with ideology or political leaning.

5. CHAPTER

Dynamic Political Leaning Inference in Social Media

An ability to infer the political leaning of social media users can help in gathering opinion polls thereby leading to a better understanding of public opinion. While there has been a body of research attempting to infer the political leaning of social media users, this has been typically simplified as a binary classification problem (e.g. left vs right or conservative vs liberal) and has been limited to a single location, leading to a dearth of investigation into more complex, multiclass classification and its generalizability to different locations, particularly those with multi-party systems. Our work performs the first such effort by studying political leaning inference in diverse regions, each of which has a different political landscape composed of multiple political parties.

This study proposes a two-step dynamic approach, starting with unsupervised user representations leveraging retweets, followed by classification to infer the political leaning of a target user. To accomplish this task, we collect and release different datasets comprising users labelled by their political leaning as well as interactions with one another. On the one hand, with the intention of comparing multi-party and binary approaches, we consider three diverse regions from Spain (Basque Country, Galicia and Catalonia) due to the emergence of new political actors and plurinationality that makes them complex. On the other hand, in order to delve with different levels of political engagement, we contemplate three of the UK's nations (Scotland, Wales and Northern Ireland), each of which has also a different political landscape composed of multiple parties.

We investigate the ability to predict the political leaning of users by leveraging these interactions in challenging scenarios such as few-shot learning, where

training data is scarce, as well as assessing the applicability to users with different levels of political engagement. We show that interactions in the form of retweets between users can be a very powerful feature to enable political leaning inference, leading to consistent and robust results across different regions with multi-party systems. In addition, error analysis and visualizations show that Relational Embedding are able to capture inner-group and inter-group political affinities.

However, while interaction-based data might not always be accessible, we aim to broaden our scope by incorporating textual data in order to represent users. Access to textual and interaction-based data not only avoids reliance on specific data types but also allows us to compare these data sources. We show that, while state-of-the-art text-based representations on their own are not able to improve over interaction-based Relational Embedding representations, a combination of text-based and interaction-based modeling using hybrid approaches considerably improves performance, specially with users who are less engaged in politics.

5.1 Motivation and Contributions

Ideology is a set of people's beliefs that can be understood as ways of thinking and acting in society. Those beliefs can generally be represented by political parties, acting like social hubs of coordinated thoughts and actions. However, ideology is often presented and simplified into binary frameworks based on individuals stance over left/right or conservative/liberal orientations. Political leaning inference is proposed as a way of representing individual actors by the closest political party, analyzing ideology from a richer perspective. To better understand society, social researchers can benefit from the development of efficient methods capable of generalizing political leaning inference across different regions (Imhoff *et al.* 2022). We address this challenge by investigating new tools and techniques for conducting deeper and more accurate social and political research with the aim of improving opinion polls, political polarization studies, stance and propaganda detection or disinformation analysis among others.

The vast majority of research on political leaning inference has been limited to binary classification between the two prevailing parties or stances (Conover *et al.* 2011b: a; Barberá and Rivero 2015; Barberá 2015; Garimella and Weber 2017; Barberá *et al.* 2015; Pennacchiotti and Popescu 2011a; Hua *et al.* 2020; Xiao *et al.* 2020). The few works that conducted multiclass classification (Boutet *et al.* 2012; Makazhanov and Rafiei 2013; Rashed *et al.* 2021) were constrained to a single scenario or region. This however limits both the applicability of such methods

and the insights learned from such studies. Indeed, each social context has its own political reality which is typically reflected in more than two ideological options.

In order to broaden the study on the ability to infer the political leaning of social media users, we identify four key limitations in previous work. First, the widely used binary frameworks can be limiting and imprecise as individuals hold a wider range of political beliefs, which instead calls for multiclass classification. Second, political ideologies and beliefs can vary substantially across different regions since each community has its own idiosyncrasy, which calls for the study of applicability across regions. Third, to achieve generalizability, it is crucial to include scenarios where the labeled data available for training is limited, which makes critical the study of few-shot learning approaches. Fourth, it is important to consider that not every social media user is as engaged in politics and/or is vocal about their beliefs, which posits the importance of assessing the ability to predict the political leaning of all kinds of users regardless of their level of engagement.

By addressing the above limitations, we aim to further research in political leaning inference by providing the first study that focuses on generalizing a multiclass political leaning inference model across different scenarios or regions. With this goal in mind, we propose and evaluate a range of techniques for data extraction and user representation for multiclass political leaning inference. The aim is to independently represent Twitter users by leveraging their interactions, effectively transforming content sharing actions on Twitter into vector spaces. By accomplishing this, we seek to achieve adaptability across diverse situations, in turn opening an avenue for further exploration in other tasks. Thus, retweet interactions are selected to represent users, known for their effectiveness in achieving user classification (Conover *et al.* 2011a; Magdy *et al.* 2016; Darwish *et al.* 2020; Stefanov *et al.* 2020; Fernandez de Landa and Agerri 2022). We conducted experiments using four distinct unsupervised techniques, namely ForceAtlas2 (Jacomy *et al.* 2014), DeepWalk (Perozzi *et al.* 2014), Node2vec (Grover and Leskovec 2016) and previously presented Relational Embeddings (Fernandez de Landa and Agerri 2022). First, the named user representations are evaluated through different regions from Spain, each with different political parties, including also the first study on the ability of those techniques to rely on scarce training data. Second, we continue enquiring into these settings over diverse regions from the UK, including also the ability to determine the political leaning of users with different levels of engagement, delving into a more realistic performance of the methods. Third, we perform political learning inference by incorporating textual data to represent users, not to be reliant on a specific data type, thereby enabling comparison and integration with previous interaction-based approaches.

To the best of our knowledge, our work is the first to study the political leaning inference task across diverse multiclass political realities, which in turn leads to new insights into tackling this challenging setting. More specifically, this chapter makes the following novel contributions: (i) we devise and experiment with a pluralistic framework that includes multiple political leanings, which proves adaptable to different regions since it is grounded on localized political actors; (ii) we evaluate a range of methodologies to make the most of retweet interactions among social media users to infer their political leaning, showing that Relational Embedding based approach is effective even in weakly-supervised and realistic scenarios; (iii) we perform a comprehensive error analysis and feature visualizations in order to show the ability of the proposed methodology to capture socio-political information coherently and in alignment with the specific political context; (iv) we conduct political learning inference without relying on one specific data type. However, our results demonstrate that interaction-based Relational Embeddings outperform textual approaches; (v) we propose hybrid modeling of social media users leveraging textual and interaction-based features, showing that the combination of both data types is required for optimal performance.

5.2 Methods

We conduct experiments using various unsupervised methods to extract user representations from text and interaction data. On the one hand, interaction-based features are employed to represent users using data derived from retweet interactions. On the other hand, text-based features are employed to represent users using textual data. Finally we propose a combination of textual and interaction features in our hybrid approach. User representations are then used to perform political leaning inference via alignment to the political parties included in our dataset. We also evaluate the impact of dimensionality reduction techniques on interaction-based features.

5.2.1 Interaction-based User Representation Methods

We experiment with a set of user representation methods (presented in 2.1.4 and 4.4.1) based on leveraging retweet interactions that can represent users' preferences and behavior. Thus, retweet based interactions on their own have been shown to be effective for user classification tasks (Conover *et al.* 2011a; Magdy *et al.* 2016; Darwish *et al.* 2020; Stefanov *et al.* 2020; Fernandez de Landa and

Agerri 2022) and have been used to represent users as done in different approaches (Cignarella *et al.* 2020; Agerri *et al.* 2021; Fernandez de Landa and Agerri 2022; Darwish *et al.* 2020). Next, we describe the proposed user representation methods which are suited to transform large and heterogeneous data sources:

ForceAtlas2 (Jacomy *et al.* 2014) is a continuous graph layout algorithm, that transforms a network into a 2-dimensional space to obtain a readable shape. As a result of an approximation-repulsion process, the nodes that are unrelated repulse each other, while related ones will attract each other.

DeepWalk (Perozzi *et al.* 2014) algorithm simulates random walks among connected nodes in a network to learn feature representations. It predicts the context or neighbors of an instance using the Skip-gram method Mikolov *et al.*, 2013a. The context is generated by random walks among surrounding connected data points, with the length and number of walks determining the context.

Node2vec (Grover and Leskovec 2016) method, similar to DeepWalk, introduces two parameters, to determine the likelihood of revisiting nodes and control the probability of exploring unexplored graph areas. Thus, the representations can vary to capture either homophily or structural equivalence.

Relational Embeddings (Fernandez de Landa and Agerri 2022) are based on a single hidden-layer neural network trying to predict who retweeted whom for all the gathered interaction pairs.

Retweet-based interactions are used to feed the aforementioned methods to train the unsupervised models. We use all the available users in order to embed as much information as possible. The user author of the retweet is considered the source of the interaction, while the user receiving the retweet would be the target. We generate source-target user pairs in this manner to provide the input to the user representation methods. With the intention of generating low dimensional but meaningful representations and based on previous work (Darwish *et al.* 2020; Stefanov *et al.* 2020; Fernandez de Landa and Agerri 2022), dimensions were set on 20 dims for DeepWalk, Node2vec and Relational Embeddings. The hyperparameters for node2vec and DeepWalk were set to the default values typically used by these algorithms: walks_per_node = 10, walk_length = 80, window or context_size = 10, and the optimization is executed for a single epoch (Perozzi

et al. 2014; Grover and Leskovec 2016). For node2vec, we set p=1 and q=0.5 in order to enhance network community related information (Grover and Leskovec 2016). Default parameters were set for ForceAtlas2 and independent models were trained for each of the regions.

5.2.2 Text-based User Representation Methods

We explore a range of text-based feature extraction techniques presented in Section 2.1.3 in order to represent users. The extraction process involves generating user vectors based on the textual content associated with each user, namely, their tweets. By utilizing tweets as input data, we aim to capture users' preferences and behaviors. With that purpose we will employ text-based user representations to predict users' political leaning, following similar methodologies to those employed in previous studies (Fagni and Cresci 2022; Fernandez de Landa and Agerri 2022; Hallac *et al.* 2019).

Term Frequency Inverse Document Frequency The tfidf statistical measure assesses the relevance of a word to a document within a set of documents. By lowering the impact of words that occur too frequently in the selected text collection the most salient features are selected. Our use of this approach is motivated by the fact that a limited set of words significantly impact the final predictions (Shen *et al.* 2018) and the remarkable results obtained in other similar text classification approaches (Fernandez de Landa and Agerri 2022; García-Díaz *et al.* 2022). In this case, all the tweet collections from each user are considered as individual documents. The obtained tfidf vectors for each author or user are used to learn a classifier, proposing a user level classification.

Word Embeddings We use word2vec (Mikolov *et al.* 2013a) (w2v) to encode each of the tweets into text-based vector representations. To accomplish this we train our own models from scratch in order to fit all the words from our datasets into the vocabulary. A separate language model is trained for each region, aiming to capture the prevailing expressions within specific communities. These models are trained using default hyperparameters (C-BOW, negative sampling, 5 epochs, 5 words window size) but different dimensions are considered to select the optimal configuration. For our purposes, we extract tweet vectors one by one, representing each tweet as the average of its word vectors (Kenter *et al.* 2016; Fagni and Cresci 2022).

Transformers We use pre-trained Transformer models (Vaswani *et al.* 2017) as they have garnered considerable attention in recent years for text classification tasks, including user profiling through textual data (García-Díaz *et al.* 2022). These models enable context and meaning representation by analyzing the relationships among tokens in a text sequence. Transformer-based contextualized embeddings values are modified depending on the surrounding words and their order, while static embedding methods such as word2vec represent words with fixed vector values. Four multilingual models have been selected for our experiments:

- **mBERT** (Devlin *et al.* 2019) model is the multilingual version of BERT (Devlin *et al.* 2019) pre-trained with the largest 104 languages in Wikipedia. Rather than simply predicting the next word in the sequence, the BERT model takes into consideration all of the words in the sequence, thereby developing a more profound comprehension of the context. BERT pre-trains bidirectional representations from unlabeled text by considering both left and right context in all layers based on two pre-training objectives, namely, mask-language modeling and next sentence prediction.

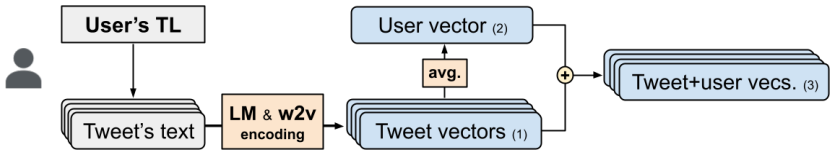
- **DistilmBERT** (Sanh *et al.* 2019) as the multilingual version of DistilBERT is a smaller and faster Transformer model distilled from BERT, with 40% fewer parameters and 60% faster performance while maintaining over 95% of BERT's performance on the GLUE benchmark.

- **XLM-RoBERTa** (Conneau *et al.* 2020) is a multilingual version of RoBERTa-base (Liu *et al.* 2019) pre-trained on a large multilingual corpus containing 100 languages. As a robustly optimized BERT approach, it is trained on a 10 times larger dataset than the used in BERT and using a dynamic masking technique, byte-pair encoding tokenization and without the next-sentence prediction objective.

- **XLM-T** (Barbieri *et al.* 2022) is an extension of the XLM-RoBERTa base model further trained on 198 million multilingual tweets. Given its focus on Twitter-based data, it is particularly relevant to assess its performance in tasks that are specific to this social media platform.

To ensure consistency across all categories and prevent overfitting, we employ the Transformers models without fine-tuning, meaning that we use the default frozen weights as done in other approaches (Fernandez de Landa and Agerri

2023). Text features are extracted separately for each tweet, treating each individual tweet as a sequence. We explored three distinct approaches for representing the text of tweets using transformers: (a) *start-of-sequence* initial token embedding is used as the entire tweet representation (Devlin *et al.* 2019); (b) *average* value of the output embeddings of all words in the tweet to represent each tweet as the average of its word vectors (Kenter *et al.* 2016; Fagni and Cresci 2022); (c) *max-pool* value of all the words in a tweet to extract the most salient features from every word-embedding, by taking the maximum values among all the word vectors (Zhelezniak *et al.* 2019).

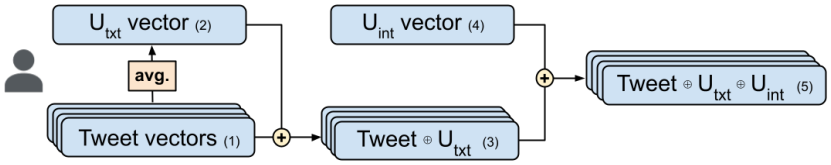


5.1 Figure – Text-based user representations. Pre-trained Transformer-based Language Models (LM) and word2vec embeddings (w2v) usage to extract text-based user representations: (1) at Tweet level, (2) at user level and (3) the combination between tweet and user features.

While the tfidf method is capable of directly representing all of a user's content at once through user-level features, word2vec and Transformers are able to extract information only at the tweet level. As the tweet level information is not sufficient to represent users due to the small amount of text, we generate and add user-level textual features to each of the tweets representation as shown in Figure 5.1. So, we obtain the textual representations of all the tweets for each user (1) which are then concatenated and averaged to obtain a user-based vector representation (2), emulating similar approaches (Hallac *et al.* 2019; Kenter *et al.* 2016; Le and Mikolov 2014; Rashed *et al.* 2021). The final representation (3) consists of the concatenation of each tweet vector with the user-based vector representation. The combined user and tweet representations for each tweet are used to train a classifier. After predicting the labels at tweet level, a majority voting strategy is employed to infer the user label by considering the various tweet labels associated with the same author (García-Díaz *et al.* 2022; Fernandez de Landa and Agerri 2023).

5.2.3 Hybrid User Representation Methods

We propose a hybrid approach which integrates both the textual content expressed by a user and their corresponding social media interactions. When using text representation methods at tweet level (Figure 5.2), the textual representation of the tweets for each of the users (1) are averaged to obtain a vector characterizing each user (2) and concatenated with each of the tweets written by the author (3). Afterwards, (4) we concatenate them with the interaction-based representations to obtain a final (5) hybrid feature for each of the tweets. Those hybrid representations are used to train a classifier and to predict the labels at tweet level, and we subsequently use a majority voting strategy to infer the user label by considering the various tweet labels associated with the same author (García-Díaz *et al.* 2022; Fernandez de Landa and Agerri 2023).



5.2 Figure – Hybrid User Representation Method for each user tweet by tweet: (1) Tweet level text representation, (2) user level text representation, (3) the combination between tweet and user text features, (4) user level interaction representation and (5) final hybrid representation concatenating all vectors.

Alternatively, when text representation methods (i.e. tfidf) facilitate the extraction of text features at the user-level, hybrid features are created at user level, concatenating user-level text and interaction features. User-level hybrid representations are used to train a classifier and to predict the labels at user level.

5.2.4 Dimensionality Reduction Techniques

We aim to study the performance of user representation methods for settings in which only limited labeled data is available, namely, in what we refer to as a weakly supervised scenario. We employ various dimension reduction techniques to compress information for weakly supervised settings with the aim of forcing our model to generalize more from the characteristics seen in the training data. We use 3 different dimensionality reduction methods:

PCA or Principal Component Analysis is a linear algorithm for dimensionality reduction which aims to represent the data in a low-dimensional space while

preserving the original global structure of the data. As this is a linear algorithm, it will only find linear dependencies or relationships in the data, without considering the neighbours on their own.

t-SNE (Van der Maaten and Hinton 2008) is a nonlinear dimensionality reduction technique that embeds high-dimensional data into a lower dimensional space. A probability distribution is built over data point pairs, giving similar data points a high probability while dissimilar ones are assigned a lower one. The algorithm computes the probability that pairs of data points in the higher dimensional space are related, and then chooses low-dimensional embeddings which produce a similar distribution based on the Kullback–Leibler divergence.

UMAP (McInnes *et al.* 2018) or Uniform manifold approximation and projection is also a nonlinear dimensionality reduction technique. This technique is similar to t-SNE, but it assumes that the data is uniformly distributed on a locally connected Riemannian manifold. UMAP creates a fuzzy graph that reflects the topology of the high dimensional graph based on the nearest neighbours of each data point. Then the low dimensional graph is built based on the fuzzy graph. This dimension reduction technique is able to reflect the large scale global structure, while also preserving the local structure.

The inputs are user vectors derived from the selected user representation method as presented in the previous subsection. Every dimension reduction techniques are used with their default hyper-parameters. Dimensionality is set to 2, following previous work (Darwish *et al.* 2020; Stefanov *et al.* 2020) and separate models are trained for each of the regions.

5.3 From binary to multy-party political leaning

Traditional two-party political systems are evolving into scenarios where new political actors are emerging (Lisi 2018) in response to new ideological conflicts (Ford and Jennings 2020) or socio-economic consequences (Kotroyannos *et al.* 2018; Morlino and Raniolo 2017) by suggesting innovative political proposals and novel approaches (Rama *et al.* 2021). Furthermore, transnational integration is leading to the emergence of plurinational states, where the presence of a singular national sentiment is not readily apparent (Keating 2001), and diverse political leanings representing distinct national sentiments are arising (McGann *et al.* 2019). Consequently, each political region develops its own political parties tailored to suit specific socio-political circumstances, with the aim of obtaining sup-

port and sympathy among the population. Hence, characterizing political leaning in terms of proximity to a specific political party would offer a more accurate representation of political nuances in contrast to oversimplifying frameworks such as left-right or conservative-liberal. In reality, every social context has its unique political landscape, characterized by more than just two ideological choices, which evolve in response to societal requirements. Moreover, a dynamic approach is essential for tailoring political leaning inference to specific times and regions, specially on complex political scenarios characterized by numerous and evolving political options.

To effectively deal with the complexities of real-world politics, it's important to consider that political reality needs to be covered as a complex and volatile environment. Therefore, our investigation proposes a multi-party framework grounded on political parties, aiming to capture political leaning beyond static binary perspectives and embrace adaptability to diverse and complex political contexts. The objective is to discover effective and robust user representation methods that do not rely on manual annotation and can capture general socio-political information for subsequent specific classifications, including left-right or multi-party political leaning, among others. Those user representations are evaluated for their effectiveness in inferring binary and multi-party political leanings within three different politically complex regions from Spain as well as on scenarios with scarce training data.

5.3.1 Datasets: Spain

In order to study the proposed multi-party frameworks we generate our own datasets. First, we select specific political regions and the most relevant political parties for each of them. Subsequently, we extract data from Twitter, manually labeling users and collecting interaction data to build the user representations.

Political Region

Given our interest in analyzing complex political contexts, we have chosen the Kingdom of Spain on 2020 summer as our case of study. Spain is a complex political context, characterized not only by its status as a plurinational country (Keating 2001) but also because of the emergence of new political actors in the last years (Rama *et al.* 2021). Thus, new political actors (UP, Cs, and Vox) burst into Spanish political scenario ruled by traditional forces (PSOE and PP) suggesting updated political proposals and novel approaches (Rama *et al.* 2021). Moreover,

our study focuses on the Basque Country, Galicia and Catalonia, which are considered stateless nations within the multinational state of Spain (Keating 2001). These regions house a greater number of political parties than other areas, each with its own regional parliaments and distinct nationalist orientations that drive a wide range of political choices. For each of the selected regions we have selected the political parties that have (or potentially may have) political representation on the regional parliaments. In order to compare multi-party to a binary framework, we also annotate each of the political parties with left-right labels.

Basque Country (EUS): *Basque Nationalist Party* (Partido Nacionalista Vasco - PNV ●) Basque nationalist and Christian-democratic political party; *Unite* (Bildu ●) left-wing Basque pro-independence coalition; *Socialist Party* (Partido Socialista Obrero Español - PSOE ●) social-democratic Spanish political party; *Together We Can* (Unidos Podemos - UP ●) democratic socialist Spanish electoral alliance; *People's Party* (Partido Popular - PP ●) conservative and Christian-democratic Spanish political party; *Citizens* (Ciudadanos - Cs ●) liberal Spanish political party; *Voice* (Vox ●) conservative Spanish political party.

Galicia (GAL): *Galician Nationalist Bloc* (Bloque Nacionalista Galego - BNG ●) left-wing Galician nationalist coalition; *Galicianist Tide* (Marea Galeguista - MG ●) left-wing Galician electoral alliance. Despite some punctual differences definitions for PSOE, UP, PP, Cs and Vox are the same for this region, whereas representatives and party accounts are specific for the region.

Catalonia (CAT): *Republican Left of Catalonia* (Esquerra Republicana de Catalunya - ERC ●) social-democratic Catalan pro-independence political party; *Together for Catalonia* (Junts per Catalunya - JxC ●) progressives Catalan pro-independence political party; *Popular Unity Candidacy* (Candidatura d'Unitat Popular - CUP ●) left-wing Catalan pro-independence political party. Despite some punctual differences definitions for PSOE, UP, PP and Cs are the same for this region, whereas representatives and party accounts are specific for the region. Note that in this case Vox party is not represented, since being a new party, it had not yet been presented in this region.

Binary framework: In addition to the political party identification, we have incorporated a binary categorization to determine each party's left-right alignment. This was done to facilitate a comparison between the proposed multi-party framework, grounded in political parties, and a more straightforward binary framework. To categorize each political party, we utilized data from opinion polls (CIS 2019: 2020a: b), which gauge public perceptions of the left-right orientations of these parties.

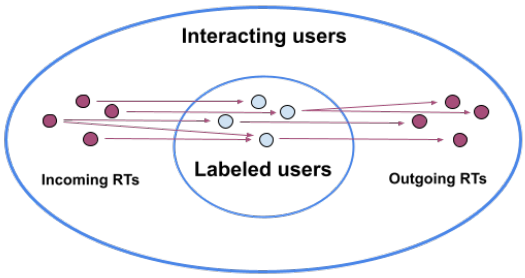
Data collection strategy

Data collection strategy consists of three steps, each applied to the selected regions. The initial step involves labeling users for supervised learning, while the other two steps focus on gathering interaction data to generate user representations for use as input data. Data collection was carried out in August 2020.
(1) **Manual labeling:** Our starting point relies on a user's seed list that consists of users related to the selected political parties of each region. Thus, we are selecting a sample of users in order to collect data, following the same technique as done in other works (Makazhanov and Rafiei 2013; Barberá 2015; Garimella and Weber 2017; Hua *et al.* 2020). The selected users are related to political parties of each region, such as the political organizations, elected members, candidates or even political militants. The identified user's are labeled by its political party, forming our labeled sample. Apart for the party level categorization for the multi-party framework, we add a binary categorization as left (L) or right (R) leaning for the binary framework (see Table 5.1). In the following steps, this user list is going to be used to extract the interactions to build user features for the labeled users.

EUS		GAL		CAT	
party	n	party	n	party	n
PNV (R) ●	146	BNG (L) ●	39	ERC (L) ●	18
Bildu (L) ●	134	MG (L) ●	7	JxC (R) ●	18
UP (L) ●	177	UP (L) ●	48	UP (L) ●	16
PSOE (L) ●	157	PSOE (L) ●	35	PSOE (L) ●	18
PP (R) ●	132	PP (R) ●	45	PP (R) ●	14
Cs (R) ●	40	Cs (R) ●	12	Cs (R) ●	14
Vox (R) ●	8	Vox (R) ●	7	CUP (L) ●	11
TOTAL:	794	TOTAL:	193	TOTAL:	109

5.1 Table – Labeled users from Spain for each region. Columns correspond to political party and number of users (n) per region. The labels between parenthesis correspond to the binary categorization left (L) or right (R) leaning for the binary framework.

(2) **Twitter data retrieval:** For every labeled user we first retrieve a history of their retweet interactions. The purpose of this initial retrieval is not to gather the retweets themselves but to identify all users who have interacted with the labeled users through retweets (see Figure 5.3). Afterwards, we gathered all retweets accessible from the timelines of both the labeled users and the interacting, extracting a substantial volume of interactions involving the sharing of content among users.



5.3 Figure – Interacting user's identification scheme. Central circle represents users manually or automatically labeled. External circle illustrates all the users interacting with the labeled users by retweeting them (incoming) or being retweeted by them (outgoing).

Thus, we have enough data to later represent the labeled users as well as the users interacting with them employing interactions from their timelines as done in other approaches (Fernandez de Landa and Agerri 2022). Table 5.2 shows the number of retweets retrieved from the labeled and interacting users during the summer of 2020.

	EUS	GAL	CAT
Labeled users	794	193	109
Interacting users	155 k	50 k	144 k
Retweets	58 M	13 M	41 M

5.2 Table – Final dataset composition for each Spanish region.

5.3.2 Experiment #1: Strongly Supervised scenario

Experimental Settings: The main goal of this experiment is to compare the performance of various interaction-based user representation methods when applying the same classifiers. Furthermore, we aim to compare our approach, which defines political leaning within a dynamic multi-party framework, with the conventional approach of treating it as a binary categorization. Thus, we experiment with different interaction-based userrepresentation methods while inferring binary or multi-class political leanings: (i) *Binary framework*, where only two classes are used to define users' political orientation as left or right. These same categories

are used across regions, making it a generalizable and uniform approach. (ii) *Multi-party framework*, where users' political leaning is defined as the closest political party. Political parties vary by region, with seven parties in each region, as specified in Section 5.3.1.

For each region and framework, we conduct experiments using a leave-one-out (LOO) cross-validation approach. This means that one user is held out for testing while all the remaining users are utilized for training. Therefore, a model is trained and tested individually for every user in the dataset, a feasible task due to the low dimensionality of the representations. The user representations obtained from each of the methods are used to train six different classification algorithms for each region and framework: Logistic Regression (LogReg), Random Forest (RF), Naive Bayes (NB) and linear, polynomial and RBF-kernel Support Vector Machines (SVM). Scikit-learn implementation (Pedregosa *et al.* 2011) with default configuration is used with that purpose.

Results: We compare the performance of the diverse user representations combined with different classifiers in a strongly supervised scenario (leave-one-out cross-validation). Besides, results are also compared between binary (Table 5.3) and multi-party (Table 5.4) frameworks, empirically showing the challenges that involves shifting from binary to multi-class inference.
(i) *Binary framework:* Looking at the results reported in Table 5.3, we notice that each user representation model effectively captures political orientation through the left-right categorization, yielding high-performance results that depend on the employed classifier. However, it is evident that models trained using RE representations demonstrate superior performance and consistency in all regions, regardless of the classifier used.
(ii) *Multi-party framework:* When analyzing the results presented in Table 5.4, it is evident that models trained using RE representations consistently demonstrate superior performance and stability across all regions. Among the models obtained with RE representations, SVM-linear classifiers obtain the best results. However, despite its popularity, the FA2 representations yield the lowest performance scores, indicating that they are the least suitable for this task. Both N2V and DW outperform FA2, but they are still surpassed by the models generated using RE representations.

As mentioned, FA2, DW and N2V can effectively capture information related to left-right orientation while they struggle when dealing with multi-class classifications. The notably superior results achieved within the binary framework

	EUS				GAL				CAT			
	N2V	DW	FA2	RE	N2V	DW	FA2	RE	N2V	DW	FA2	RE
LogReg	83.7	87.2	76.3	**96.0**	87.1	91.4	97.7	**98.2**	97.2	**99.1**	69.5	98.1
RF	82.7	87.3	83.0	**96.5**	91.9	97.1	**98.8**	98.2	96.2	**98.1**	89.7	96.2
NB	49.9	50.3	74.5	**93.0**	59.0	60.1	97.6	**98.2**	69.3	63.6	74.6	**98.1**
SVM-lin	80.5	85.5	77.2	**95.4**	85.0	89.4	**98.3**	98.2	96.2	**98.1**	73.7	97.2
SVM-pol	57.4	59.7	82.9	**93.9**	41.7	43.3	**98.8**	96.4	88.0	87.1	74.6	**98.1**
SVM-rbf	69.5	73.9	83.5	**95.1**	64.4	76.1	**98.3**	98.2	84.0	86.8	74.6	**98.1**
average	70.6	74.0	79.6	**95.0**	71.5	76.2	**98.3**	97.9	88.5	88.8	76.1	**97.6**

5.3 Table – BINARY FRAMEWORK (left-right). F1 macro score results for strongly supervised scenario (LOO CV) on EUS, GAL and CAT datasets. Algorithms used to generate the user representations: N2V (Node2vec), DW (DeepWalk), FA2 (ForceAtlas2), RE (Relational Embeddings). Values in **bold** represent best results for each classifier, while underlined values represent best overall results for each dataset.

	EUS				GAL				CAT			
	N2V	DW	FA2	RE	N2V	DW	FA2	RE	N2V	DW	FA2	RE
LogReg	63.7	66.1	33.3	**94.0**	42.2	56.4	46.6	**92.6**	86.7	88.0	33.6	**95.1**
RF	59.8	70.1	50.6	**92.8**	62.2	71.3	63.2	**90.1**	77.2	80.3	47.8	**91.6**
NB	38.5	41.5	42.6	**86.7**	46.4	46.8	50.3	**87.9**	39.5	53.9	63.1	**91.8**
SVM-lin	61.3	64.6	33.8	**93.1**	38.6	53.9	47.0	**92.8**	82.5	86.4	21.0	**96.2**
SVM-pol	36.1	37.1	42.2	**87.7**	06.2	08.4	47.3	**89.9**	59.4	71.1	07.1	**93.7**
SVM-rbf	39.6	42.1	48.6	**92.3**	20.5	28.1	47.4	**91.7**	52.2	59.2	18.7	**94.2**
average	49.8	53.6	41.9	**91.1**	36.0	44.2	50.3	**90.8**	66.3	73.2	31.9	**93.8**

5.4 Table – MULTI-PARTY FRAMEWORK (7 political parties). F1 macro score results for strongly supervised scenario (LOO CV) on EUS, GAL and CAT datasets. Algorithms used to generate the user representations: N2V (Node2vec), DW (DeepWalk), FA2 (ForceAtlas2), RE (Relational Embeddings). Values in **bold** represent best results for each classifier, while underlined values represent best overall results for each dataset.

compared to the multi-party framework illustrate that bipolar approaches to define political leaning lead to higher overall accuracy at the cost of essential nuances required to comprehend the specific political and social context. On the contrary, RE stands out as the most effective method for capturing finer-grained information related to more specific party-based political leanings, achieving high performance results in both the challenging multi-party context and the binary framework.

5.3.3 Experiment #2: Weakly Supervised scenario

Experimental Settings: Next we experiment in a more challenging, weakly supervised scenario, where the classifiers are provided with limited training data for each region in two different settings: (i) *One-shot*, where only one item per class is selected for training. The selected item for each class is manually selected, being the item representing each of the political parties. (ii) *Three-shot*, a few-shot setting where three items per class are selected for training. In this occasion, for each class we will select a single user corresponding to the political party, as well as two users representing the most referential candidates. The remainder of the users are left for the test set. In this scenario, the inference will be conducted at the political party level within the multi-party framework.

For this setting we only use the RE user representations and SVM-linear classifier, which achieved the best results at multi-party framework in the strongly supervised scenario (94.0 f1 macro score, Table 5.4). Additionally, we provide results obtained both with and without employing dimensionality reduction techniques. The dimensions have been reduced to 2, in accordance with prior research (Darwish *et al.* 2020; Stefanov *et al.* 2020), while the remaining hyperparameters are set to their default values. 3 different dimensionality reduction techniques are used for this purpose, such as, PCA, t-SNE (Van der Maaten and Hinton 2008) and UMAP (McInnes *et al.* 2018).

Results: As a second step, RE model is evaluated on the weakly supervised scenario to see the performance when considerably reducing training data. Table 5.5 demonstrates that 2 dimensional representations obtained from REs through t-SNE and UMAP dimension reduction techniques yield superior results compared to representations generated by REs without any dimensionality reduction for both one-shot and three-shot settings. PCA may not be the most appropriate method for dimension reduction in this context, as it is unable to retain information specific to each community and yields inferior results compared to using the full-dimensional representations. In contrast, UMAP and t-SNE reductions outperform full-dimensional representations as they may preserve community related information due to their architecture based on nearest-neighbours.

Furthermore, the results obtained by combining REs with UMAP and t-SNE dimension reduction techniques in the weakly supervised scenario (Table 5.5) outperform the performance of any other model from the strongly supervised scenario (Table 5.4). Specifically when REs are combined with UMAP, only one data point is necessary for training (one-shot) to outperform any other model of the strongly

	EUS			GAL			CAT		
Dim. Red.	LOO	3-shot	1-shot	LOO	3-shot	1-shot	LOO	3-shot	1-shot
none	**93.1**	*76.5	55.4	**92.8**	*81.4	***88.5**	**96.2**	***94.2**	*89.2
UMAP 2d	89.5	***90.0**	***81.6**	80.2	*83.5	*82.4	95.4	***94.2**	***95.1**
t-SNE 2d	90.5	*77.7	*72.9	85.1	***85.5**	*86.8	94.3	*93.3	***95.1**
PCA 2d	49.9	28.2	26.6	59.3	45.5	40.1	79.9	61.8	66.6

5.5 Table – F1 macro score results for SVM-linear classifier on multi-party framework with RE features on strongly (LOO CV) and weakly (3- and 1-shot) supervised scenarios. Values with * represent when the results of REs with 1- or 3-shot training are higher than the best overall result of non RE methods for multi-party framework on strongly supervised scenario: DW with RF for EUS (*70.1*) and GAL (*71.3*); DW with LogReg for CAT (*88.0*).

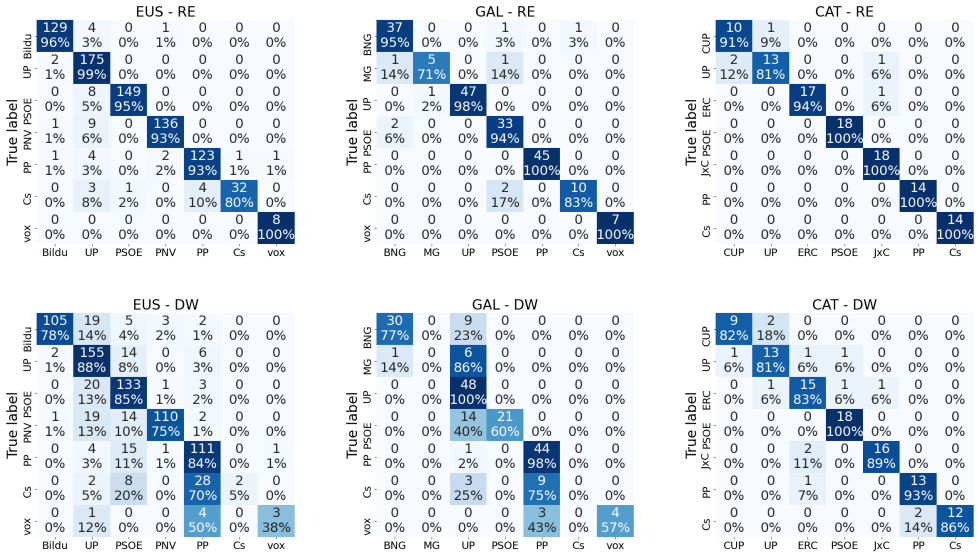
supervised scenario in more than 10 points at EUS and GAL and 7 points in CAT. We confirm that compressing RE representations into 2 dimensional features with UMAP or t-SNE is a good strategy to handle situations with very few annotated data, obtaining similar scores to those of RE on the strongly supervised scenario. Moreover, the results demonstrate consistency across all three regions, indicating that RE representations reach competitive and robust results even where only the user belonging to the political party is annotated.

5.3.4 Discussion

In this section, we will delve into a deeper analysis of the reported results by conducting an error analysis and generating visualizations of the user representations for the best method.

Error analysis: In order to understand the considerable differences in performance among user representation methods, we are conducting a detailed comparison of the best and second best methods among themselves. The confusion matrices presented in Figure 5.4 report the errors performed by the Logistic Regression classifier using RE and DW user representation models for multi-party framework at strongly supervised scenario.

In relation to EUS dataset, RE (Figure 5.4 top left) model shows incidental classification errors that occur among classes such as Bildu-UP, PP-Cs and PSOE or PNV as UP. This errors take place among parties that are ideologically close, showing that the model fails between simmilar classes. On the other hand, DW



5.4 Figure – Confusion matrices for Logistic Regresion classifier using RE (top) and DW (bottom) user representation models in the strongly supervised scenario on EUS (left), GAL (center) and CAT (right) datasets.

(Figure 5.4 bottom left) model show considerable errors among UP, PSOE and PNV, failing to classify users arround this orientations. Moreover, DW generally fails to classify right-wing unionist party members, grouping them all as PP users.

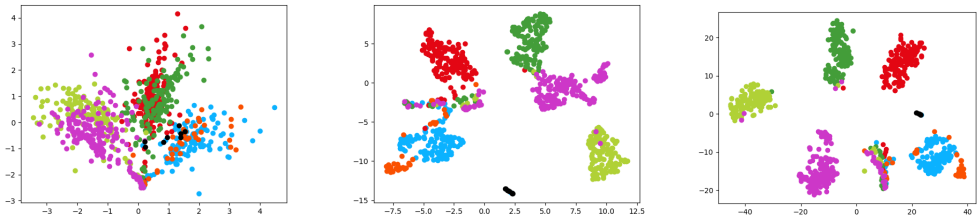
When considering GAL dataset, DW (Figure 5.4 top center) model misclassifies BNG, MG and PSOE users as UP members, grouping many left-wing representatives under UP political orientation. Furthermore, some users from vox and Cs are classified as PP users, being the representative of the right-wing. DW model seems to underrepresent political options in a left-right dichotomy, simplifying political complexity. On the other hand, RE (Figure 5.4 bottom center) model has very few classification errors occurring among ideologically similar orientations, demonstrating the capacity of RE's to represent parties as well as political orientations.

Finally, with respect to CAT dataset, we can see that RE (Figure 5.4 top right) gets better results than DW (Figure 5.4 bottom right), however both models achieve high performance despite minor deviations. This can be attributed to factors such as the smaller dataset size or balanced classes contributing to their outstanding performance.

Summarizing, classification errors usually take place among parties that are ideologically close, showing that Retweet-based DW and RE models can capture general ideological tendencies. This conclusion aligns with the high results obtained by DW within the binary framework, confirming these models' ability to capture general ideological traits. However, RE models show a high accuracy with a very low error rate, demonstrating that this model can also capture previously specified political parties besides the general ideological alignments.

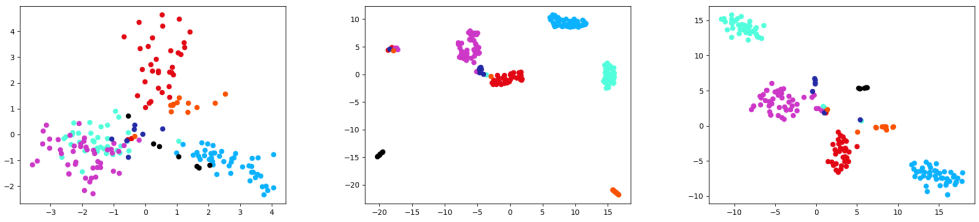
Data visualization: In order to better understand the effectiveness of the RE user representation techniques, we visualize labeled users' RE representations for the three regions, EUS, GAL and CAT by performing PCA, UMAP and t-SNE dimensionality reduction as done in weakly supervised scenario. This visualization assists us in conducting a qualitative evaluation of RE user representation method, correlating common-sense political knowledge to our mathematically constructed model.

***EUS** (Figure 5.5):* In EUS dataset UMAP (Figure 5.5 center) and t-SNE (Figure 5.5 right) visualizations have similar groupings with clear clusters for specific political organizations, which are arranged based on their political proximity. Thus, the centered positions of the graph are taken by the parties in the Basque government, formed by PNV (●) and PSOE (●). The leftist positions are represented by UP (●) and Bildu (●), located on one side. Whereas right winged positions represented by PP (●) and Cs (●), are pictured on the other side, while alt-right Vox (●) is represented as an outlier. PCA visualization (Figure 5.5 left) is fuzzier but able to group users on 3 different groups corresponding to the previous global views; PNV and PSOE representing centered positions; Bildu and UP representing the left and progressives; PP, Cs and Vox representing the right and the conservatives.



5.5 Figure – Visualization of PCA, UMAP and t-SNE 2 dimension reduction for EUS Relational Embedding representation.

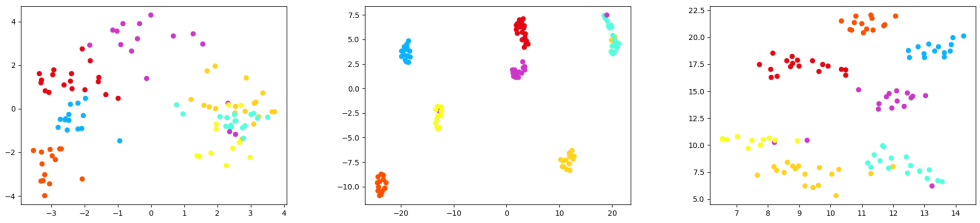
***GAL** (Figure 5.6):* In this dataset t-SNE visualizations (Figure 5.6 right) show clear clusters for the political parties, drawing them on a singular axis. On one end we have BNG (●) representing a pro-independence left-wing followed by UP (●) and MG (●) representing the left; PSOE (●) and Cs (●) represent centered positions next to PP (●) representing the right and the conservatives on the oposite end. PCA visualization (Figure 5.6 left) also groups the users on 3 different groups mixing political parties into a more simple layout; UP, BNG and MG as left and progressives; PSOE and Cs grouped on centered positions; PP (●) and Vox (●) grouped as right and the conservatives.



5.6 Figure – Visualization of PCA, UMAP and t-SNE 2 dimension reduction for GAL Relational Embedding representation.

***CAT** (Figure 5.7):* In CAT dataset PCA (Figure 5.7 left) and t-SNE (Figure 5.7 right) visualizations provide symbolic representations of the political reality, wherein political parties are positioned close to each other based on their stance regarding the independence process. On one side, parties like Cs (●) and PP (●), which strongly advocate for unity with the Spanish kingdom, are clustered in the left (PCA) or top (t-SNE) portions of the plot. Conversely, pro-independence parties such as CUP (●), JxC (●), and ERC (●) are clustered on the right (PCA) or bottom (t-SNE) side. In more central positions, UP (●) and PSOE (●) are situated between both groupings, serving as a connecting links bridging the divide between the two sides.

It is evident that the dataset size has an impact on the visualizations, with representations becoming fuzzier as the dataset size increases. Being the biggest dataset, EUS exhibits the largest degree of fuzziness, while CAT is the smallest dataset and has the most defined communities. Additionally, the clarity of the communities depending on the dimensionality reduction technique is in accordance with the results achieved by these dimension reduction techniques on the weakly supervised scenario. On the one hand, the PCA dimension reduction



5.7 Figure – Visualization of PCA, UMAP and t-SNE 2 dimension reduction for CAT Relational Embedding representation.

technique does not clearly define the communities related to the specific political parties. However, it can effectively display information related to more general political orientations by clustering ideologically similar parties closely together in the same euclidean space. On the other hand, in the graphs arisen from UMAP and t-SNE dimension reduction techniques, it can be seen that the communities are clearly defined and situated depending on their ideological similarities. Regardless of the dimension reduction technique or the dataset size, REs can effectively represent the political communities as well as the ideological similarities and disparities among them. These findings suggest that RE user representations have the capacity to embed knowledge about the socio-political environment leaning on retweet based user interactions.

5.4 Different levels of political engagement

We have demonstrated the efficacy of incorporating multi-party political leaning to create a more comprehensive representation of political nuances adaptable to diverse regions, even in situations with limited training data. The success achieved with Relational Embeddings motivates us to extend our methodology to more challenging and realistic contexts. Our subsequent objective involves analyzing different levels of political engagement within the previously described multi-party and multi-region frameworks. Political participation encompasses different levels of involvement (Almond and Verba 2015) as in democratic systems citizens have to express their demands, and these expressions require at least some minimum level of engagement (Van Deth *et al.* 2007). Hence, political involvement spans a continuum, ranging from passive observers to highly engaged activists, each harboring distinct perspectives and motivations. Delving into research that

considers these different levels of involvement facilitates a nuanced understanding of political behavior. Our focus lies in exploring the ability to infer the political leaning of users with different levels of engagement, providing a more realistic assessment of our methods. To systematically investigate different degrees of political involvement, we have generated a new dataset spanning three distinct regions in the UK. This dataset includes users with diverse levels of political engagement, such as members, supporters, and sympathizers, labeled according to their alignment with a specific political party. This research facilitates a more granular analysis of political landscapes, enhancing our comprehension of the multifaceted nature of political leaning by incorporating different levels of political involvement. Our goal remains the discovery of stable and robust user representation methods capable of capturing socio-political information for further classification. This includes multi-party political leaning across diverse regions and realistic scenarios that consider different levels of political engagement. Consequently, we evaluate user representations for their effectiveness in inferring multi-party political leanings within three different regions in the UK. We also explore challenging scenarios such as few-shot learning, where training data is scarce, and assess the applicability of our methods to users with different levels of political engagement.

5.4.1 Datasets: United Kingdom

In order to extend research on multi-party frameworks into different political landscapes, we create additional datasets. Thus, we select new political regions from the United Kingdom and the most relevant political parties for each of them. Afterwards, we follow similar data extraction techniques as in Section 5.3.1 adding specific data collection in order to include different political leanings.

Political Region

Given our interest in analyzing the socio-political context of the United Kingdom as a multi-party system (Lynch 2007), our study focuses on political parties in Scotland (5.5M citizens), Wales (3.1M) and Northern Ireland (1.8M) during autumn of 2022. These regions form politically diverse contexts, each with its own devolved government and strong nationalist sentiments that foster many political options. The UK's political landscape has evolved substantially in recent decades, from being dominated by two parties in the 1950s (Conservative and Labour parties attaining over 95% of the votes) to a more diverse, multi-party landscape (75% across both parties in 2019).

Scotland (SCT): *Scottish National Party* (SNP ●) is a Scottish nationalist and social democratic political party; positioned on the center-left, pro-independence and pro-European. *Scottish Conservative & Unionist Party* (SCU ●) is a conservative party in Scotland, Nationally affiliated with the Conservative Party; center-right and unionist. *Scottish Labour Party* (SL ●) is a Scottish social democratic political party, an autonomous section of the UK Labour Party; considered to be center-left and unionist. *Scottish Green Party* (SGP ●) is a Scottish green political party, affiliated with the Global Greens and associated mainly with environmentalist policies; positioned on the left, pro-independence and pro-European. *Scottish Liberal Democrats* (SLD ●) is a Scottish liberal and federalist political party, part of the United Kingdom Liberal Democrats; positioned on the political center, pro-European and unionist.

Wales (WAL): *Welsh Labour* (WL ●) is a Welsh social democratic political party, and formally part of the UK Labour Party; center-left and unionist. *Welsh Conservatives* (WC ●) is a conservative party in Wales, a branch of the UK's Conservative Party; ideology is center-right and unionist. *Plaid Cymru* (PC ●) is the principal Welsh nationalist political party; positioned on the left and pro-independence. *Welsh Liberal Democrats* (WLD ●) is a Welsh liberal and federalist political party, branch of the UK's Liberal Democrats; positioned on the political center, pro-European and unionist.

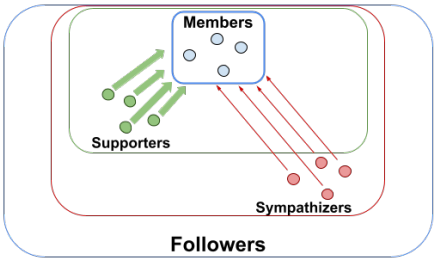
Northern Ireland (NIR): *Sinn Féin* (SF ●) is an Irish republican and democratic socialist political party; considered to be left-wing, pro-unification and pro-independence. *Democratic Unionist Party* (DUP ●) is a conservative and loyalist political party in Northern Ireland; positioned on the right-wing and unionist. *Alliance Party of Northern Ireland* (APNI ●) is a liberal political party in Northern Ireland, aligned with the UK's Liberal Democrats; positioned on the political center, pro-European and they consider themselves outside of Nationalism and Unionism. *Ulster Unionist Party* (UUP ●) is a unionist political party in Northern Ireland, a branch of the UK's Conservative Party; positioned on the center-right and unionist. *Social Democratic and Labour Party* (SDLP ●) is a social-democratic and Irish nationalist political party in Northern Ireland; center-left and pro-Irish.

Data collection strategy

We have developed a generalizable data collection methodology for multi-party regions, expanding upon the previous proposal (Section 5.3.1) to encompass different levels of political engagement. Our methodology was tested in the UK but

can be adapted for use in other regions. Once the regions of interest and the relevant political parties have been identified, our methodology consists of three steps to collect: (i) an initial seed of users (members), (ii) other users with different levels of engagement or interest (supporters and sympathizers) and, (iii) interactions and timelines pertaining to those users. Data collection was done between September and October 2022.

(1) **Manual labeling:** In line with data collection strategy followed in the previous section (5.3.1), we start by collecting an initial seed of users. For each of the political parties in our datasets, we identify party members with Twitter accounts including members of parliament (MPs) or members of regional parlaments (MSPs, MSs and MLAs). This leads to a collection of users where each user is linked to a specific region and party (details in Table 5.6, column 'Members').



5.8 Figure – Creation scheme for Supporter and Sympathizer evaluation sets. Supporters: more engaged, following 5 or more member users. Sympathizers: less involved, following up to 2 member users.

(2) **Snowball collection of friends and followers:** To expand the datasets beyond direct members of the party, we look for less engaged users including supporters and sympathizers (see Figure 5.8). We rely on *follower* connections with member users as a proxy to collect less engaged users (Barberá 2015; Xiao *et al.* 2020), where the level of engagement or political interest is determined by the number of members they follow (Xiao *et al.* 2020). Engaged users with a strong interest in politics are referred to as "supporters" if they follow 5 or more members of a party. On the other hand, users with less vested interest, such as those who follow 2 or fewer members, are called "sympathizers". The specific thresholds of 2 and 5 were empirically determined by looking at the frequencies in the data so that we could obtain a balanced number of supporter and sympathizer user groups. We retrieve up to 100 users per party for each of the supporter and sympathizer groups, filtering out those for which interaction data (see step 3) is not available, leading to the counts shown in the columns "Supporter" and "Sympathizer" of

Table 5.6.

Region	Party	Member	Supporter	Sympathizer
SCT	SNP	184	96	85
	SCU	59	97	84
	SL	52	95	86
	SGP	42	99	88
	SLD	24	98	94
	total	361	485	437
WAL	WL	55	97	88
	WC	42	98	85
	PC	42	99	85
	WLD	27	100	95
	total	166	394	353
NIR	SF	80	98	63
	DUP	65	75	67
	APNI	52	83	79
	UUP	58	73	68
	SDLP	59	76	72
	total	314	405	349

5.6 Table – Labeled users from UK for each region. Manually labeled Member users and automatically labeled Supporter and Sympathizer users, by region and class.

(3) **Twitter data retrieval:** For every user in the member, supporter and sympathizer groups we retrieve a history of their retweet interactions, regardless of whether these interactions are with users included in our datasets (see Figure 5.3). For every user in the member, supporter and sympathizer groups we retrieve a history of their retweet interactions, following the same method as in the previous section (5.3.1). Table 5.7 shows the final statistics of retweets retrieved and the number of total users performing those interactions. Data collection was performed during 2022 autumn.

	SCT	WAL	NIR
Member users	361	166	314
Supporter users	485	394	405
Sympathizer users	437	353	349
Interacting users	87k	62k	21k
Retweets	19M	21M	4M
All users	937k	933k	426k

5.7 Table – Final dataset composition for each UK region.

5.4.2 Experiment #1: Strongly Supervised scenario

Experimental Settings: To compare the performance of various interaction-based user representation methods on a strongly supervised setting, we will employ the same settings as used in Section 5.3.2. Therefore, for each region, we will experiment with different classifiers using a leave-one-out (LOO) cross-validation setting for the users within the member group in order to infer the political party they align with.

	SCT				WAL				NIR			
	N2V	DW	FA2	RE	N2V	DW	FA2	RE	N2V	DW	FA2	RE
LogReg	71.5	68.0	31.0	**99.4**	55.2	62.8	24.6	**99.2**	51.5	64.8	28.7	**97.7**
RF	81.8	78.2	63.5	**99.2**	67.5	78.2	68.9	**96.2**	66.4	80.9	76.3	**97.4**
NB	37.5	35.5	51.1	**99.5**	30.3	32.7	42.3	**98.7**	28.0	32.1	34.0	**97.7**
SVM-lin	73.0	69.0	30.6	**99.4**	33.2	54.7	37.9	**96.4**	35.4	46.9	41.2	**97.7**
SVM-pol	39.7	43.2	30.0	**96.4**	22.2	25.9	12.6	**96.4**	08.1	10.3	10.3	**94.5**
SVM-rbf	41.3	43.1	61.6	**99.9**	26.1	28.6	38.4	**98.6**	17.8	29.7	50.6	**97.4**
average	57.5	56.2	44.6	**98.9**	39.1	47.2	37.5	**97.6**	34.5	44.1	40.2	**97.0**

5.8 Table – Results for Strongly Supervised scenario. F1 macro score results leave-one-out CV on SCT, WAL and NIR member datasets. Algorithms used to generate the representations: N2V (Node2vec), DW (DeepWalk), FA2 (ForceAtlas2), RE (Relational Embeddings). Values in **bold** represent best results for each classifier.

Results: Looking at the results reported in Table 5.8, we observe that the models trained with RE representations achieve best results for all the classifiers across every region. Among the models trained with RE representations, Logistic Regression consistently obtains the best results. On the other hand, and despite its popularity, the FA2 representations lead to the lowest performance scores, showing that it is the least suitable for this task. Both N2V and DW are clearly better than FA2, but still are clearly outperformed by the models obtained with RE representations. The RE method also behaves more robustly in multi-party scenarios and across regions. Finally, N2V, DW and FA2 show substantial variability across the different regions, while RE is the most stable method, showing robustness and adaptability.

5.4.3 Experiment #2: Weakly Supervised scenario

Experimental Settings: To do experimentation on a weakly supervised scenario, we follow same settings as used in Section 5.3.3, using one-shot and three-

shot approaches. For this setting we only use the RE user representations and Logistic Regression method, which was the best combination in Experiment #1 above. Furthermore, we also provide results obtained with and without dimensionality reduction techniques.

	SCT			WAL			NIR		
Dim. Red.	LOO	3-shot	1-shot	LOO	3-shot	1-shot	LOO	3-shot	1-shot
none	99.4	71.8	*91.9	**99.2**	*96.9	***98.5**	**97.7**	*97.0	*94.6
UMAP 2d	**99.9**	*91.2	*90.8	98.5	***97.8**	*94.0	97.3	***97.2**	*94.7
t-SNE 2d	99.4	***95.3**	***92.2**	98.5	***97.7**	*96.2	97.0	*94.8	***94.9**
PCA 2d	87.3	71.2	67.4	93.9	*89.5	75.3	69.9	65.5	49.8

5.9 Table – Results for Weakly Supervised scenario. F1 macro score results on *Member* datasets using logistic regression classifier with RE features on strongly (LOO CV) and weakly (3- and 1-shot) supervised scenarios. Values with * represent when the results of REs with 1- or 3-shot training are higher than the best overall result of non RE methods for multi-party framework on strongly supervised scenario: N2V with RF for SCT (*81.8*); DW with RF for WAL (*78.2*) and NIR (*80.9*).

Results: Table 5.9 shows that 2 dimensional representations derived from t-SNE and UMAP dimension reduction technique get better results than RE without any dimensionality reduction for one-shot and few-shot settings. Results with PCA reduction are substantially worse across evaluation settings and regions. Compressing RE user representations into 2 dimensional representations with UMAP or t-SNE can be a good solution to handle community detection on weakly supervised scenarios as they can highlight communities due to their architecture based on nearest-neighbours. Interestingly, despite being evaluated on few-shot and one-shot settings, these methods obtain scores similar to those of RE on the strongly supervised scenario. Furthermore, results are consistent across the 3 regions, showing that RE representations reach competitive and robust results even in weakly supervised scenarios.

5.4.4 Experiment #3: Realistic scenario

Experimental Settings: We define a more realistic, challenging scenario, in which we test the ability of the interaction-based models to predict the political leaning of less engaged users, namely, of supporters and sympathizers. This assessment can offer insights into the model's applicability in a real-world context,

where individuals may not be directly affiliated with political parties and have different levels of attachment. In order to do this, we use *members* for training and *supporters* and *sympathizers* for testing. We break down the performance of the models for each of the groups –supporters and sympathizers– to evaluate the impact of the level of political engagement of users to infer political leaning. For these experiments we use the two best overall classifiers in the strongly supervised scenario: Logistic Regression and Random Forest.

	SCT				WAL				NIR			
SUP	N2V	DW	FA2	RE	N2V	DW	FA2	RE	N2V	DW	FA2	RE
LogReg	23.0	24.7	21.6	**90.8**	38.8	48.2	27.2	**95.3**	21.2	20.9	31.1	**75.4**
RF	50.8	44.1	40.3	**81.4**	50.0	59.9	56.1	**93.8**	30.5	33.3	36.8	**65.4**
SYM	N2V	DW	FA2	RE	N2V	DW	FA2	RE	N2V	DW	FA2	RE
LogReg	8.3	09.4	18.0	**63.3**	19.8	17.6	23.3	**60.6**	6.5	6.5	21.6	**41.2**
RF	22.8	20.7	25.6	**51.7**	17.8	17.2	23.8	**60.2**	9.3	7.7	26.3	**38.1**
avg.	26.2	24.7	26.4	**71.8**	31.6	35.7	32.6	**77.5**	16.9	17.1	28.9	**55.0**

5.10 Table – Results for Realistic scenario. F1-score results on SCT, WAL and NIR Supporter (SUP) and Sympathizer (SYM) users datasets. Algorithms used to generate the user representations: N2V (Node2vec), DW (DeepWalk), FA2 (ForceAtlas2) RE (Relational Embeddings). Values in **bold** represent best results for each classifier.

Results: Table 5.10 shows that the performance for this scenario is considerably lower. Overall, supporter users get better results than sympathizer users, showing that models suffer more when trying to learn political leaning of users that do not engage so much with political parties (which is only natural, in a way). N2V and DW fail to produce satisfactory results for this realistic scenario. We hypothesize that random walks may generate noisy or irrelevant paths that can negatively affect the quality of the embeddings. The FA2 method also fails to infer political leaning, showing that the use of a two-dimensional vector space for the approximation-repulsion process is insufficient to embed complex socio-political information. In any case, as it has been the case for previous experiments, the pair-based RE user representations are significantly better for every evaluation setting across every region. All the methods without exception show higher results on WAL datasets (4 classes) than on SCT and NIR datasets (5 classes). It seems that including one class more in a multiclass approach to political leaning inference has negative influence to classify users which are less engaged in politics (supporters and sympathizers).

5.4.5 Discussion

In this section we discuss the analysis of political leaning in light of the reported results. We also consider different ways of making our results explainable and provide an error analysis to identify possible weaknesses of our approach.

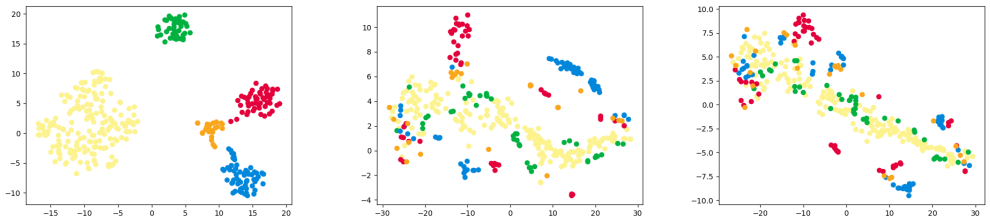
Data Visualization: In order to better understand and explain the effectiveness of the different user representation techniques, we visualize RE, N2V and DW user representations for the three regions, SCT, WAL and NIR by performing t-SNE dimensionality reduction into 2 dimensions.

The first noticeable point looking at Figures 5.9, 5.10 and 5.11 is that, in contrast to N2V and DW, the visualizations obtained from the RE representations are clearly able to discriminate the multiparty political communities represented by the member users for each of the countries. In fact, the clear communities obtained in the visualization of the RE user representations is arguably in accordance with them outperforming other methods in the experimental evaluations reported in Tables 5.8 and 5.10.

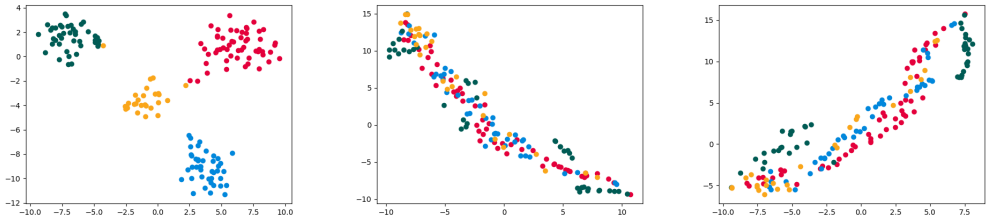
Taking a closer look at the SCT visualization (left plot in Figure 5.9), we can see parties represented by clearly distinguishable communities. Thus, SNP (●) takes a big part of the figure, mainly isolated from the others. Unionist parties are forming their own cluster at the right side of the chart, separated from the other two parties. Inside the unionist community, it can be seen that SLD (●) acts like a link between SCU (●) and SL (●), showing its position as a central political actor. SLD also takes a central role in the which highlights its centrist political outlook. The representation locates SGP (●) apart from SNP and the unionists but between SNP and SL, showing a proximity to those. Furthermore, it can be seen that the pro-independence (SNP and SGP) and unionist parties (SL, SLD and SCU) are represented in different positions, showing a high polarization in the dispute across national identities.

Moving on to Figure 5.10 which shows the visualization obtained for WAL, it is possible to note that WLD (●) is situated in the center of the political spectrum, which is interesting as in reality they are considered to be positioned in the political center. The other 3 parties are surrounding WLD, but WL (●) is between PC (●) and WC (●), showing that the last two are the most opposed poles (left/right or pro-independence/unionist).

Finally, for NIR we can see in Figure 5.11 that political parties are grouped in their own clusters, except for a few instances that may have been incorrectly classified. These few errors seem to align with to those causing slightly lower



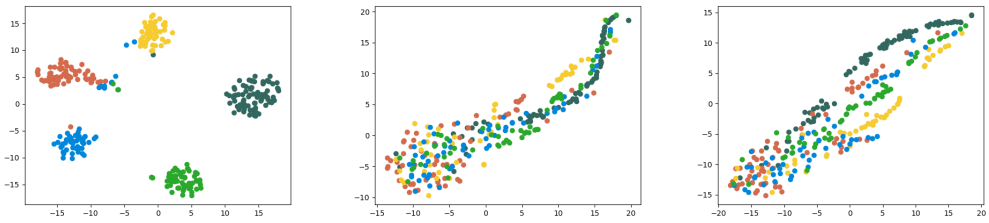
5.9 Figure – Visualization of t-SNE 2 dimension reduction of Relational Embeddings (left), node2vec (center) and Deep Walk (right) representations for SCT Member users.



5.10 Figure – Visualization of t-SNE 2 dimension reduction of Relational Embeddings (left), node2vec (center) and Deep Walk (right) representations for WAL Member users.

performance, as shown in Table 5.8, of the RE user representations for NIR when compared to SCT and WAL. Looking at the positions of the parties, we can see that DUP (●) is located next to UUP (●) forming a conservative and unionist pole. Besides, SF (●) and SDLP (●) define the left-wing and pro-Irish pole. As a centralist political actor APNI (●) is located between both main groups, but much closer to the conservative-unionist pole forming with them a wider liberal-conservative, right-wing pole at the left of the chart. Summarizing, we believe that REs capture well multiple ideological disparities (left/right or pro-Irish/unionist) among these parties.

The ability of RE user representations to depict distinct communities in all three cases is noteworthy, especially when compared to the inability of N2V and DW. Furthermore, REs are able to locate the communities following a pattern of ideological similarities and disparities. These observations lead us to infer that RE user representations have the potential to embed socio-political information within the generated features.

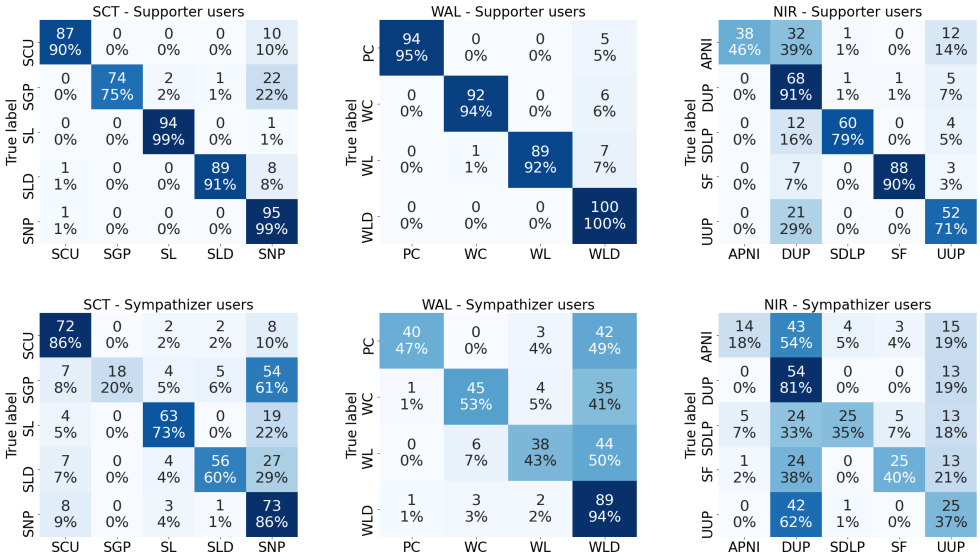


5.11 Figure – Visualization of t-SNE 2 dimension reduction of Relational Embeddings (left), node2vec (center) and Deep Walk (right) representations for NIR Member users.

Data quantity: Relating the results with the number of the interactions collected in the datasets, we can see that there are some similarities among the user representation methods. The RE user representation method has better results for strongly supervised approach at SCT (19M RTs) and WAL (21M RTs) comparing to a small drop for NIR (4M RTs) which contains considerably less interactions (Table 5.7). The same occurs in the weakly supervised scenario, in which SCT and WAL obtain better evaluation results in comparison to NIR. This not only occurs with REs, but also with other user representation methods such as N2V and DW. The take out message is seems to be that the larger the number of interactions the better the results. The results for the realistic scenario reported in Table 5.9 further confirm this trend. Thus, the larger the timelines and the amount of users from which to extract the retweets, the better representations we get for all the user representation methods.

Error analysis: Considering the almost perfect scores obtained by REs on the strongly and weakly supervised scenarios, our error analysis will focus only on the realistic scenario, which consists of users less engaged politically, namely, supporters and sympathizers. The confusion matrices presented in Figure 5.12 report the main errors performed by the classifiers based on REs on each particular region.

With respect to SCT, it can be observed that most misclassified instances correspond to models predicting SNP instead of the correct label. Thus, for supporter users (Figure 5.12 top left), 22% of the errors in predicting SGP users (75% acc.) are wrongly predicted as SNP. This is even more pronounced for sympathizer users (Figure 5.12 bottom left) given that the classifiers performs substantially worse in this evaluation setting. Thus, 61% of the errors in classifying SGP cor-



5.12 Figure – Confusion matrices of Supporter (top) and Sympathizer (bottom) users of Logistic Regression trained with RE user representations for SCT (left), WAL (center) and NIR (right).

respond to the model predicting SNP instead. We hypothesize that this may be explained by the fact that SNP and SGP have certain ideological similarities and have been in a cooperation agreement since 2021.

The model trained for WAL, compared to the SCT case, has a lower error rate when predicting supporter's political leaning (Figure 5.12 top center). There, the few errors correspond to predicting WLD instead of the correct classes. If we look at the sympathizers errors (Figure 5.12 bottom center), they substantially amplify the pattern seen for in the supporters setting. We believe that the source of errors may be caused by the central role played by WLD in Wales's politics and by the plurality in policies across the political spectrum.

With respect to NIR, we can see that the model has more problems discriminating between the different political options. If we look at the confusion matrix for supporters (Figure 5.12 top right), DUP gets 32% of UUP and 39% of APNI instances. Moreover, 12% of the APNI errors correspond to UUP, which shows that the model is not able to detect well APNI users. This might be due to the centralist and liberal profile of APNI, which makes it difficult to discriminate from other political options. The same pattern can be observed for the sympa-

thizer users (Figure 5.12 bottom right) although amplified by the larger number of classification errors. It is particularly interesting the difficulties of the model in distinguishing UUP and APNI from DUP, which seems to indicate that the DUP is seen as the main reference of the right unionist space.

The confusion matrices in the realistic scenario show that the best model suffers to clearly classify multi-party political leaning especially for users less engaged politically (sympathizers). It is noticeable the large amount of errors when trying to discriminate UUP and APNI from DUP, which is indicative perhaps of the dominance by DUP of the right wing agenda. In the case of WAL, the main source errors seem to be due to the centralist position of WLD with respect to other political options. Finally, in SCT we can see the phenomenon of a smaller party (SGP) cooperating with a larger one (SNP) and getting assimilated as a result of this collaboration. Summarizing, our political party-based outlook may capture sociopolitical information since errors commonly occur among ideologically adjacent classes and increase when targeting less politically engaged users. In any case, we confirm the necessity of addressing political leaning as a multipolar classification task which, despite being more difficult, would provide a more representative analysis of the social reality.

5.5 Hybrid text-interaction modeling for political leaning inference

Previous approaches consider political leaning inference in a multi-party multi-region framework, although they are limited to the application of social interaction data. Hence, our objective is to represent individual actors leveraging different data sources, including published texts, alongside social media interactions. This facilitates political leaning inference even in scenarios where no interactions are available, such as news media or political speeches. By doing so, we aim to incorporate wider and more accurate representations into the ongoing exploration of public opinions for diverse social science studies, including hate speech, disinformation or propaganda detection (Akhtar *et al.* 2019; Hristakieva *et al.* 2022).

In the previous section (Section 5.4), we showed that the use of features derived from interactions between users can lead to high performance in inferring the political leaning of social media users that are actively engaged in politics. However, a pure interaction-based approach proved to have clear limitations in making predictions for users who are less engaged in politics, hence hindering

the broader applicability of the approach. To achieve this, we introduce a hybrid approach, which integrates textual and interaction-based information into a hybrid model for improved political leaning inference across broader groups of social media users of varying degrees of political engagement. In doing so, we present the first attempt at addressing political leaning inference across a range of multiclass political realities fusioning text and interaction data.

Our proposed hybrid approach is flexible in that it can incorporate different encoding techniques. We explore with different hybrid methods combining text-based approaches like TF-IDF, word2Vec (Mikolov *et al.* 2013a), and Transformers (Vaswani *et al.* 2017), with interaction-based methods such as DeepWalk (Perozzi *et al.* 2014), Node2vec (Grover and Leskovec 2016) and Relational Embedding (Fernandez de Landa and Agerri 2022). We evaluate the resulting hybrid models in three datasets pertaining to three different regions of the UK, each with different political parties. These datasets include textual data specific to each user, in addition to the interaction data collected thus far. We study the performance of our models on users with different levels of engagement in politics, with a particular focus on users with lower levels of engagement, whose posting of political content and interactions with content and users relevant to politics are predicted to be less frequent, which poses an additional challenge.

5.5.1 Datasets: United Kingdom hybrid

In this phase, our aim is to gather data to characterize labeled Twitter users based on their text and interactions. To achieve this, we utilize the dataset that we have previously developed in Section 5.4.1. This dataset comprises users from various regions of the United Kingdom, each with differing levels of political engagement. These users are annotated according to the political party they align with, along with associated interaction-based data. In order to incorporate textual data to the mentioned dataset, we collected 120 tweets per Member user and 60 tweets per Supporter and Sympathizer user. To ensure a balance between the number of users and the available data, we excluded labeled users with insufficient data and discarded tweets containing fewer than 10 tokens. This process resulted in a reduction of users from the previous dataset (Section 5.4.1), as some users did not meet the specified criteria, leading to a different distribution (see Table 5.11).

Twitter data extraction was undertaken during October 2022, collecting the timelines of all the identified users. In Table 5.12 it can be seen the shape of the final corpus for each region: (i) Member users and gathered text-based data (120 tweets per user); (ii) Supporter and Sympathizer users and the corresponding

Region	Party	Member	Supporter	Sympathizer
SCT	SNP	181	91	74
	SCU	59	86	81
	SL	52	88	72
	SGP	42	82	77
	SLD	24	90	84
	total	358	437	388
WAL	WL	55	92	77
	WC	42	91	75
	PC	42	91	72
	WLD	27	98	81
	total	166	372	305
NIR	SF	79	92	37
	DUP	61	44	54
	APNI	52	62	66
	UUP	57	52	48
	SDLP	58	54	65
	total	307	304	270

5.11 Table – Labeled users from UK with text and interaction data.

text-based data (60 tweets per user); (iii) interaction-based retweet data.

		SCT	**WAL**	**NIR**
Members	users	358	166	307
	tweets	42,960	19,920	36,840
	tokens	1,400k	789k	653k
Supporters	users	437	372	304
	tweets	26,220	22,320	18,240
	tokens	654k	676k	497k
Sympathizers	users	388	305	270
	tweets	23,280	18,300	16,200
	tokens	1,194k	523k	436k
Interactions	interacting users	87k	62k	21k
	retweets	19M	21M	4M

5.12 Table – Final text and interaction dataset composition for each UK region.

5.5.2 Experimental Setup

We leverage the interaction-based, text-based and hybrid user representations in order to conduct political leaning inference. We do so in two different sets of

experiments. First, we focus on determining the optimal configuration to obtain good quality textual representations. Second, we will apply the best textual representations, together with interaction-based representations, in the hybrid approach, evaluating them on the Member, Supporter and Sympathizer datasets across the 3 UK regions.

	SCT				WAL				NIR			
Dims.	50	100	200	300	50	100	200	300	50	100	200	300
tfidf	22.4	39.3	48.4	57.2	45.3	54.8	64.2	64.8	30.3	40.4	55.5	60.2
w2v	57.9	62.3	63.5	64.0	61.8	61.3	61.3	64.0	54.2	56.3	57.5	57.9

5.13 Table – F1 macro score results 10 fold CV on SCT, WAL and NIR Members datasets. Algorithms used to generate the features: tfidf and w2v. Underlined values represent best result for each algorithm in each dataset.

	SCT				WAL				NIR			
Transformers	B	dB	R	Rt	B	dB	R	Rt	B	dB	R	Rt
start-of-sequence	35.9	33.7	07.2	37.2	55.7	55.3	41.0	56.4	45.6	42.0	10.4	39.0
average	54.2	48.6	12.7	39.7	68.4	64.0	52.0	56.1	56.7	49.3	20.1	40.1
max-pool	67.6	73.4	47.6	60.7	75.6	75.1	58.0	64.4	68.4	72.1	50.7	57.4

5.14 Table – F1 macro score results for 10 fold CV on SCT, WAL and NIR Members datasets. Algorithms used to generate the features: mBERT (B), DistilmBERT (dB), XLM-RoBERTa (R), XLM-T (Rt). Underlined values represent best results for each algorithm in each dataset.

Textual Methods Selection: For each region, we experiment with the users in the group of *Members* using a 10 fold cross-validation (CV) setting, i.e., a 10% of the users are left for evaluation while all the others are used for training. The primary objective of this experiment is to compare the performance of text-based user representation methods. The representations obtained using the different methods are used to train a RBF-kernel Support Vector Machine (SVM) classification algorithm, following the same idea as in Chapter 4. We use the scikit-learn implementation (Pedregosa *et al.* 2011) with default configuration.

With respect to tfidf and w2v, different dimension values were tested, as shown in Table 5.13. Ultimately, the best dimension value for both methods is set to 300. Particularly, for tfidf, as the dimension value increases, the results tend to improve. Regarding Transformer-based methods, the results in Table 5.14 demonstrate that

max-pooling proves to be the most effective strategy. The best textual user representation methods are used to be compared to and combined with interaction-based user representation methods.

Experimental Settings: We will extract representations from texts and interactions to train independent user classification models for each of the political regions: SCT, WAL and NIR. Each of the regions will have its own user representations to observe the performance of the methods on different scenarios.

In order to compare the quality of our proposed user representation methods considering the level of political engagement of the users, we evaluate separately on the Member, Supporter and Sympathizer datasets. On the one hand, we train and evaluate a SVM (RBF kernel) classifier (Pedregosa *et al.* 2011) using 10 fold CV with the Members dataset. On the other hand, we use the Members dataset to train another SVM (RBF kernel) classifier and then evaluate the model on the Supporter and Sympathizer datasets. To ensure consistency across all categories and prevent overfitting, we have employed the default or automatic hyperparameters for all the aforementioned classifiers. In addition, majority and random label predictors are added as baselines for each of the datasets.

5.5.3 Analysis of Results

In this section we analyze the results obtained for different regions (SCT, WAL and NIR) on Member, Supporter and Sympathizer datasets, testing the methods through regions and different levels of political attachment. On the one hand, we are interested in evaluating standalone text-based and interaction-based methods. On the other hand, we compare standalone methods with hybrid approaches that combine interactions and text.

Text vs interaction representations: In Table 5.15 we can compare the results for text and interaction-based representations on their own, showing which data type is better to represent political leaning. Regarding text-based representations, Bert-based (B and dB) approaches generally yield superior results compared to Roberta-based (R and Rt), w2v and tfidf methods. Interestingly, despite being a smaller model, dB achieves superior results compared to B, being the best text based approach. Among Roberta-based approaches, Rt is significantly superior to the R method, given that the former is an extended version of R that has been retrained using Tweets. Bert-based (B and dB) approaches also outper-

		SCT			WAL			NIR			
		Mem.	Sup.	Sym.	Mem.	Sup.	Sym.	Mem.	Sup.	Sym.	ALL
Baselines	majority	13.4	06.9	06.4	12.4	09.9	10.1	08.2	09.3	04.8	09.0
	random	17.8	21.6	18.7	21.5	27.9	28.1	21.2	21.9	16.1	21.6
Interactions	RE	99.4	91.5	62.7	97.9	95.8	59.6	97.6	72.9	33.0	78.9
	N2V	80.8	62.0	17.4	67.5	48.5	11.1	60.8	21.6	07.0	41.9
	DW	80.9	61.6	22.9	77.5	57.0	14.7	72.2	23.6	06.6	46.3
Text	tfidf	57.2	52.4	29.5	64.8	59.5	26.1	60.2	45.8	25.9	46.8
	w2v	64.0	40.5	27.5	64.0	51.2	31.5	57.9	37.9	26.7	44.6
	B	67.6	47.8	27.7	75.6	55.2	33.9	68.4	48.5	34.4	51.0
	dB	73.4	59.0	36.9	75.1	64.1	33.6	72.1	49.8	28.5	54.7
	R	47.6	43.7	29.8	58.0	42.0	28.0	50.7	42.0	24.3	40.7
	Rt	60.7	48.3	34.5	64.4	52.2	32.7	57.4	44.0	27.8	46.9
Hybrid	RE + tfidf	**99.7***	**97.7***	**74.2***	**99.2***	**98.4***	**67.0***	97.3	**82.8***	**48.1***	**84.9***
	RE + w2v	99.4	95.6*	66.0*	98.5*	96.0*	64.4*	98.2*	72.7	37.2*	80.9*
	RE + B	98.5	94.0*	63.1*	**99.2***	94.9	60.8*	97.7*	78.5*	45.8*	81.4*
	RE + dB	99.4	95.4*	64.3*	**99.2***	94.7	61.6*	**98.4***	80.2*	42.9*	81.8*
	RE + R	99.4	94.6*	66.0*	98.6*	96.5*	64.7*	97.4	78.5*	41.1*	81.9*
	RE + Rt	99.4	94.4*	66.3*	**99.2***	96.0*	62.3*	98.1*	78.4*	44.7*	82.1*
	N2V + tfidf	74.5	44.5	11.3	42.2	26.7	11.8	32.5	10.3	06.7	29.5
	N2V + w2v	89.8*	56.0	28.6*	75.1*	60.2*	32.7*	73.1*	45.2*	25.4	54.0*
	N2V + B	77.6	52.1	27.7	74.5	55.9	35.6*	68.0	48.3	34.6*	52.7*
	N2V + dB	83.2*	60.5	35.0	74.8	65.1*	34.4*	71.6	50.1*	29.2*	56.0*
	N2V + R	75.4	47.1	28.2	58.4	45.6	26.7	52.5	37.3	22.6	43.8*
	N2V + Rt	79.4	51.0	32.8	65.5	57.6*	30.6	57.9	42.8	28.1	49.5*
	DW + tfidf	71.9	38.1	13.1	57.1	28.8	10.5	45.7	13.2	06.7	32.1
	DW + w2v	87.5*	58.0	30.1*	79.4*	64.2*	31.0	78.5*	46.2*	29.2*	56.0*
	DW + B	76.5	53.3	27.5	75.5	56.4	35.6*	69.2	49.2*	34.3	53.1*
	DW + dB	83.8*	62.0*	35.6	75.9	65.0*	33.9*	72.1	50.2*	29.4*	56.4*
	DW + R	74.6	43.4	25.1	63.9	50.5	26.8	62.0	37.9	22.0	45.1*
	DW + Rt	78.6	48.9	33.5	67.8	57.3	29.8	61.8	42.4	27.7	49.8*

5.15 Table – Evaluation results. Macro-F1 scores on SCT, WAL and NIR from Member (10 fold CV), Supporter and Sympathizer datasets. Algorithms used to generate the representations: Relational Embeddings (RE), Node2vec (N2V), DeepWalk (DW), tfidf, word2vec (w2v), mBERT (B), DistilmBERT (dB), XLM-RoBERTa (R), XLM-T (Rt). Values in **bold** represent best overall results for each dataset, while underlined values represent best results on text-only and interactions-only framework. Values with * represent when the combination of text and interactions gets better results than each on its own.

form interaction-based N2V and DW on politically less engaged users. However, interaction-based approaches tend to be better with politically engaged Member users.

In conclusion, max-pooled DistilmBERT Transformer model is the best approach to tackle political leaning inference with textual data, outperforming other

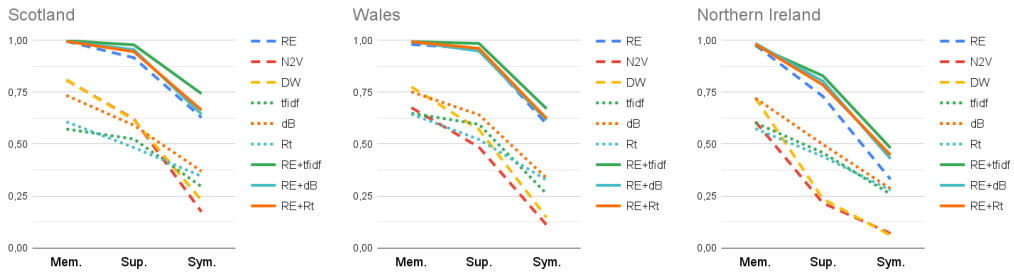
Transformer configurations as well as w2v and tfidf baselines. Furthermore, we have to remark that none of the textual approaches come close to surpassing the performance of interaction-based RE representations, which consistently yield superior results across all regions and political attachments. However, RE is not able to perform well with politically less engaged users, as the performance drops as the political involvement decreases.

Hybrid representations: Taking a wider look into the whole Table 5.15, the results indicate that the use of hybrid text and interaction representations leads to a notable improvement in results as compared to their independent usage (ALL column on Table 5.15). The RE method consistently yields better results when compared to other approaches based on interactions or text on their own. However, incorporating any of the extracted text representations alongside the RE representations often leads to improved results. This is particularly beneficial for politically less engaged users since their interactions alone may not provide enough information to determine their orientation accurately.

Thus, RE combined with Transformer-based representations (B, dB, R, Rt) generally perform better than when the RE are on their own. Especially, RE combined with Rt (RE+Rt) outperformed RE method for the whole 3 regions and all the political engagements. Hence, the combination of RE with Transformer-based representations (B, dB, R, Rt) typically yields superior performance compared to standalone RE. Furthermore, tfidf representations, which do not perform well when combined with *DW* and *N2V* representations, considerably enhance the results achieved by *RE* when combined with them. Thus, results on the Sympathizer datasets improve more than 10 points in average across the three regions, while for Supporter they increased more than 5 points. Considering that the added *tfidf* representations are based on the most significant terms per user, we can hypothesize that certain referential terms may serve as anchor terms for specific political parties.

Results for different levels of engagement: We next look at the performance scores with a focus on the three levels of engagement, i.e. members, supporters and sympathizers. Results for these three groups as shown in Figure 5.13 indicate that, as hypothesized, politically less engaged users are more difficult to predict for all the selected approaches. This observation is supported by the downward trend of all the lines, visually showing a steep performance decrease as engagement fades. This trend in turn reinforces the need for a data collection strategy

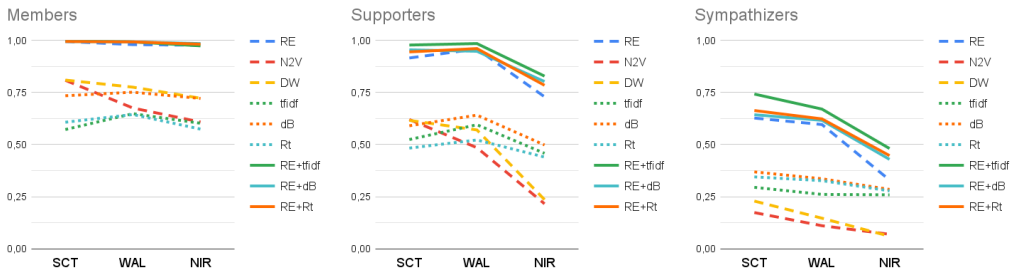
like the one we defined here to collect not only actively engaged users, but also those with more modest levels of engagement which do need to be considered in these analyses. Furthermore, we also observe that political leaning inference of highly engaged users is best achieved by exploiting their content sharing actions or interactions (N2V and DW). This however changes when we shift our focus towards less engaged users, where the use of textual content becomes more crucial as the interactions alone do not suffice (i.e. tfidf, dB and Rt). Interestingly, the combination of both data types further improves the results, especially when the hard-to-beat RE method underperforms on politically less engaged users.



5.13 Figure – Performance variations for interaction-based approaches (RE, N2V and DW), best text-based approaches (tfidf, dB and Rt) and corresponding hybrid approaches (RE+tfidf, RE+dB and RE+Rt) across different levels of political engagement on SCT (left), WAL (center) and NIR (right) datasets.

Results across regions: We are also interested in looking at how results differ across the three regions under study, i.e. Scotland (SCT), Wales (WAL) and Northern Ireland (NIR). We show results for each region in Figure 5.14, which shows that performance scores across regions also vary depending on the level of political engagement. Performance is consistenly high and comparable across all three regions when we look at highly engaged users (i.e. members). These trends vary more across regions when we look at less engaged users. As performance decreases for less engaged users across all three regions, we see that this drop is more prominent to some extent for WAL but particularly for NIR. Performance is more stable across regions when text-based methods (tfidf, dB and Rt) are used than when interaction-based methods (N2V and DW) are used, given that the availability of interaction data varies across regions. The weakest overall results occur within the NIR region, not least when interaction-based approaches are used on less engaged users. These weak results are however mitigated through the use of

hybrid representations, especially when used in combination with RE+tfidf, which again proves to be a more robust strategy to be used, both to ensure consistency across regions as well as to better generalize on less engaged users.



5.14 Figure – Performance variations for interaction based approaches (RE, N2V and DW), best text based approaches (tfidf, dB and Rt) and corresponding hybrid approaches (RE+tfidf, RE+dB and RE+Rt) across regions for members (left), supporters (center) and sympathizers (right) datasets.

5.6 Conclusion

In this work we look at the ability to infer the political leaning of social media users across multiple regions with multi-party systems, a challenging scenario that, to the best of our knowledge, has not been studied before. To conduct political leaning inference dynamically we proposed a two-step approach, starting with different unsupervised user representations through retweets, followed by political party classification. With the of comparing binary and multi-party frameworks, we collect data from three politically complex areas in Spain, characterized by plurinationality and the emergence of new political actors.

Furthermore, we present the first systematic analysis of political leaning inference, considering different levels of involvement. We collected a dataset spanning three UK regions, where users with different levels of political engagement (members, supporters, sympathizers) are labelled by the political party they align with. By conducting a set of experiments with different datasets, we find that a model leveraging user interactions based on Relational Embeddings, achieves the best results. Unlike the other methods, RE use real user interactions without generating any artificial user connections (see Chapter 4). Experimental results are consistent across regions and with scarce training data, demonstrating its stability and robustness in different scenarios and situations.

However, these representations tend to underperform when dealing with users who have weaker political engagements. To overcome this limitation, the proposed hybrid approaches to combine RE interaction representations with all the proposed textual representations results in considerable improvements across all datasets, especially with politically less attached users. The results are consistent across the regions and the different levels of political engagement, demonstrating its robustness. All in all, we demonstrate that our proposed hybrid approach achieves consistently improved performances, with a slight improvement on highly engaged users, but a remarkable improvement with those less engaged.

Considering the improvement obtained from the combination of textual and interaction data, we need to conduct further research to extract hybrid representations. We can experiment with various models using different datasets with missing information to address more realistic scenarios. As the collected datasets include interaction and textual data, we are able to try different configurations. Furthermore, we want to implement the proposed data extraction and user representation techniques in various other tasks, including hate-speech, disinformation or propaganda detection.

6. CHAPTER

Conclusions

In this thesis we have developed data collection and user representation methods to perform more accurate and generalizable social research. We show that hybrid methods, based on text and user interactions, are beneficial for a number of tasks, including stance detection and political learning inference. More specifically, and in relation to the different research lines outlined in Section 1.1, the main **contributions** of this thesis are the following:

- We proposed a new methodology to investigate how Artificial Intelligence can be applied to social research. Thus, we explored data collection and labeling techniques, as well as classification models to infer social media user's demographic traits. In order to do so, we first collected *Heldugazte-osoa* corpus, consisting of 6M publications in Basque, enabling further analysis of this under-resourced language. Second, we annotated *Heldugazte* and *Heldugazte-age* datasets to infer the life stage of Basque users (young or adult). Leveraging these datasets, we developed and evaluated different methods for life stage classification, including experiments with monolingual and multilingual Transformer-based language models. Third, we applied our methods to the raw corpus to qualitatively analyze and evaluate their performance in real-world scenarios. Finally, we empirically demonstrated the potential of interactions to convey socio-political information. This contribution corresponds to the **L1** research line.
- We collected and annotated the **VaxxStance dataset**, a comprehensive public dataset designed for stance detection on the vaccines topic. This dataset

includes both text and interaction data in two different languages (Basque and Spanish), centered on the same topic. Therefore, we encourage experimentation using both social and textual features in multilingual and crosslingual settings. This contribution corresponds to the **L2.1** research line.

- We presented a novel method, **Relational Embedding**, to represent and exploit interaction data based on one-to-one relations, such as *friends* and/or *retweets*. We experimented on seven publicly available stance detection datasets, showing that our method behaved robustly across various targets and languages without any specific manual engineering. Furthermore, combining our method with textual data systematically improved the results, outperforming even ensembles of large pre-trained language models. This contribution is related to the **L2.2** research line.

- We proposed a **multi-party framework** to better capture political leaning based on institutional political parties, which proves adaptable to different regions since it is grounded on localized political actors. We annotated a dataset containing labeled users by political party and left-right orientation alongside their retweets from the regions of Basque Country, Catalonia and Galicia. Subsequently, comprehensive experimentation with multi-party (7 political parties) and binary (left-right) frameworks showed that Relational Embeddings outperform other user representation methods even with scarce training data. This contribution refers to the **L3.1** research line.

- We delved into political learning inference based on the previously described multi-party framework by focusing on **different levels of user's political engagement**. We annotated another dataset containing labeled users by political party alongside their retweets from the regions of Wales, Scotland and Northern Ireland. Therefore, we include three datasets per region regarding different levels of implication with a political party: members, supporters or sympathizers. We evaluate a range of methodologies to make the most of retweet interactions among social media users to infer their political leaning, showing again that Relational Embedding based approach is effective even along the different levels of engagement. This contribution corresponds to the **L3.2** research line.

- Finally, we addressed political learning inference without relying on any specific data type although our results demonstrate that interaction-based Relational Embeddings outperformed state-of-the-art textual approaches. In

addition, we proposed **hybrid modeling** of social media users merging text and interaction features, showing that the combination of both data types is required for optimal performance. This contribution is related to the **L3.3** research line.

In terms of **publications**, this thesis contains 3 papers published in JCR indexed journals, namely, *Journal of Multilingual and Multicultural Development*, *Information* and *Procesamiento Del Lenguaje Natural*. In addition, we published 2 congress papers at IberLef and SIGUL workshops in SEPLN and COLING. Finally, we published 7 other peer reviewed papers during this PhD (2 Basque scientific journals, 2 SEPLN, 3 IkerGazte), including a most relevant research for the development of the Basque Country **award at IkerGazte** 2021. Four other papers are currently under review for different high impact journals.

During the development of the thesis, we also contributed to the generation of **valuable resources** for the Basque language. On the one hand, we have released *heldugazte-osoa*[1] and *Basque Twitter Corpus*[2] (Fernandez de Landa *et al.* 2024a), two datasets containing textual publications (6M and 8M respectively) on Basque language from social media users. On the other hand we developed two annotated datasets in order to infer the writing style (*heldugazte*[3]) and the life stage (*heldugazte-age*[4]) of Basque speaking social media users. Furthermore, the *VaxxStance*[5] stance detection dataset for Basque and Spanish was used as a natural language understanding (NLU) benchmark in Basque (Urbizu *et al.* 2022; Artetxe *et al.* 2022). The aforementioned resources will be hosted in the Clariaheus strategic network with the aim of contributing to the social sciences and humanities for the Basque language and culture (Alkorta *et al.* 2024). Additionally, we generated 3 other datasets for political leaning inference in different settings and regions. In total, we have created 2 raw corpora and 6 annotated datasets for different tasks and languages.

Our research field, in which we made several contributions, has undergone substantial changes since this thesis began. On one hand, the appearance of Large Language Models (LLMs) (Touvron *et al.* 2023; Achiam *et al.* 2023; Team *et al.* 2023; Wu *et al.* 2023) has significantly benefited individuals in addressing social science-related tasks, serving as valuable resources for automatic labeling (Chang

[1] http://ixa2.si.ehu.es/heldugazte-corpus/heldugazte.osoa.tar.gz
[2] https://github.com/joseba-fdl/basque_twitter_covid19_corpus
[3] https://github.com/ixa-ehu/heldugazte-corpus
[4] https://github.com/joseba-fdl/heldugazte-age-corpus
[5] https://vaxxstance.github.io/

et al. 2023) among others. However, supervised approaches achieve better evaluation results than generative LLMs for stance (Etxaniz *et al.* 2024) or ideology (Ziems *et al.* 2023) classification tasks. On the other hand, social media data from platforms like Twitter and Reddit is no longer easily accessible, prompting researchers to explore other data sources that leverage not only textual data but also data related to user interactions.

As future work, and in order to develop a richer social analysis, we might apply the data extraction and user representation techniques for other tasks, including hate-speech, disinformation or propaganda detection. Furthermore, we are eager to apply the methodologies developed in this thesis to identify possible social or political biases in discriminative or generative LLMs. By doing so, we aim to improve the fairness and accountability of these models thereby ensuring their contributions without perpetuating harmful biases.

Bibliography

Abul-Fottouh D. and Fetner T. Solidarity or schism: ideological congruence and the twitter networks of egyptian activists. *Mobilization: An International Quarterly*, 23(1):23–44, 2018.

Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Aleman F.L., Almeida D., Altenschmidt J., Altman S., Anadkat S., *et al.*. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agerri R., Barnes J., Bengoetxea J., Calvo B., Fernandez de Landa J., García-Ferrero I., Toporkov O., and Zubiaga I. Hitz@ disargue: Few-shot learning and argumentation to detect and fight misinformation in social media. *Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations (SEPLN-CEDI-PD 2024)*. CEUR-WS.org, 2024.

Agerri R., Bermudez J., and Rigau G. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2014 lib., 3823–3828, 2014.

Agerri R., Centeno R., Espinosa M., de Landa J.F., and Álvaro Rodrigo. Vaxxstance@iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181, 2021.

Agerri R. and Rigau G. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238(2):63–82, 2016.

Agerri R. and Rigau G. Language independent sequence labelling for Opinion Target Extraction. *Artificial Intelligence*, 268:85–95, 2019.

Agerri R., San Vicente I., Campos J.A., Barrena A., Saralegi X., Soroa A., and Agirre E. Give your Text Representation Models some Love: the Case for

Basque. *Proceedings of The 12th Language Resources and Evaluation Conference*, 4781–4788, 2020.

Akbik A., Blythe D., and Vollgraf R. Contextual string embeddings for sequence labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, 1638–1649, 2018.

Akhtar S., Basile V., and Patti V. A new measure of polarization in the annotation of hate speech. *AI* IA 2019–Advances in Artificial Intelligence: XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings 18*, 588–603. Springer, 2019.

Akoglu L. Quantifying political polarity based on bipartite opinion networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 2–11, 2014.

Al Zamal F., Liu W., and Ruths D. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. *Proceedings of the International AAAI Conference on Web and Social Media*, 270:2012, 2012.

AlDayel A. and Magdy W. Stance Detection on Social Media: State of the Art and Trends. *Information Processing & Management*, 58(4):102597, 2021.

Alegria I., Aranberri N., Comas P.R., Fresno V., Gamallo P., Padró L., San Vicente I., Turmo J., and Zubiaga A. TweetNorm: a benchmark for lexical normalization of Spanish tweets. *Language Resources and Evaluation*, 49(4):883–905, 2015.

Alkhalifa R. and Zubiaga A. QMUL-SDS@SardiStance: Leveraging Network Inter-actions to Boost Performance on Stance Detection using Knowledge Graphs. *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR Workshop Proceedings, 2020.

Alkorta J., Farwell A., Fernandez de Landa J., Altuna B., Estarrona A., Iruskieta M., Arregi X., Goenaga X., and Arriola J.M. Clariah-eus: a cross-border clariah node for the basque language and culture. *Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations (SEPLN-CEDI-PD 2024)*. CEUR-WS.org, 2024.

Almond G.A. and Verba S. *The civic culture: Political attitudes and democracy in five nations*. Princeton university press, 2015.

Alpaydin E. *Machine learning*. 2021.

Armengol-Estapé J., Carrino C.P., Rodriguez-Penagos C., de Gibert Bonet O., Armentano-Oller C., Gonzalez-Agirre A., Melero M., and Villegas M. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4933–4946. Association for Computational Linguistics, 2021.

Artetxe M., Aldabe I., Agerri R., Perez-de Viñaspre O., and Soroa A. Does corpus quality really matter for low-resource languages? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7383–7390, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

Atanasov A., De Francisci Morales G., and Nakov P. Predicting the role of political trolls in social media. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 1023–1034, 2019.

Augenstein I. Towards explainable fact checking. *ArXiv*, abs/2108.10274, 2021.

Augenstein I., Rocktäschel T., Vlachos A., and Bontcheva K. Stance detection with bidirectional conditional encoding. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 876–885, Austin, Texas, November 2016. Association for Computational Linguistics.

Ayyoubzadeh S.M., Ayyoubzadeh S.M., Zahedi H., Ahmadi M., and Kalhori S.R.N. Predicting covid-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study. *JMIR public health and surveillance*, 6(2):e18828, 2020.

Bail C.A. *Terrified: How anti-Muslim fringe organizations became mainstream*. Princeton University Press, 2015.

Baldwin T., de Marneffe M.C., Han B., Kim Y.B., Ritter A., and Xu W. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *Proceedings of the Workshop on Noisy User-generated Text*, 126–135, 2015.

Barberá P. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23:76 – 91, 2015.

Barberá P., Jost J.T., Nagler J., Tucker J.A., and Bonneau R. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

Barberá P. and Rivero G. Understanding the political representativeness of twitter users. *Social Science Computer Review*, 33:712 – 729, 2015.

Barbieri F., Anke L.E., and Camacho-Collados J. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 258–266, 2022.

Basile V., Bosco C., Fersini E., Nozza D., Patti V., Rangel Pardo F.M., Rosso P., and Sanguinetti M. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, 54–63, 2019.

Bauman Z. *Liquid modernity*. John Wiley & Sons, 2000.

Beck U. *Risk society: Towards a new modernity*. sage, 1992.

Belli R.F., Traugott M.W., Young M., and McGonagle K.A. Reducing vote overreporting in surveys: Social desirability, memory failure, and source monitoring. *The Public Opinion Quarterly*, 63(1):90–108, 1999.

Blondel V.D., Guillaume J.L., Lambiotte R., and Lefebvre E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Bojanowski P., Grave E., Joulin A., and Mikolov T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

Bonikowski B. and Gidron N. The populist style in american politics: Presidential campaign discourse, 1952–1996. *Social Forces*, 94(4):1593–1621, 2016.

Boutet A., Kim H., and Yoneki E. What's in your tweets? i know who you supported in the uk 2010 general election. *Proceedings of ICWSM*, 2012.

Boutyline A. and Willer R. The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political psychology*, 38(3): 551–569, 2017.

Boyd D. and Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

Brown P.F., Desouza P.V., Mercer R.L., Pietra V.J.D., and Lai J.C. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

Bruns A., Harrington S., and Hurcombe E. <? covid19?>'corona? 5g? or both?': the dynamics of covid-19/5g conspiracy theories on facebook. *Media International Australia*, 177(1):12–29, 2020.

Bruns A. and Highfield T. Is habermas on twitter?: Social media and the public sphere. *The Routledge companion to social media and politics*, 56–73. Routledge, 2015.

Castells M. The network society revisited. *American Behavioral Scientist*, 67(7): 940–946, 2023.

Caton S. and Haas C. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020.

Celli F., Stepanov E., Poesio M., and Riccardi G. Predicting brexit: Classifying agreement is better than sentiment and pollsters. *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, 110–118, 2016.

Cesare N., Grant C., and Nsoesie E.O. Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807*, 2017.

Cesare N., Lee H., McCormick T., Spiro E., and Zagheni E. Promises and pitfalls of using digital traces for demographic research. *Demography*, 55(5):1979–1999, 2018.

Chancellor S. and De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43, 2020.

Chang Y., Wang X., Wang J., Wu Y., Yang L., Zhu K., Chen H., Yi X., Wang C., Wang Y., *et al.*. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Chen S.F. and Goodman J. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

Cignarella A.T., Lai M., Bosco C., Patti V., and Rosso P. SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In Basile V., Croce D., Di Maro M., and Passaro L.C., editors, *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org, 2020.

CIS. *Barómetro de diciembre 2019. Postelectoral Elecciones Generales 2019*, 3269 lib. Centro de Investigaciones Sociológicas, 11 2019.

CIS. *Preelectoral de Galicia. Elecciones Autonómicas julio 2020*, 3287 lib. Centro de Investigaciones Sociológicas, 06 2020a.

CIS. *Preelectoral del País Vasco. Elecciones Autonómicas julio 2020*, 3286 lib. Centro de Investigaciones Sociológicas, 06 2020b.

Clark A. Combining distributional and morphological information for part of speech induction. *10th Conference of the European Chapter of the Association for Computational Linguistics*, 2003.

Colleoni E., Rozza A., and Arvidsson A. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332, 2014.

Collins M. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1–8, 2002.

Conforti C., Berndt J., Pilehvar M.T., Giannitsarou C., Toxvaerd F., and Collier N. Will-they-won't-they: A very large dataset for stance detection on twitter. *ACL*, 2020.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., and Stoyanov V. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Conover M.D., Gonçalves B., Ratkiewicz J., Flammini A., and Menczer F. Predicting the political alignment of twitter users. *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, 192–199. IEEE, 2011a.

Conover M.D., Ratkiewicz J., Francisco M., Gonçalves B., Menczer F., and Flammini A. Political Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 2011b.

Corley C.D., Cook D.J., Mikler A.R., and Singh K.P. Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public health*, 7(2):596–615, 2010.

Culotta A. Towards detecting influenza epidemics by analyzing twitter messages. *Proceedings of the first workshop on social media analytics*, 115–122, 2010.

Darwish K., Stefanov P., Aupetit M., and Nakov P. Unsupervised user stance detection on twitter. *Proceedings of ICWSM*, 14 lib., 141–152, 2020.

De Choudhury M., Gamon M., Counts S., and Horvitz E. Predicting depression via social media. *Proceedings of the international AAAI conference on web and social media*, 128–137, 2013.

Defays D. An efficient algorithm for a complete link method. *The computer journal*, 20(4):364–366, 1977.

Del Tredici M., Marcheggiani D., Schulte im Walde S., and Fernández R. You shall know a user by the company it keeps: Dynamic representations for social media users in NLP. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4707–4717. Association for Computational Linguistics, 2019.

DellaPosta D., Shi Y., and Macy M. Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511, 2015.

Derczynski L., Bontcheva K., Liakata M., Procter R., Hoi G.W.S., and Zubiaga A. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 69–76, 2017.

Devlin J., Chang M., Lee K., and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186, 2019.

Eagle N., Macy M., and Claxton R. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

Eckert P. Age as a sociolinguistic variable. *The handbook of sociolinguistics*, 151–167, 2017.

Edelmann A., Wolff T., Montagne D., and Bail C.A. Computational social science and sociology. *Annual Review of Sociology*, 46:61–81, 2020.

Espinosa M.S., Agerri R., Rodrigo A., and Centeno R. DeepReading@SardiStance: Combining Textual, Social and Emotional Features. *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR Workshop Proceedings, 2020.

Etxaniz J., Sainz O., Perez N., Aldabe I., Rigau G., Agirre E., Ormazabal A., Artetxe M., and Soroa A. Latxa: An open language model and evaluation suite for basque, 2024.

Eurostat. *Being young in Europe today - digital world*. Eurostat. European Statistical Office, 2023a.

Eurostat. *Digitalisation in Europe - 2023 edition*. Eurostat. European Statistical Office, 2023b.

Eusko Jaurlaritza E., Gobernua N., and de la Langue Basque O.P. VI. Inkesta Soziolinguistikoa. *irekia.euskadi.eus*, 2016.

Eustat. *Panorama de da sociedad de la información - 2022*, 220126 lib. Eustat - Euskal Estatistika Erakundea / Instituto Vasco de Estadística, 2022.

Fagni T. and Cresci S. Fine-grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning. *Journal of Artificial Intelligence Research*, 73:633–672, 2022.

Fatehkia M., Kashyap R., and Weber I. Using facebook ad data to track the global digital gender gap. *World Development*, 107:189–209, 2018.

Fernandez de Landa J. Sare sozialen erabilera moduak eta maiztasunak gasteizko nerabeen kolektiboaren baitan. *II. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Gizarte Zientziak eta Zuzenbidea*, 2017.

Fernandez de Landa J. Gazteak eta euskara sare sozialetan. zer, nori, nork: euskarazko txio formal eta informalak sailkatuz eta konparatuz. *Eusko Ikaskuntzaren XVIII. Kongresua Geroa Elkar-Ekin: Mendeurreneko Kongresua*, (18):348–355, 2019.

Fernandez de Landa J. and Agerri R. Euskarazko on-line artikuluetan aipatutako izendun entitate nabarmenen identifikazioa denbora errealean. *EKAIA EHUko Zientzia eta Teknologia aldizkaria*, (40), 2021a.

Fernandez de Landa J. and Agerri R. Social analysis of young basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, 0(0):1–15, 2021b.

Fernandez de Landa J. and Agerri R. Relational embeddings for language independent stance detection. *arXiv preprint arXiv:2210.05715*, 2022.

Fernandez de Landa J. and Agerri R. Hitz-ixa at politices-iberlef2023: Document and sentence level text representations for demographic characteristics and political ideology detection. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023)*. CEUR-WS.org, 2023.

Fernandez de Landa J. and Agerri R. Political leaning inference through plurinational scenarios. *arXiv preprint arXiv:2406.07964*, 2024.

Fernandez de Landa J., Agerri R., and Alegria I. Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, 10(6):212, 2019a.

Fernandez de Landa J., Alegria I., and Agerri R. Euskaldun gazte eta helduen harremanak twitterren. *III. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Gizarte Zientziak eta Zuzenbidea*, 2019b.

Fernandez de Landa J., García-Ferrero I., Salaberria A., and Campos J.A. Uncovering social changes of the basque speaking twitter community during covid-19 pandemic. *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, 363–371, Torino, Italia, 2024a. ELRA and ICCL.

Fernandez de Landa J., García Ferrero I., Salaberria Saizar A., and Campos Tejedor J.A. Twitterreko euskal komunitatearen eduki azterketa pandemia garaian. *IV. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Ingeniaritza eta Arkitektura*, 2021.

Fernandez de Landa J., Zubiaga A., and Agerri R. Generalizing political leaning inference to multi-party systems: Insights from the uk political landscape. *arXiv preprint arXiv:2312.01738*, 2023.

Fernandez de Landa J., Zubiaga A., and Agerri R. Htim: Hybrid text-interaction modeling for broadening political leaning inference in social media. *arXiv preprint arXiv:2406.08201*, 2024b.

Ferraccioli F., Sciandra A., Pont M.D., Girardi P., Solari D., and Finos L. TextWiller@SardiStance, HaSpeede2: Text or Context? A smart use of social network data in predicting polarization. *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR Workshop Proceedings, 2020.

Ferrara E., Chang H., Chen E., Muric G., and Patel J. Characterizing social media manipulation in the 2020 us presidential election. *First Monday*, 2020.

Flores R.D. Do anti-immigrant laws shape public sentiment? a study of arizona's sb 1070 using twitter data. *American Journal of Sociology*, 123(2):333–384, 2017.

Ford R. and Jennings W. The changing cleavage politics of western europe. *Annual review of political science*, 23:295–314, 2020.

Forgy E.W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

Fruchterman T.M.J. and Reingold E.M. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21, 1991.

Gamallo P., Pichel J.R., and Alegria I. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152–162, 2017.

Garcia-Díaz J.A., Jiménez Zafra S.M., Martín Valdivia M.T., García-Sánchez F., Ureña López L.A., and Valencia García R. Overview of politices at iberlef 2023: Political ideology detection in spanish texts. 2023.

García-Díaz J.A., Jiménez-Zafra S.M., Valdivia M.T.M., García-Sánchez F., Ureña-López L.A., and Valencia-García R. Overview of politices 2022: Spanish author profiling for political ideology. *Procesamiento del Lenguaje Natural*, 69(0):265–272, 2022. ISSN 1989-7553.

Garimella K., Morales G.D.F., Gionis A., and Mathioudakis M. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018.

Garimella V.R.K. and Weber I. A long-term analysis of polarization on twitter. *Proceedings of ICWSM*, 11 lib., 2017.

Ghosh S., Singhania P., Singh S., Rudra K., and Ghosh S. Stance detection in web and social media: a comparative study. *International Conference of the Cross-Language Evaluation Forum for European Languages*, 75–87. Springer, 2019.

Giddens A. *The consequences of modernity*. 1990.

Giorgioni S., Politi M., Salman S., Basili R., and Croce D. Unitor @ sardistance2020: Combining transformer-based architectures and transfer learning for robust stance detection. *EVALITA*, 2020.

Glandt K., Khanal S., Li Y., Caragea D., and Caragea C. Stance detection in covid-19 tweets. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 1 lib., 2021.

Golder S.A. and Macy M.W. Digital footprints: Opportunities and challenges for online social research. *Annual review of sociology*, 40:129–152, 2014.

González-Bailón S., Borge-Holthoefer J., and Moreno Y. Broadcasters and hidden influentials in online protest diffusion. *American behavioral scientist*, 57(7): 943–965, 2013.

González-Bailón S. and Wang N. Networked discontent: The anatomy of protest campaigns in social media. *Social networks*, 44:95–104, 2016.

González Bermúdez M. An analysis of Twitter corpora and the differences between formal and colloquial tweets. *Proceedings of the Tweet Translation Workshop 2015*, 1–7. CEUR-WS. org, 2015.

Goodfellow I., Bengio Y., and Courville A. *Deep learning*. MIT press, 2016.

Grave E., Bojanowski P., Gupta P., Joulin A., and Mikolov T. Learning word vectors for 157 languages. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Graves A., Mohamed A.r., and Hinton G. Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649. Ieee, 2013.

Grover A. and Leskovec J. node2vec: Scalable feature learning for networks. *Proceedings of KDD*, 2016.

Hallac I.R., Makinist S., Ay B., and Aydin G. user2vec: Social media user representation based on distributed document embeddings. *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–5, 2019.

Hanna A. Computer-aided content analysis of digitally enabled movements. *Mobilization: An International Quarterly*, 18(4):367–388, 2013.

Hardalov M., Arora A., Nakov P., and Augenstein I. Cross-domain label-adaptive stance detection. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9011–9028, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Hardalov M., Arora A., Nakov P., and Augenstein I. Few-Shot Cross-Lingual Stance Detection with Sentiment-Based Pre-Training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 lib., 10729–10737, 2022.

Hofman J.M., Watts D.J., Athey S., Garip F., Griffiths T.L., Kleinberg J., Margetts H., Mullainathan S., Salganik M.J., Vazire S., *et al.*. Integrating explanation and prediction in computational social science. *Nature*, 595(7866):181–188, 2021.

Hristakieva K., Cresci S., Da San Martino G., Conti M., and Nakov P. The spread of propaganda by coordinated communities on social media. *Proceedings of the 14th ACM Web Science Conference 2022*, 191–201, 2022.

Hua Y., Ristenpart T., and Naaman M. Towards measuring adversarial twitter interactions against candidates in the us midterm elections. *Proceedings of ICWSM*, 2020.

Huang J., Kornfield R., Szczypka G., and Emery S.L. A cross-sectional examination of marketing of electronic cigarettes on twitter. *Tobacco control*, 23(suppl 3):iii26–iii30, 2014.

Huang Z., Xu W., and Yu K. Bidirectional lstm-crf models for sequence tagging. *Proceedings of the 21st International Conference on Asian Language Processing.*, 2015.

Imhoff R., Zimmer F., Klein O., António J.H., Babinska M., Bangerter A., Bilewicz M., Blanuša N., Bovan K., Bužarovska R., *et al.*. Conspiracy mentality and political orientation across 26 countries. *Nature human behaviour*, 6 (3):392–403, 2022.

Jacomy M., Venturini T., Heymann S., and Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6):e98679, 2014.

Jones R.J., Cunliffe D., and Honeycutt Z.R. Twitter and the Welsh language. *Journal of Multilingual and Multicultural Development*, 34(7):653–671, 2013.

Jordan M.I. and Mitchell T.M. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Jusup M., Holme P., Kanazawa K., Takayasu M., Romić I., Wang Z., Geček S., Lipić T., Podobnik B., Wang L., Luo W., Klanjšček T., Fan J., Boccaletti S., and Perc M. Social physics. *Physics Reports*, 948:1–148, 2022. Social physics.

Karthikeyan K., Wang Z., Mayhew S., and Roth D. Cross-lingual ability of multilingual BERT: An empirical study. *International Conference on Learning Representations*, 2020.

Keating M. *Plurinational Democracy: Stateless Nations in a Post-Sovereignty Era*. Oxford University Press, 11 2001.

Kennedy C., Blumenthal M., Clement S., Clinton J.D., Durand C., Franklin C., McGeeney K., Miringoff L., Olson K., Rivers D., *et al.*. An evaluation of the 2016 election polls in the united states. *Public Opinion Quarterly*, 82(1):1–33, 2018.

Kenter T., Borisov A., and de Rijke M. Siamese CBOW: Optimizing word embeddings for sentence representations. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 941–951, Berlin, Germany, August 2016. Association for Computational Linguistics.

Kipf T. and Welling M. Semi-Supervised Classification with Graph Convolutional Networks. *International Conference on Learning Representations*, 2017.

Kitchin R. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

Kotroyannos D., Tzagkarakis S.I., and Pappas I. South european populism as a consequence of the multidimensional crisis? the cases of syriza, podemos and m5s. *European Quarterly of Political Attitudes and Mentalities*, 7(4):1–18, 2018.

Kotsiantis S.B., Zaharakis I., Pintelas P., *et al.*. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

Küçük D. and Can F. Stance Detection: a Survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37, 2020.

Kulshrestha J., Eslami M., Messias J., Zafar M.B., Ghosh S., Gummadi K.P., and Karahalios K. Quantifying search bias: Investigating sources of bias for political searches in social media. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017.

Lahoti P., Garimella V.R.K., and Gionis A. Joint non-negative matrix factorization for learning ideological leaning on twitter. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2017.

Lai M., Cignarella A.T., Farías D.I.H., Bosco C., Patti V., and Rosso P. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075, 2020a.

Lai M., Cignarella A.T., Finos L., and Sciandra A. Wordup! at vaxxstance 2021: Combining contextual information with textual and dependency-based syntactic features for stance detection. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, 2021.

Lai M., Patti V., Ruffo G., and Rosso P. #brexit: Leave or remain? the role of user's community and diachronic evolution on stance detection. *Journal of Intelligent & Fuzzy Systems*, 31(2):2341–2352, 2020b.

Lampos V. and Cristianini N. Tracking the flu pandemic by monitoring the social web. *2010 2nd international workshop on cognitive information processing*, 411–416. IEEE, 2010.

Larson H.J., Smith D.M., Paterson P., Cumming M., Eckersberger E., Freifeld C.C., Ghinai I., Jarrett C., Paushter L., Brownstein J.S., *et al.*. Measuring vaccine confidence: analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines. *The Lancet infectious diseases*, 13 (7):606–613, 2013.

Lazer D., Pentland A., Adamic L., Aral S., Barabási A.L., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M., Jebara T., King G., Macy M., Roy D., and Alstyne M.V. Computational social science. *Science*, 323(5915):721–723, 2009.

Lazer D.M.J., Pentland A., Watts D.J., Aral S., Athey S., Contractor N., Freelon D., Gonzalez-Bailon S., King G., Margetts H., Nelson A., Salganik M.J., Strohmaier M., Vespignani A., and Wagner C. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062, 2020.

Le Q. and Mikolov T. Distributed representations of sentences and documents. *International conference on machine learning*, 1188–1196. PMLR, 2014.

LeCun Y., Bengio Y., and Hinton G. Deep learning. *nature*, 521(7553):436–444, 2015.

Leturia I. Evaluating different methods for automatically collecting large general corpora for basque from the web. *Proceedings of COLING 2012*, 1553–1570, 2012.

Li Y., Zhao C., and Caragea C. Improving stance detection with multi-dataset learning and knowledge distillation. *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, 6332–6345, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Lisi M. *Party system change, the European crisis and the state of democracy*. Routledge, 2018.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., and Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lynch P. Party system change in britain: Multi-party politics in a multi-level polity. *British Politics*, 2(3):323–346, 2007.

Ma X., Wu J., Xue S., Yang J., Zhou C., Sheng Q.Z., Xiong H., and Akoglu L. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Magdy W., Darwish K., Abokhodair N., Rahimi A., and Baldwin T. #isisisnotislam or #deportallmuslims? predicting unspoken views. *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, page 95–106, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342087.

Makazhanov A. and Rafiei D. Predicting political preference of twitter users. *Social Network Analysis and Mining*, 4:1–15, 2013.

Marquardt J., Farnadi G., Vasudevan G., Moens M.F., Davalos S., Teredesai A., and De Cock M. Age and gender identification in social media. *Proceedings of CLEF 2014 Evaluation Labs*, 1129–1136, 2014.

McGann A., Dellepiane-Avellaneda S., and Bartle J. Parallel lines? policy mood in a plurinational democracy. *Electoral Studies*, 58:48–57, 2019.

McInnes L., Healy J., Saul N., and Großberger L. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.*, 3:861, 2018.

McMonagle S., Cunliffe D., Jongbloed-Faber L., and Jarvis P. What can hashtags tell us about minority languages on Twitter? A comparison of #cymraeg, #frysk, and #gaeilge. *Journal of Multilingual and Multicultural Development*, 40(1):32–49, 2019.

Mhichíl M.N.G., Lynn T., and Rosati P. Twitter and the Irish language, #Gaeilge – agents and activities: exploring a data set with micro-implementers in social media. *Journal of Multilingual and Multicultural Development*, 39(10):868–881, 2018.

Mikolov T., Chen K., Corrado G., and Dean J. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013a.

Mikolov T., Grave E., Bojanowski P., Puhrsch C., and Joulin A. Advances in Pre-Training Distributed Word Representations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, 2018.

Mikolov T., Sutskever I., Chen K., Corrado G.S., and Dean J. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119, 2013b.

Mishra P., Del Tredici M., Yannakoudakis H., and Shutova E. Author profiling for abuse detection. *Proceedings of the 27th International Conference on Computational Linguistics*, 1088–1098. Association for Computational Linguistics, 2018.

Mohammad S., Kiritchenko S., Sobhani P., Zhu X., and Cherry C. SemEval-2016 task 6: Detecting stance in tweets. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41, San Diego, California, June 2016. Association for Computational Linguistics.

Montes M., Rosso P., Gonzalo J., Aragón E., Agerri R., Ángel Álvarez Carmona M., Álvarez Mellado E., de Albornoz J.C., Chiruzzo L., Freitas L., Adorno H.G., Gutiérrez Y., Zafra S.M.J., Lima S., de Arco F.M.P., and (eds.) M.T. Proceedings of the iberian languages evaluation forum (iberlef 2021). CEUR Workshop Proceedings, 2021.

Morgan-Lopez A.A., Kim A.E., Chew R.F., and Ruddle P. Predicting age groups of Twitter users based on language and metadata features. *PloS one*, 12(8): e0183537, 2017.

Morlino L. and Raniolo F. *The impact of the economic crisis on South European democracies*. Springer, 2017.

Myslín M., Zhu S.H., Chapman W., Conway M., *et al.*. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15(8):e2534, 2013.

Nguyen D., Doğruöz A.S., Rosé C.P., and de Jong F. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593, 2016.

Nguyen D., Gravel R., Trieschnigg D., and Meder T. "How Old Do You Think I Am?" A Study of Language and Age in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 7 lib., 439–448, 2013.

Nguyen D., Trieschnigg D., Doğruöz A.S., Gravel R., Theune M., Meder T., and de Jong F. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. *25th International Conference on Computational Linguistics (COLING 2014)*, 1950–1961. Dublin City University and Association for Computational Linguistics, 2014.

Palmer J.R., Espenshade T.J., Bartumeus F., Chung C.Y., Ozgencil N.E., and Li K. New approaches to human mobility: Using mobile phones for demographic research. *Demography*, 50(3):1105–1128, 2013.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., and Duchesnay E. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pennacchiotti M. and Popescu A.M. Democrats, Republicans and Starbucks Aficionados: User Classification in Twitter. *Association for Computing Machinery*, 430–438. ACM, 2011a.

Pennacchiotti M. and Popescu A.M. Democrats, republicans and starbucks aficionados: user classification in twitter. *KDD*, 2011b.

Pennebaker J.W., Francis M.E., and Booth R.J. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

Pennington J., Socher R., and Manning C. Glove: Global vectors for word representation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics.

Perozzi B., Al-Rfou R., and Skiena S. Deepwalk: Online learning of social representations. KDD '14, page 701–710. Association for Computing Machinery, 2014.

Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., and Zettlemoyer L. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. Association for Computational Linguistics, 2018.

Plà F. and Hurtado L.F. Political tendency identification in twitter using sentiment analysis techniques. *International Conference on Computational Linguistics*, 2014.

Preotiuc-Pietro D., Liu Y., Hopkins D.J., and Ungar L.H. Beyond binary labels: Political ideology prediction of twitter users. *ACL*, 2017.

Qiu J., Tang J., Ma H., Dong Y., Wang K., and Tang J. Deepinf: Social influence prediction with deep learning. *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2110–2119, 2018.

Rajadesingan A. and Liu H. Identifying users with opposing opinions in twitter debates. *Social Computing, Behavioral-Cultural Modeling and Prediction: 7th International Conference, SBP 2014, Washington, DC, USA, April 1-4, 2014. Proceedings 7*, 153–160. Springer, 2014.

Rama J., Cordero G., and Zagórski P. Three is a crowd? podemos, ciudadanos, and vox: The end of bipartisanship in spain. *Frontiers in Political Science*, 3, 2021.

Rao D., Yarowsky D., Shreevats A., and Gupta M. Classifying latent user attributes in Twitter. *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 37–44. ACM, 2010.

Rashed A., Kutlu M., Darwish K., Elsayed T., and Bayrak C. Embeddings-based clustering for target specific stances: The case of a polarized turkey. *Proceedings of ICWSM*, 2021.

Recuero R., Zago G., and Soares F. Using social network analysis and social capital to identify user roles on polarized political conversations on twitter. *Social Media+ Society*, 5(2):2056305119848745, 2019.

Ritter A., Clark S., and Etzioni O. Named entity recognition in tweets: an experimental study. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1524–1534, 2011.

Rosenthal S., Farra N., and Nakov P. SemEval-2017 task 4: Sentiment analysis in Twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 502–518, 2017.

Rosenthal S. and McKeown K. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 763–772. Association for Computational Linguistics, 2011.

Russell S.J. and Norvig P. *Artificial intelligence: a modern approach*. Pearson, 2016.

Salaberria A., Campos J.A., Garcıa I., and Fernandez de Landa J. Twitterreko euskal komunitatearen eduki azterketa pandemia garaian. *IV. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Ingeniaritza eta Arkitektura*, 2021.

Salathé M., Freifeld C.C., Mekaru S.R., Tomasulo A.F., and Brownstein J.S. Influenza a (h7n9) and the importance of digital epidemiology. *The New England journal of medicine*, 369(5):401, 2013.

Salathé M. and Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology*, 7(10):e1002199, 2011.

Salganik M.J. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

Salton G. and Yu C.T. On the construction of effective vocabularies for information retrieval. *Acm Sigplan Notices*, 10(1):48–60, 1973.

Sanh V., Debut L., Chaumond J., and Wolf T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Proceedings of NeurIPS EMC2 Workshop*, 2019.

Santillana M., Nguyen A.T., Dredze M., Paul M.J., Nsoesie E.O., and Brownstein J.S. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10):e1004513, 2015.

Schiller B., Daxenberger J., and Gurevych I. Stance Detection Benchmark: How Robust Is Your Stance Detection? *KI-Künstliche Intelligenz*, 1–13, 2021.

Shen D., Wang G., Wang W., Min M.R., Su Q., Zhang Y., Li C., Henao R., and Carin L. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 440–450, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Shu K., Sliva A., Wang S., Tang J., and Liu H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19 (1):22–36, 2017.

Singh A., Thakur N., and Sharma A. A review of supervised machine learning algorithms. *2016 3rd international conference on computing for sustainable global development (INDIACom)*, 1310–1315. Ieee, 2016.

Sobhani P., Inkpen D., and Zhu X. A dataset for multi-target stance detection. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 551–557, 2017.

Stefanov P., Darwish K., Atanasov A., and Nakov P. Predicting the topical stance and political leaning of media using tweets. *Proceedings of ACL*, 2020.

Stewart I., Flores R.D., Riffe T., Weber I., and Zagheni E. Rock, rap, or reggaeton?: Assessing mexican immigrants' cultural assimilation using facebook data. *The world wide web conference*, 3258–3264, 2019.

Taulé M., Pardo F.M.R., Martí M.A., and Rosso P. Overview of the Task on Multimodal Stance Detection in Tweets on Catalan# 1oct Referendum. *IberEval@ SEPLN*, 149–166, 2018.

Team G., Anil R., Borgeaud S., Wu Y., Alayrac J.B., Yu J., Soricut R., Schalkwyk J., Dai A.M., Hauth A., *et al.*. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., Bashlykov N., Batra S., Bhargava P., Bhosale S., *et al.*. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Urbizu G., San Vicente I., Saralegi X., Agerri R., and Soroa A. BasqueGLUE: A natural language understanding benchmark for Basque. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1603–1612, Marseille, France, 2022. European Language Resources Association.

Van der Maaten L. and Hinton G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Van Deth J.W., Montero J.R., and Westholm A. *Citizenship and involvement in European democracies: A comparative analysis*, 17 lib. Routledge, 2007.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., and Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008, 2017.

Vayena E., Salathé M., Madoff L.C., and Brownstein J.S. Ethical challenges of big data in public health, 2015.

Vaz de Melo P.O.S. How many political parties should brazil have? a data-driven method to assess and reduce fragmentation in multi-party political systems. *PLOS ONE*, 10(10):1–24, 10 2015. URL `https://doi.org/10.1371/journal.pone.0140217`.

Velickovic P., Cucurull G., Casanova A., Romero A., Lio' P., and Bengio Y. Graph Attention Networks. *International Conference on Learning Representations*, 2018.

Villena J., Lana S., Martínez E., and González J.C. TASS-workshop on sentiment analysis at SEPLN. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 2013.

Wang W., Rothschild D., Goel S., and Gelman A. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

Webster F. *Theories of the information society*. Routledge, 2014.

Wicke P. and Bolognesi M.M. Framing covid-19: How we conceptualize and discuss the pandemic on twitter. *PloS one*, 15(9):e0240010, 2020.

Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.W., da Silva Santos L.B., Bourne P.E., *et al.*. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

Wong F.M.F., Tan C.W., Sen S., and Chiang M. Quantifying political leaning from tweets and retweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 2013.

Wu L. and Liu H. Tracing fake-news footprints: Characterizing social media messages by how they propagate. *Proceedings of the eleventh ACM international conference on Web Search and Data Mining*, 637–645, 2018.

Wu S., Koo M., Blum L., Black A., Kao L., Scalzo F., and Kurtz I. A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology. *arXiv preprint arXiv:2308.04709*, 2023.

Xiao Z., Song W., Xu H., Ren Z., and Sun Y. Timme: Twitter ideology-detection via multi-task multi-relational embedding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 2258–2268, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984.

Yan H., Das S., Lavoie A., Li S., and Sinclair B. The congressional classification challenge: Domain specificity and partisan intensity. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2019.

Yang D., Qu B., Yang J., and Cudre-Mauroux P. Revisiting user mobility and social relationships in lbsns: a hypergraph embedding approach. *The world wide web conference*, 2147–2157, 2019.

Zagheni E., Weber I., and Gummadi K. Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 721–734, 2017.

Zaghouani W. and Charfi A. Arap-Tweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

Zhang H. and Pan J. Casm: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, 49(1):1–57, 2019.

Zhelezniak V., Savkov A., Shen A., Moramarco F., Flann J., and Hammerla N.Y. Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors. *International Conference on Learning Representations*, 2019.

Ziems C., Held W., Shaikh O., Chen J., Zhang Z., and Yang D. Can large language models transform computational social science? *Computational Linguistics*, 50 (1):237–291, 2023.

Zotova E., Agerri R., Nuñez M., and Rigau G. Multilingual stance detection in tweets: The Catalonia independence corpus. *Proceedings of the 12th Language Resources and Evaluation Conference*, 1368–1375, Marseille, France, May 2020. European Language Resources Association.

Zotova E., Agerri R., and Rigau G. Semi-automatic generation of multilingual datasets for stance detection in Twitter. *Expert Systems with Applications*, 170: 114547, 2021.

Zubiaga A., San Vicente I., Gamallo P., Pichel J.R., Alegria I., Aranberri N., Ezeiza A., and Fresno V. TweetLID: a benchmark for tweet language identification. *Language Resources and Evaluation*, 50(4):729–766, 2016.

Zubiaga A., Wang B., Liakata M., and Procter R. Stance classification of social media users in independence movements. *Catalonia*, 2(8,599):10–960, 2017.

Zubiaga A., Wang B., Liakata M., and Procter R. Political homophily in independence movements: Analyzing and classifying social media users by national identity. *IEEE Intelligent Systems*, 34:34–42, 2019.

Glosategia

adimen artifizial *Artificial Intelligence*

adostasun maila *agreement rate*

aurkigarritasun *findability*

artearen egoera *state of the art*

aurre-entrenatu *pre-trained*

aurrerantza elikatzen den sare *feedforward network*

ausazko aldagai *random variable*

ausazko ibilaldi *random walk*

autonormalizazio *self-normalization*

azpilaginketa *subsampling*

azpihitz *subtoken*

datu-multzo *dataset*

doikuntza *fine-tuning*

doitasun *precision*

domeinuz kanpo *out-of-domain*

egiturarik gabeko datuak *unstructured data*

entropia gurutzatu *cross-entropy*

estaldura *recall*

euskarri bektoredun makina *support vector machine*

gainbeiratu *supervised*

gainbegiratze indartsu *strongly supervised*

gainbegiratze arin *weakly supervised*

gainlaginketa *upsampling*

geruza anitzeko perzeptroi *multilayer perceptron*

geruza ezkutu *hidden layer*

gizarte zientzia konputazionala *computational social science*

goiz-eten *early stopping*

hitz-bektore *word embedding*

hitzen bektore errepresentazio jarraituak *Continuous Vector Representations of words*

Hizkuntza Naturalaren Inferentzia *Natural Language Inference*

hizkuntza-arteko *crosslingual*

hizkuntza-eredu *language model*

indarrak-zuzendutako *force-directed*

interoperabilitate *interoperability*

ikasketa automatiko *Machine Learning*

ikasketa sakon *Deep Learning*

irisgarritasun *accessibility*

laginketa *sampling*

sailkapen geruza *classification layer*

Eranskinak

Eranskin honetan tesia osatzen duten artikuluak aurkitu daitezke, gomendatutako irakurketa ordenean aurkeztua.

A.1 Fernandez de Landa *et al.* (2019a)

information

MDPI

Article

Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case

Joseba Fernandez de Landa * and Rodrigo Agerri * and Iñaki Alegria

IXA NLP Group, University of the Basque Country UPV/EHU, 20018 Donostia-San Sebastian, Spain; i.alegria@ehu.eus

* Correspondence: joseba.fdl@gmail.com (J.F.d.L.); rodrigo.agerri@ehu.eus (R.A.)

Received: 30 April 2019; Accepted: 11 June 2019; Published: 13 June 2019

Abstract: Social networks like Twitter are increasingly important in the creation of new ways of communication. They have also become useful tools for social and linguistic research due to the massive amounts of public textual data available. This is particularly important for less resourced languages, as it allows to apply current natural language processing techniques to large amounts of unstructured data. In this work, we study the linguistic and social aspects of young and adult people's behaviour based on their tweets' contents and the social relations that arise from them. With this objective in mind, we have gathered over 10 million tweets from more than 8000 users. First, we classified each user in terms of its life stage (young/adult) according to the writing style of their tweets. Second, we applied topic modelling techniques to the personal tweets to find the most popular topics according to life stages. Third, we established the relations and communities that emerge based on the retweets. We conclude that using large amounts of unstructured data provided by Twitter facilitates social research using computational techniques such as natural language processing, giving the opportunity both to segment communities based on demographic characteristics and to discover how they interact or relate to them.

Keywords: social informatics; social networks; topic modelling; relations; less resourced languages; text classification; information extraction; natural language processing

1. Introduction

In the last few years, Twitter has become one of the most used social networks, creating new ways of consuming information but also of communicating and creating both leisure- and work-related explicit relationships. Furthermore, such relationships may be formed implicitly via the information we shared with our followers via retweets. Furthermore, Twitter has become a source of spontaneously generated textual data for many human languages, including less resourced languages such as Basque [1,2]. Thus, this data is becoming more and more useful for doing social research [3–5] which may complement traditional methods traditionally used in sociology. Furthermore, Twitter allows to obtain massive amounts of textual data to apply modern techniques of natural language processing (NLP) based on machine and deep learning, for tasks such as automatic analysis of opinions (opinion mining or sentiment analysis) [4,6], named entity recognition and lexical normalization [5], fake news and rumour detection [7], among others.

Taking the concept of liquid society [8] as a starting point, in this paper we present a multidisciplinary work that aims to contribute novel insights in social research by using and investigating techniques from natural language processing. Thus, the objective of this work is to provide a detailed social and demographic-based view of the most important topics and relationships

that are formed between the members of a specific community, namely, the community of Basque Twitter users classified by life stage (young and adult). By "Basque Twitter users" we refer to those users that are geolocalized in the area of the Basque Country (cultural) and that write at least the 20% of their tweets using the Basque language. This analysis will be entirely based on the automatically collected tweets.

More specifically, the work presented in this paper will consist of the following steps. First, we will identify the relevant Basque Twitter users. Second, a large corpus of tweets will be extracted using the identified users' timelines. Third, we will classify users in terms of life stage (adult/young) by classifying the writing style of the tweets in their timelines as formal or informal. This is due to the fact that we do not have metadata information about the users' age, which means that we have to somehow try to infer the life stage based on features about writing style. Five different classification methods will be investigated, including the use of machine learning techniques based on both feature-based and neural network architectures. We will use the best classifier to label the large corpus of Basque tweets by their writing style thereby classifying their users in terms of young or adult. In the final step, this classification will be used to investigate the most relevant topics among users of Twitter writing in Basque as well as for researching the social relations that emerge between each user type (young or adult). The final result will be a comprehensive, automatically obtained, sociological picture of the most important aspects within the community of Basque users of Twitter.

The main contributions of this work are the following. First, we collect and publish a corpus of 6 M tweets written in Basque and a manually annotated gold-standard of 1000 tweets in terms of the writing style, namely, formal or informal. Both corpora are made available to facilitate both NLP and social research in Basque (https://github.com/ixa-ehu/heldugazte-corpus), a less resourced high-inflected language. Second, we investigate the classification of Basque tweets as formal or informal using five different approaches including novel NLP techniques based on word embeddings, contextual character embeddings and neural networks. We believe that our work is the most comprehensive experimentation for the detection of users' age both in terms of number of users and NLP techniques used. An interesting insight from this part is the fact that, contrary to common views, the neural approach [9] obtains very competitive results despite being trained on a very small dataset. We believe that is due mostly to the combination of character-based contextual embeddings with static word embeddings. Third, the models of the best system will be made publicly available to facilitate their use and reproducibility of results (https://github.com/ixa-ehu/ixa-pipe-doc). Fourth, we perform the largest investigation of social relationships in terms of age stages for Basque communities based on an automatically classified corpus of almost 8000 users and six million tweets. Finally, we believe that this work shows how to make it possible to do meaningful and large scale social and NLP research for less resourced languages using texts from social media. For us, this work is just the beginning.

The rest of the paper is as follows. Next, we present the related work. Section 3 describes the process to collect our corpus while Section 4 contains the experiments performed for the classification of users by life stage. We use the obtained results to annotate our corpus and perform topic detection in Section 5. Section 6 analyzes the relations and communities of Twitter users extracted from the data and we offer some concluding remarks in Section 7.

2. Related Work

Analyzing demographic characteristics in social media is receiving increasing attention in the area of social media mining [3,10]. The popularity of Twitter has in fact benefited such approaches as it is possible now to mine spontaneous contributions and opinions of users about any kind of topic in many languages, including less resourced languages. Both social and linguistic aspects are important for our work as we will try to perform social and demographic analysis via large scale linguistic processing of tweets. The main techniques used here are those related to topic modelling, the study of social relationships and text processing or NLP.

With respect to topic modelling, many different approaches have tried to extract common information contained in large amounts of tweets scattered through the network, such as topic extraction with respect to an event and their related tweets [11], real-time classification of twitter trends [12] or to compare the content of Twitter with traditional news media [13]. Applying topic modelling to tweets, like for other tasks, requires some pre-processing to make the task appropriate for such short texts or documents [14].

In relation to the study of the social relationships that are generated within the network, closer to us are those studies that have aimed to identify communities of users based on their retweets. Among these, one can find studies about political polarization [15], political affiliation detection [16] or even studies about identifying communities in movements for independence [17]. In these studies, based on the retweets made by the user, it is shown that the identification of communities or groups is quite feasible. In this paper we will use similar methodologies to the ones mentioned here with the objective of uncovering latent communities.

The use of Twitter is very popular in NLP, where tweets are used for many tasks such as mining opinions about specific products or topics [4,18], analyzing stance detection and fake news [6,17] or in more low level NLP tasks such as POS tagging [19], Named Entity Recognition [5], normalization [20] and language identification [21]. Previous NLP work is relevant to us as we will work with tweets to perform text classification with the aim of labelling tweets according to their writing style (conventional-informal or formal).

Of particular interest to us is the body of work performed with the objective of age or life stage detection for Twitter users. Table 1 lists the most relevant work on this task. Most of them create their own datasets with manually annotated data to learn life stages or age ranges. As it can be seen the number of users varies from 300 to 3000. It is also important to mention that the best performing systems are those that use at most two [22,23] or three [10,24] labels denoting the life stages for the classification task.

Table 1. Most relevant systems for age detection in Twitter.

Reference	Corpus Size # Users	# Labels	Language
Rao et al. (2010)	1000	2	en
Al Zamal et al. (2012)	400	2	en
Marquart et al. (2014)	306	5	es
Nguyen et al. (2013)	3110	3	nl
Morgan-Lopez et al. (2017)	3184	3	en

For the task of age detection, most of previous works use both the text and the metadata provided by the Twitter API, most importantly, the age of each user. Previous work is based on logistic regression [10,24] and support vector machines [22,23,25]. Best results in this area have obtained around 86 word accuracy [24] for three age or life stages, although others scored well below that, most of them ranging around 74–80 word accuracy. These comparatively lower scores make it more difficult to perform meaningful social and demographic research using such automatic classifiers. Therefore, an important contribution of this work will be to strive in providing good performing classifiers for Basque tweets but also with the aim of developing robust and general enough models to be applicable for different text classification tasks across different languages.

3. Extracting a Large Corpus of Tweets from Basque Users

As we have already commented in the introduction, Twitter allows to obtain relatively large datasets also for less resourced languages. Before starting to collect the data for our work, we defined the community of users that will be the subject (or, in other words, the universe) of our study. In our case the choice was quite simple, namely, we wanted to use tweets from every Twitter user that publishes tweets in Basque. The task of identifying such users was facilitated thanks to UMAP,

a platform that monitors every tweet written in Basque. More specifically, UMAP includes a list of users who publish at least 20% of their tweets in Basque (https://umap.eus/). We used such a source to obtain an initial list of 8189 users.

The Twitter API was used via the tweepy package, choosing the timeline extraction mode to gather the last 3200 available tweets of each of the 8189 users in our sample. The data collection was performed during the days 30 and 31 of May 2018, gathering, after discarding some users due to API errors, more than 10 million (multilingual) tweets from 7980 users. We call this data the large corpus. Next, we classify the tweets by language using the metadata provided by the Twitter API, identifying those that are written in Basque. The result is a dataset of around six million tweets that we refer to as the Heldugazte corpus, which we split into personal tweets and retweets. The main statistics of the Heldugazte corpus are provided by Table 2 (The corpus is publicly available at https://github.com/ixa-ehu/heldugazte-corpus.).

Table 2. Characteristics of the Heldugazte corpus.

	Personal Tweets in Basque	Retweets in Basque
Tweets	3,171,785	2,891,136
Terms	1,434,050	813,833
Tokens	37,350,268	39,329,204

4. Classifying Users by Age Stage

In this section we present our work to classify Basque Twitter users as young or adult by analyzing their tweets' writing style. As we have seen in Section 2, previous work includes automatic approaches to infer demographic characteristics in the context of social networks and media, the most common being those related with genre and age [26]. Our objective of developing a classifier to characterizing Basque Twitter users according to two stages of life (young/adult) is therefore placed within that research line.

Furthermore, Section 2 also shows that most of previous work addresses the problem of classifying age stages using supervised machine learning techniques. This implies that some training data must be annotated according to the age or age stage of each user. Moreover, state of the art results indicate that classifying by age stages allows one to obtain better results than when the problem is formulated in terms of age ranges [24]. This seems to be due to the fact that shared experiences are easier to relate along time by using age stages rather than age ranges [3,27]. Therefore, in this work we decided to focus on classifying users in terms of two general age stages, namely, young and adult.

However, this decision encountered an early methodological problem. In order to annotate tweets written in Basque according to the age stages of their authors it is obviously necessary to have available some data providing such information. Unfortunately, for the large majority of users from which we mined tweets this information is not available. Thus, our decision to overcome this problem consisted of focusing on the information available on the tweets themselves by taking into account writing style features. After all, most of the previous work widely uses writing style based features in order to classify users in terms of age stages [10,22–24].

It has been suggested that adults tend to use more conventional language [24] whereas for young people it is more common to display repetitions and out of vocabulary words [10,22,28]. Generalizing on this idea, we will assume that writing style changes according to the age of the Twitter user. In other words, we will consider that adult writing style is more formal whereas young people's style can be seen as more informal. Therefore, we will be classifying users as young/adult by classifying their tweets in terms of formal or informal language.

4.1. Experimental Framework

Following the setting established above, we addressed the problem of classifying each tweet as formal or informal by means of several approaches. First, we applied a statistical method based on

perplexity [29]. Second, we experimented with several supervised methodologies: (i) a baseline method using sparse, one-hot word representations [30]; (ii) a feature-based method using word representations as continuous word vectors, namely, word embeddings, on top of the baseline developed in (i) [31–33]; (iii) an off-the-shelf system which uses clustering features for representing the documents [34,35] and (iv), a deep learning approach leveraging character contextual word embeddings and a neural network architecture [9].

We developed a classifier with each of the methods listed above and evaluate it on a held-out, gold-standard dataset manually annotated at tweet level for the categories of formal and informal. The method that obtained the best results in the experiments was then used to annotate personal tweets of the six million Heldugazte corpus (see Table 2) thereby classifying their authors as young or adult. In this sense, users were classified according to the writing style of the tweets in their timelines.

In the rest of this section we will describe the process of manual annotation of the training and evaluation data. After that, we describe the configuration settings of each of the systems used or implemented for the experiments.

4.1.1. The Heldugazte Gold Standard Corpus

In order to focus on learning to classify tweets according to writing style, we decided to clean the tweets keeping only those words that contained alphanumeric characters. Thus, we removed emoticons, hashtags, users' names (@) and URL links. Furthermore, we only considered those tweets that contained more than four tokens. We then randomly selected 1000 tweets sent from personal accounts from the 6 million tweet dataset collected as described in Section 3. One annotator manually labeled every tweet as formal or informal.

To manually label the tweets, a qualitative methodology was followed: all those tweets that contained out of vocabulary words or colloquial expressions were classified as informal. This methodology has been motivated by previous work on classifying formal and colloquial tweets [36]. Table 3 shows the main features of our manually annotated dataset. For the experiments described in the next five subsections, we split the gold standard corpus in three sets, leaving 65% for training and 35% for testing (the Heldugazte gold-standard corpus is publicly available at https://github.com/ixa-ehu/heldugazte-corpus). In the following we present two examples of the type of tweets we classify in this work. In both cases it can be seen the differences in writing style. Thus, in the informal tweets appear dialectal and/or slang forms (ein, examin, bau, det), whereas the formal tweet displays quite a standard Basque grammar.

Table 3. The Heldugazte gold standard corpus.

Total number of tweets	1000
Formal	492
Informal	508
Tokens in shortest tweet	5
Tokens in longest tweet	34
Token avg.	9.66

(1) Informal tweet : "inoizz ezdet ein mateko examin bau au baino okerro" (Informal: This is worst exam I have ever done.).
(2) Formal tweet: "killian jornet fenomenoa da zegamaaizkorri irabazi du beste behin non dago mendizale gazte honen muga" (Formal: Killian Jornet is a phenomenon he won the Zegama-Aizkorri again. Where is this mountaineer's limit?).

4.1.2. Perplexity-Based Distance

Perplexity is a widely-used evaluation metric for language models built with n-grams extracted from text corpora [37]. Furthermore, it has been used for specific tasks, namely, to classify between

formal and colloquial tweets [36] or for language identification between similar languages [29]. Both works inspired our approach to classifying tweets as formal or informal using perplexity. The former showed that perplexity could be useful to indirectly detect out-of-vocabulary words in tweets [36]. More importantly for us, the latter formally proposed the concept of perplexity-based distance between two languages.

A perplexity-based distance between two languages is established by "comparing the n-grams of a text in one language with the n-gram language model trained for the other language. Then, the perplexity of a test text T in language L2, given the language model LM of language L1, can be used to define the distance, between L1 and L2" [29]. According to this definition, low perplexity indicates close proximity between languages L2 and L1.

In order to classify tweets as formal or informal, we calculated the perplexity-based distance between a tweet with respect to a character-based seven-gram language model. If the perplexity-based distance between each tweet and the character-based n-gram language model is low, then we will classify it as formal and vice versa. Of course, for this approach to work we need to establish a threshold so that if the perplexity-based distance is lower than the threshold the tweet is labeled as formal and if higher, as informal.

Our method to calculate such threshold proceeds in the following manner. First, we built a seven-gram character-based language model using the data from a corpus composed of texts from the Basque Egunkaria and Berria newspapers. Second, we took the tweets in the training set (65% of the dataset) and classify each of them according to the perplexity-based distance with respect to the language model. We tried every value in the [0, 10] range with an increment of 0.1 as threshold. Third, we computed precision, recall and accuracy comparing the prediction of the threshold according to the perplexity value with respect to the gold-standard labels in the training data. Finally, the threshold chosen will be the value in which all the three metrics converge. As shown by Figure 1, we found out that the best threshold value was around 4.4. We will use this value in order to evaluate this approach in Section 4.2.

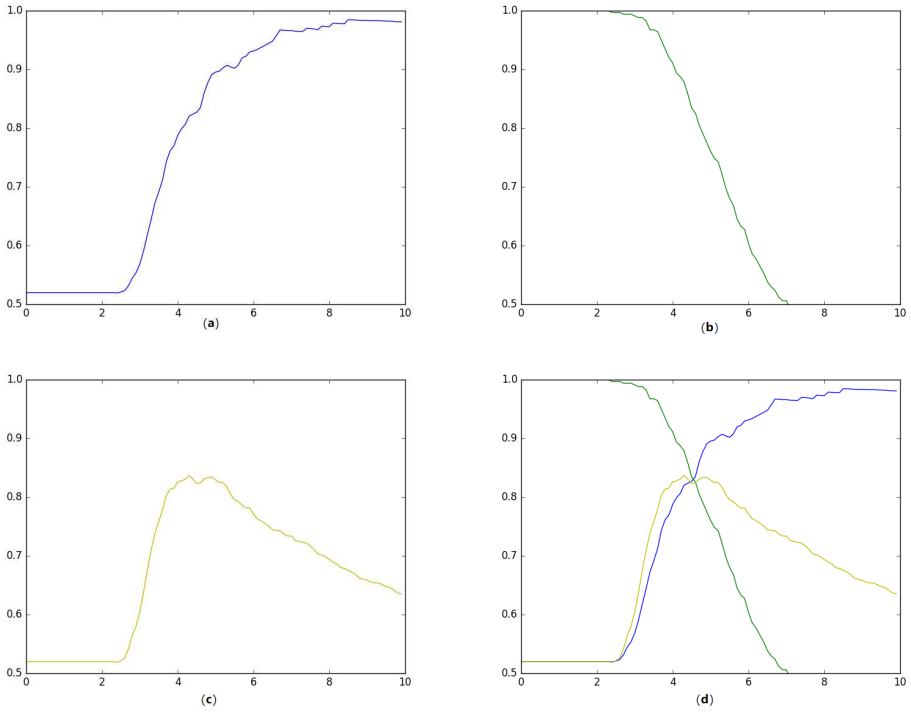


Figure 1. Finding the optimal threshold value (**d**) on the training data using precision (**a**), recall (**b**) and Accuracy (**c**) curves.

Table 4 displays the results of classifying the tweets in the training data using perplexity-based distance with threshold value at 4.4. The obtained overall accuracy using 4.4 as threshold was 0.831.

Table 4. Results of the perplexity-based approach on the training set.

Label	Error	Precision	Recall	F1
Informal	48	0.824	0.858	0.841
Formal	62	0.839	0.801	0.820

4.1.3. Supervised Baseline

The first baseline applying supervised machine learning was done to test various learning algorithms using the gold standard dataset described in Section 4.1.1. More specifically, we simply applied a bag of words representation which represented each document (tweet) according to the word frequencies in each document. The result is a very sparse word representation where the dimension of the vector representing each word is equal to the number of words in the corpus. In this setting, we apply six of the most common machine learning algorithms with default parameters using the scikit-learn library [30]. Table 5 reports their performance by means of five-fold cross validation on the training data.

Table 5. Bag of words results via five-fold cross-validation on the training set. Best result in bold.

Machine Learning Classifier (BoW)	Accuracy
5-NN (k-NN)	0.614
Decision Tree	0.677
Random Forest	0.707
Naive Bayes	0.765
Logistic Regression	0.775
SVM	**0.777**

As it can be seen, the best performing algorithms in this baseline setting were logistic regression and SVM. This is not surprising given the ability of SVM to perform well with very few labeled data. Next, we decided to experiment with a less-sparse representation of the tweets by means of pre-trained word embeddings and the SVM classifier.

4.1.4. Pre-Trained Word Embeddings

Distributed word representations or word embeddings are widely used nowadays in natural language processing. Several techniques have been proposed in order to obtain word embeddings, most of them based on the hypothesis that the meaning of a word is defined by the context in which it appears [31,32]. Thus, obtaining word embeddings usually requires large quantities of good quality training data which makes it difficult when experimenting with less resourced languages such as Basque. However, FastText provides pre-trained models for many languages, including Basque [33] by using the common crawl data (http://commoncrawl.org). The Basque model they distribute is trained on both common crawl and Wikipedia using CBOW with position-weights, in 300 dimension, with character n-grams of length 5, a window of size 5 and 10 negatives (https://fasttext.cc).

For this experiment, we mapped the words in the corpus to their real vector representation in the FastText model and average all the vectors with respect to the vocabulary. We optimized the C hyperparameter using accuracy and evaluated by five-fold cross-validation on the training data. Table 6 reports the detailed results of the five-fold cross-validation using $C = 1.1$ as hyperparameter.

Table 6. Support vector machine (SVM) (rbf and FastText embeddings) results via five-fold cross-validation on the training set.

Label	Error	Precision	Recall	F1
Informal	68	0.810	0.768	0.793
Formal	66	0.818	0.747	0.781

4.1.5. IXA Pipes

The document classification system included in the IXA pipes tools, ixa-pipe-doc, aims to establish a simple and shallow feature set, avoiding any linguistic motivated features, with the objective of removing any reliance on costly extra gold annotations and/or cascading errors if automatic annotations are used. The underlying motivation was to obtain robust models to facilitate the development of document classification systems for several languages, datasets and domains while obtaining state of the art results.

The ixa-pipe-doc, as a component of IXA pipes, includes a simple method to combine various types of clustering features induced over different data sources or corpora. This method has already obtained state of the art results in several tasks such as newswire named entity recognition [34] and opinion target extraction [35], both in out-of-domain and in-domain evaluations and for several languages, including Basque. Clusters of words provide denser document representations. Although still a one-hot vector representation, the dimensions of the representation gets reduced to the number of clustering classes used. This is done by mapping the words in the document to the words in each of the clustering lexicons [38].

We will use the three types of simple clustering features based on unigram matching that ixa-pipe-doc implements: (i) Brown [39] clusters, taking the 4th, 8th, 12th and 20th node in the path; (ii) Clark [40] clusters and, (iii) Word2vec [31] clusters, based on K-means applied over the extracted word vectors using the skip-gram algorithm. The implementation of the clustering features looks for the cluster class of the incoming token in one or more of the clustering lexicons induced following the three methods listed above. If found, then we add the class as feature. The Brown clusters only apply to the token related features, which are duplicated (More details, including examples of the features are provided in Agerri and Rigau [34,35]).

For the experiments, we used pre-trained clusters using the Elhuyar Web Corpus [41] and from a 600 M word corpus obtained from crawling local news sites (local news corpus). The number of clusters trained with each algorithm and data source was the following: 100–800 clusters using the Clark and Word2vec methods, and 1000, 2000 and 3200 classes with the Brown algorithm. The best combination of features was obtained by performing every possible permutation between them in a five-fold cross validation setting using the gold standard training data. Following this methodology, the best configuration consisted of the features listed in Table 7 (we did not use any local or lexicon-based features, just the clustering representations).

Table 7. Source data and number of clusters used with ixa-pipe-doc system. EWC: Elhuyar Web Corpus. LNC: local news corpus.

Cluster Type	Corpus–# Clusters
Brown	EWC-3200
Clark	EWC-600 & LNC-300
Word2vec	EWC-300 & LNC-500

Table 8 provides the results per class using ixa-pipe-doc. As it can be seen, they are the best results obtained so far both in terms of accuracy (0.887) and F1 measure.

Table 8. The ixa-pipe-doc results via five-fold cross-validation on the training set.

Label	Error	Precision	Recall	F1
Informal	32	0.892	0.886	0.889
Formal	30	0.883	0.889	0.886

4.1.6. Flair

Flair refers to both a deep learning toolkit based on neural networks and to a specific type of character-based contextual word embeddings [9]. Unlike static word embeddings such as those of Word2vec [31], Glove [32] or FastText [33], contextual embeddings allow one to obtain word representations in a vector space taking into account the sense of the word given the context in which it appears. Thus, while a polysemous word would be given a unique real valued representation in the FastText pre-trained word embedding model used in Section 4.1.4, Flair embeddings will aim to provide different representations depending on the contextual meaning of the word. Another important difference is that flair embeddings are not word-based, they are trained by modeling words as sequences of characters.

Flair embeddings have been successfully applied to sequence labelling tasks obtaining best results for a number of public benchmarks [9]. In this paper, we apply the Flair toolkit to train document classification systems for classifying tweets as formal or informal. Flair provides a recurrent neural network (RNN) architecture (Cho et al., 2014) to represent documents, modelling text as a sequence of characters passed to the RNN which at each point in the sequence is trained to predict the next character [9]. We used this architecture to train document classification systems using several pre-trained word embedding models: flair contextual embeddings for Basque, character embeddings and the FastText Basque embeddings used previously. The flair contextual embeddings for Basque were trained on various sources and it contains around 249M words. We performed five-fold cross-validation on the training data obtaining the best results with a combination of the flair and the FastText embeddings, obtaining 0.808 in word accuracy. Table 9 reports on the final cross-validation results for each label.

Table 9. Flair results via five-fold cross-validation on the training set.

Label	Error	Precision	Recall	F1
Informal	61	0.898	0.759	0.823
Formal	65	0.730	0.878	0.792

4.2. Experimental Results

In order to finish our experiments, we used the full training set for the systems that obtained best cross-validation results and evaluated them on the gold standard test set. Thus, we tested the following systems: the SVM model with FastText word embeddings, the perplexity-based distance method, ixa-pipe-doc and Flair. Table 10 reports the final results of our experiments to obtain a good classifier of Basque tweets according to writing style (formal/informal).

It should be noticed that we did not implement any specific features for the experiments with our gold standard data. The main reason is that we wanted to avoid including any features that might cause overfitting to the training data. By doing so, we were aiming to develop general, robust classifiers that hopefully will be equally competitive for other languages, text genres and tasks. We believe that the strong results obtained by the IXA pipes and the Flair system, despite the lack of any specific tuning to the data, show the generalization power of using combined word representations. Furthermore, it is particularly interesting the fact that the neural network architecture provided by flair performed so well with such a small training data. Our hypothesis is that, being the dataset so small, there are not so many out of vocabulary words.

Table 10. Final evaluation results on the test set. Best results in bold.

System	Accuracy	Label	Error	Precision	Recall	F1
Perplexity	0.825	Informal	26	0.805	0.847	0.825
		Formal	35	0.848	0.806	0.826
SVM-FastText	0.832	Informal	24	0.843	0.823	0.836
		Formal	33	0.834	0.828	0.829
IXA pipes	0.886	Informal	20	0.882	0.881	**0.882**
		Formal	20	0.889	0.888	**0.889**
Flair	0.866	Informal	22	0.869	0.858	0.863
		Formal	24	0.868	0.877	0.872

We believe that the developed classifiers are good enough to use them to classify the personal tweets in the six million Heldugazte corpus in terms of the formal and informal classes. We decide to use the IXA pipes model due to the results obtained and its lower requirements in terms of memory and computing power (Flair requires longer time and a GPU for training and tagging).

4.3. Labelling the Large Corpus

The young/adult classifier developed in the previous section will allow us to compare the main features of those two life stages by taking into account the topics that appear on users' tweets and the relations that arise between them. After training the best performing model of ixa-pipe-doc described in Section 4.1.5 and evaluated in Section 4.2 on the 1000 tweets of the *full gold-standard corpus*, we proceed to tag the personal tweets in the Heldugazte corpus (see Table 2). We only used for classification the (multilingual) timelines of those users which contained at least 10 tweets written in the Basque language, namely, 7.087 users out of the 7.980 that we crawled in Section 3.

Given that the classifier labels each tweet as *formal* or *informal*, we decided, after some qualitative error analysis, that those users whose timelines in which more than 45% of the tweets are labelled as *informal* be considered as young, and adult otherwise. Thus, after classifying users in terms of young and adult, we obtained 5508 which were adult users and 1579 labelled as young users, applying the obtained label to every tweet and retweet of each user's timeline, namely, to every tweet in the Large Corpus. Details of the tagging results are displayed in Table 11. Even though the resulting classification is quite unbalanced that is not a problem because we will perform our analysis of each type of user independently.

Table 11. Classifying tweets in large corpus in terms of age stage (young/adult).

	Adult	Young
Users	5508	1579
Tweets (personal)	4,046,512	1,128,124
Retweets	4,345,500	963,668
Tweets in Basque (topics)	2,634,534	530,226
Retweets in Basque (relations)	2,421,058	400,448

4.3.1. Analyzing Adult Twitter Users

The first issue that arises when looking at the adults' tweets is that more than half of them are actually retweets. This shows that there is a disposition to share information as much as to generate it. Furthermore, Figure 2 shows that Basque is the most used language, Spanish being the second. Finally, it is also interesting the fact that the use of Spanish increases for the retweets which probably reflects the fact that publicly available and shareable information in Spanish is much larger than in Basque.

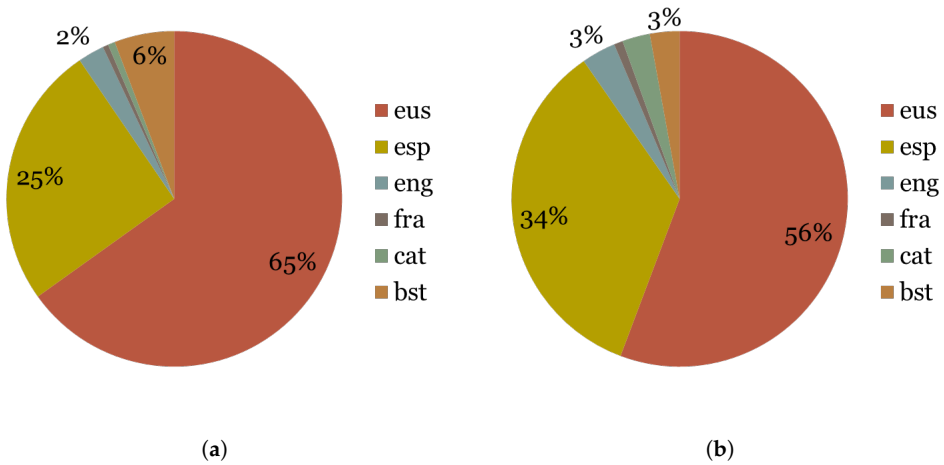


Figure 2. Adults in the large corpus: 5508 users. (**a**) Adult personal tweets. (**b**) Adult retweets.

If we take a look at the tweets written only in Basque, we find that 2,634,534 tweets have been published, containing more than 32 million tokens. Figure 3 shows their distribution according to their length in tokens. The average tweet contains 12 tokens. Standard deviation with respect to the average was 5.65. Furthermore, median corresponds to 12, mode being 14 token (more than 200,000 tweets).

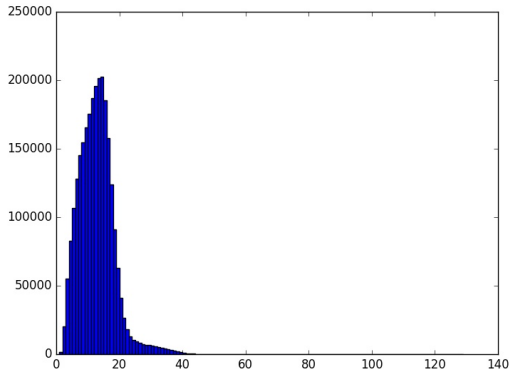
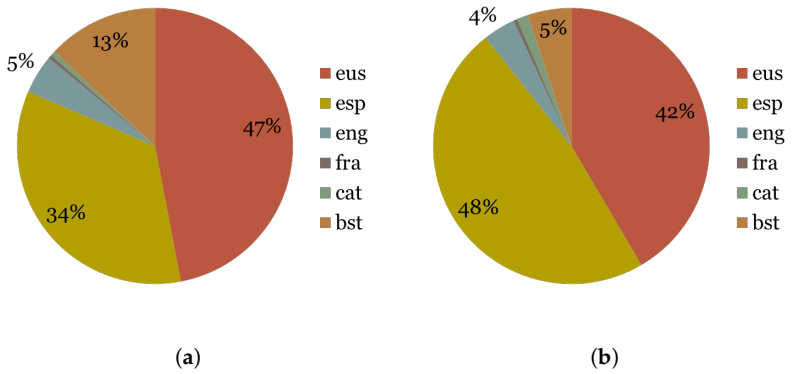


Figure 3. Distribution of Basque tweets published by adults.

4.3.2. Analyzing Young Twitter Users

The young users corpus contains just one quarter in size with respect to the corpus of adult users. In this case, it can be seen that there were fewer retweets (963,668) than original personal tweets (1,128,124). It is perhaps more noticeable the fact that the use of Basque is less common between young users: 18% lower for tweets and 24% lower for retweets. As it is expected, the use of Spanish is higher between young users, reaching 47% for retweets and 34% for personal tweets. The data in Figure 4 shows that Basque is less used between young users of Twitter.

If we take a look at the tweets written only in Basque, we find that 530,226 personal tweets have been published, containing more than five million tokens. Figure 5 shows their distribution according to their length in tokens. The average tweet contains nine tokens, whereas the standard deviation with respect to the average is 5.49. Furthermore, median corresponds to eight, mode being six token (around 45,000 tweets). In general, it is noticeable the fact that young users published much shorter tweets than adults.



(**a**) (**b**)

Figure 4. Young users in large corpus: 1579 users. (**a**) Young personal tweets. (**b**) Young retweets.

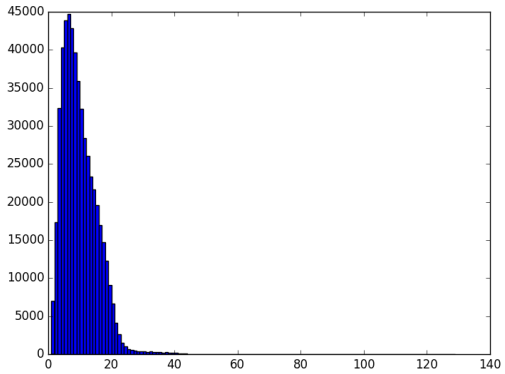


Figure 5. Distribution of Basque tweets published by young users.

5. Topics

The aim of this section is to detect the most frequent topics of Twitter users writing in Basque. In order to do so, we start with the personal tweets of users classified as adults and young in the previous section using the classifier developed in Section 4.1.5. As shown by Table 11, more than three million personal tweets in Basque were obtained. On the one hand, 2,634,534 personal tweets will be used to predict adult topics and, on the other hand, 530,226 personal tweets for young users. We will apply topic modelling via latent dirichlet allocation (LDA) [42] in order to infer relevant topics per type of user from unstructured data. Specifically, we used the implementation of LDA provided by the gensim package [43].

Topic modeling is a commonly used tool in the field of text mining. With this technique, words are grouped or clustered according to their context, thereby generating more general contents or concepts via the clustered words and allowing to identify general topics from specific words. We apply this technique to identify the main subjects that appear in the Basque users' tweets. In order words, LDA makes it possible to classify events that are observable in latent or hidden clusters. This is possible thanks to the many hidden features, such as the similarity of words.

Before applying LDA, we need to structured the documents in our dataset. It is difficult to directly apply LDA to our data due to the short length of tweets. Thus, we decide to group the tweets by user, namely, we will create one document per user where the document will contain every tweet published by that person [13,14]. Additionally, and considering that Basque is an agglutinative language, we decided to lemmatize the documents to reduce the number of terms that needed to be modelled. For this pre-processing step, the IXA pipes Basque lemmatizer was used [44].

LDA requires us to choose a number of topics beforehand. After several tests, 20 topics were used for the adults documents and 12 for the young ones. The different in topics is coherent with the number of tweets for each of user type. Although there is not a fixed correct number of topics, this choice affects the interpretability of the LDA results [45,46]. Thus, it is interesting to achieve the most dispersed possible model so that the overlap between topics is kept to a minimum. Furthermore, the resulting topics should correlate with social reality which is latent in the real data.

The results of applying LDA are displayed using LDAvis [47], which offers an easy interpretation of each topic. As it is customary, the identity or meaning of the topic is determined by the words of which it is composed [45]. Thus, Table 12 shows the topics obtained for the adult users whereas Table 13 lists the topics for the young users.

Table 12. Topics of adult users. Whenever necessary, English translation is provided below each row of representative words.

Topics of Adult Users	Representative Words in the Topic	% of Words
1 Conversation	entzun, iruditu, bizitza, pentsatu, pasatu listen, imagine, life, think, pass	10.5
2 Politics	Euskal Herri, espainia, politiko, estatu, eskubide Basque Country, Spain, politics, states, rights	10.0
3 Basque tweeters	@txargain, @berria, @boligorria, euskara, idatzi @user, @newspaper, @user, Basque, write	6.9
4 Cultural offer	lehiaketa, sarrera, ikastaro, erakusketa, antzerki competition, entry, course, exhibition, theater	6.4
5 Public administration	udal, zerbitzu, publiko, aurrekontu, euskadi municipal, services, public, budget, euskadi	6.1
6 Basque television	@euskaltelebista, urhanditan, @xabiermadariaga, herritxiki @television, TV program, @journalist, TV program	5.3
7 Tournaments	txapelketa, final, kirol, jokatu, kanporaketa championship, final, sports, play, playoffs	5.0
8 Basque prisoners	preso, herri, espetxe, iheslari, elkartasun prisoner, people, prison, fugitive, solidarity	4.9
9 Culture	liburu, literatura, filma, poesia, dokumental books, literature, film, poetry, documentary	4.8
10 Social movements	feminista, asanblada, gaztetxe, borroka, langile feminist, assembly, squatted house, fight, worker	4.8
11 Education	ikasle, hezkuntza, irakasle, ikastola, ikastetxe students, education, teachers, Basque colleges, schools	4.3
12 Science	euskara, artikulu, interesgarri, zientzia, teknologia Basque, articles, interesting, science, technology	4.1
13 Music	kontzertu, disko, talde, entzun, musika concert, disc, group, listen, music	3.9
14 Basque language	euskara, hizkuntza, euskaldun, euskal, ikasi Basque language, language, Basque speaker, Basque, learn	3.8
15 Sports	talde, real, partida, irabazi, jokatu team, real, match, win, play	3.8
16 Gipuzkoa (Province)	tolosa, andoain, hernani, ordizi, beasain (Cities in the province of Gipuzkoa)	3.7
17 Media in Basque	@berria, @euskalirratia, @argia, @zebrabidea, @iehkohitza	3.5
18 Donostia (City)	donostia, @donostiakoudala, ezagutu, gipuzkoa Donostia, City Hall of Donostia, meet, Gipuzkoa	3.5
19 Nafarroa (Province)	nafarroa, baztan, altsasu, irunerri, irun (Cities in the province of Navarre)	2.7
20 Bizkaia (Province)	larrabetzu, lekeitio, durango, bermeo, arrasate (Cities in the province of Bizkaia)	2.6

Table 13. Topics of young users. Whenever necessary, English translation is provided below each row of representative words.

Topics of Young Users	Representative Words in the Topic	% of Words
1 Gipuzkera dialect (informal chat)	in, ne, oain, atxalde, biyar do, mine, now, late, tomorrow	14.7
2 Express feelings	maite, amets, gau, bizi, bihotz love, dream, night, live, heart	11.4
3 Bizkaiera dialect (informal chat)	dau, be, ein, dot, emun, bixar is, also, do, have, give, tomorrow	10.8
4 Sports	partidu, irabazi, jokatu, txapeldun, etapa match, win, play, champion, stage	9.9
5 Cultural activities	areto, antzoki, gaztetxe, tailer, kontzertu halls, theaters, youth clubs, workshops, concerts	9.7
6 To congratulate	zorion, pasatu, animo, eskerrikasko, polit congratulations, pass, courage, thank you, nice	9.3
7 Tell the life	jajaja, bihar, ohera, partido, ikasi Hahaha, morning, to bed, party, study	7.6
8 Bizkaiera dialect (formal chat)	dot, dau, barri, barik, be have, is, new, without, too	7.1
9 Gipuzkera dialect (formal chat)	det, ne, hoi, iruditu, irakurri do, mine, that, seem, read	7.1
10 Basque prisoners	herri, euskal, etxe, preso, gazte people, Basque, house, prisoner, youth	6.4
11 Athletic CB (football team)	aupa, athletic, @athletic, san mames, bilbo	3.3
12 Rowing	sailkapen, jardunaldi, maila, txapelketa, estropada classification, event, level, championship, regatta	2.7

The topics from adult users show that they mostly talk about politics, social, cultural and linguistic (Basque language-related) issues. It is also interesting to notice that public institutions also appear in the social network, such as Basque Country regional offices from Gipuzkoa, Bizkaia or Nafarroa. However, if we look at the topics most common between young users (Table 13) we can see that they are mostly related to everyday affairs (chatting between friends, expressing feelings and emotions with respect to something, etc.). In some cases, they talk about everyday issues using their local Basque dialect (e.g., Gipuzkera and *Bizkaiera*). Additionaly, sports (athletic club, rowing) are also a recurring theme among young people. Thus, comparing the two different age stages, it should be noted that young people use Twitter for more day-to-day activities among friends or among contacts within the network. In the case of adults, it is clear that there is more political and social content.

6. Relations

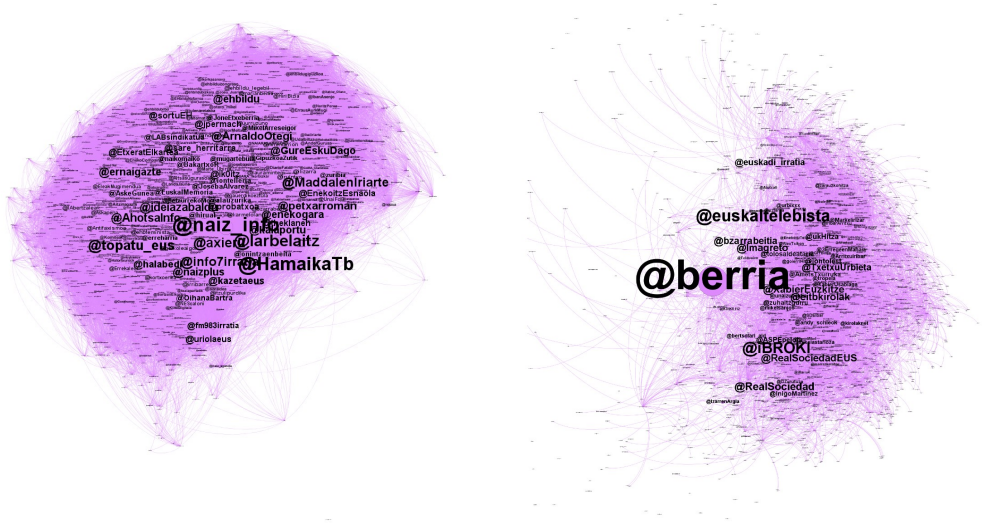
In this section we will study the relations that appear between Basque users of Twitter. As in the previous section, the starting point will be the retweets of users classified as adult or young using the classifier developed in Section 4.1.5. The number of retweets for each type of users are reported in Table 11. Specifically, 2,421,058 retweets will be used to study the relations that are created between adult users. With respect to young users, the corpus consists of 400,448 retweets. The overall objective is to uncover the most important relations formed between users by studying the links between their retweets and comparing the different behaviour and relations of young and adult users.

First, we created a giant graph using the data (retweets) for each type of user. To build the graph, two features extracted from each retweet were used: (i) who retweets and (ii) who has been

retweeted. Thus, the data source will be the source and the target of each retweet. Based on these two characteristics, the two graphs were created using the gephi program [48].

Once the graphs were created, we proceed as follows: first, we divided each graph into subgroups using the modularity algorithm [49]. Second, we gave the network a spatial structure by using the ForceAtlas2 algorithm [50], ordering the nodes according to the community to which they belong. Finally, the identity of each subgroup or community was defined, using the most important nodes of the created subgroups. By doing so it can be seen how the communities within each graph are structured.

Thus, in the adult graph we identified 33,277 nodes whereas 24,987 nodes were identified among the young users. Once the two independent network graphs, one for young users and another one for adults, were created, we split each network into subgroups to analyze how the communities or subgroups inside each one are formed. The subgroups are defined based on their most important nodes. The aim is to uncover, via their retweets, the relations that are created between the Basque users of Twitter. As an example of the graphs obtained, Figure 6 displays the two most important subgroups within the adult and young graphs, respectively.



(**a**) (**b**)

Figure 6. Example graphs of the two most important subgroups within the adult and young graphs. (**a**) Most important community (nationalist left) in the adult graph. (**b**) Most important community (sports) in the young graph.

6.1. Relationships in the Adult Graph

Table 14 shows the five main subgroups derived from the graph of adult users. Each of the subgroups displays a common characteristic, namely, that all of them have a direct relation with topics or issues related to the Basque Country. Thus, it can be seen that for adult users Basque language is used to talk mostly about Basque issues, highlighting politics (Separatist left) and current affairs (news). In other words, the main function is to talk about politics and social issues but with a clear focus on the Basque community itself (rather than on international affairs). In the following we describe the main characteristics of each of the subgroups contained in the graph of adult users.

- Nationalist left (27.92%): this subgroup is made up of nodes with a specific political orientation, mostly related to members of the Nationalist Left. In addition to the users that appear in the first column of Table 14, there are also many important nodes that refer to specific users (@ArnaldoOtegi, @jpermach, @JosebaAlvarez...) or institutions (@sortuEH, @LAB...) of

the Nationalist Left. This is the main group, joined by more than a quarter of all nodes that corresponds to the relationship of a certain political orientation.

- News (23.77%): this group, related to news, consists of almost a quarter of all users. Most of the nodes of this subgroup are related to the media, specially several users related to the Basque public television (EITB).
- Basque language (15.34%): in the third subgroup, there are topics related to the Basque language, such as communication media in Basque (@zuzeu, @Gaztezulo, @ArabakoALEA), associations for the promotion of the Basque language (@AEK_eus, @EHEbizi...) as well as several individuals related to the Basque language (@KikeAmonarriz, @KoldoTellitu, @MertxeMugika).
- Music and GED (13.56%): in the fourth subgroup there is a special phenomenon, since it brings together two different groups in the same subgroup. The first one is related to music, since we can appreciate different users related to the music scene (@EsneBeltza, @ZuriHidalgo, @ZeEsatek, @40minuturock, @hesiantaldea, @ItzrrSemeak...). The second one is related to the users of the social movement "Gure Esku Dago" (@GoierrikoGED, @GEDTolosaldea, @GureEskuDagoDon...).
- Basque tweeters (13.10%): in this last subgroup we can find popular Basque users of Twitter, which are important within the Basque community due to their large number of followers or retweets.

Table 14. Most important nodes for the subgroups in the graph of adult users.

Nationalist Left	News	Basque Language	Music and GED	Basque Users
@naiz_info	@berria	@zuzeu	@XMadariagaI	@boligorria
@HamaikaTb	@eitbAlbisteak	@KikeAmonarriz	@gaizkapenafiel	@zaldieroa
@larbelaitz	@euskaltelebista	@Sustatu	@JGGarai	@urtziurkizu
@topatu_eus	@euskadi_irratia	@Gaztezulo	@EsneBeltza	@landergarro
@axierL	@tolosaldeataria	@AEK_eus	@UrHanditan	@ielortza

6.2. Relationships in the Young Graph

The subgroups in the graph of young users (see Table 15) display both similarities and differences with respect to the adult graph.

Table 15. Most important nodes for the subgroups in the graph of young users.

Sports	Basque Language	Nationalist Left	News	Music
@berria	@enekogara	@naiz_info	@argia	@berritxarrak
@euskaltelebista	@GureEskuDago	@larbelaitz	@HamaikaTb	@gaztea
@iBROKI	@EsaldiakEuskara	@topatu_eus	@eitbAlbisteak	@izanpirata
@RealSociedad	@ZuriHidalgo	@ArnaldoOtegi	@MaddalenIriarte	@eitbeus
@XabierEuzkitze	@MeriLing1	@ernaigazte	@ielortza	@LeakoHitza

Focusing on the similarities, Basque language, Nationalist left and News are important subgroups in both graphs. These common subgroups can be related to politics and immediacy, which are basic characteristics of identity in Twitter. With respect to the differences, it noticeable that subgroups related to leisure take a more central stage, such as sports and music. Moreover, it is worth pointing out that young Basque users take Twitter as a channel to comment on everyday issues. Finally, the main topics in each of the subgroups within the young users graphs are listed in the following:

- Sports (21.61%): this subgroup, which includes most of the nodes which are considered roles models for the youths, is related to sports. The group is be made up of sports teams or organizations (@RealSociety, @RealSociedadEUS, @ASPEpelota, @SDEibar...), as well as its athletes (@InigoMartinez, @AmetsTxurruka, @XabierUsabiaga, @Markelirizar...). However, the most important nodes are sports journalists (@iBROKI, @XabierEuzkitze, @Imagreto, @bzarrabeitia,

@TxetxuUbieta...) and the media (@berria, @euskaltelebista, @eitbkirolak, @euskadi_irratia...). Once again, it can be clearly seen that the newspapers and TV media are the most important nodes.

- Basque language (20.70%): a fifth of all the nodes are in this subgroup. The most important ones are those directly related to the Basque language (@EsaldiakEuskara, @euskarazEH, @Bertsotan, @bertsolaritza, @Euskeraz_Bizi...). In the adult graph it was also found a community related to this topic, although the most important nodes are markedly different.
- Nationalist left (17.12%): this third group, composed of nodes related to the nationalist left, is perhaps the most similar in both adult and young graphs. For example, media (@naiz_info, @topatu_eus, @inform7irratia, @naizplus...), organizations (@ernaigazte, @ehbildu, @sortuEH...) and individuals (@ArnaldoOtegi, @lauramintegi...) related to the nationalist left, appear in both subgroups.
- News (14.92%): as in the previous subgroup, this community is also very similar for both young and adult users. The most important nodes correspond to general news Basque media (@argia, @HamaikaTb, @eitb Albums, @zuzeu, @Gaztezulo).
- music (11.35%): In this final subgroup, although quite heterogeneous, it can be said that the most important nodes are related to music. Among these, the music related media (@gaztea, @DidaGaztea), music bands (@berritxarrak, @muguruzafm, @Glaukomaband), as well as record companies (@BagaBigaeus) are the main nodes.

To finish this section, Table 16 offers an overview of the main subgroups per type of user. Both graphs create communities related to political and social issues, although it is more important for the graph of young users. Moreover, in both types of users Basque language is an important subgroup and it shows that users write in Basque mostly about issues or topics directly related to the Basque Country. The main difference between both types of users is the importance of the sports subgroup which is not present in the graph of adult users. This shows the influence of sports celebrities as role models among young people.

Table 16. Communities in each graph of users.

Subgroups in Graph of Adult Users	**% of Nodes**
Nationalist left	27.92
News	23.77
Basque language	15.34
Music and GED	13.56
Basque tweeters	13.10
Subgroups in Graph of Young Users	**% of Nodes**
Sports	21.61
Basque language	20.70
Nationalist left	17.12
News	14.92
Music	11.35

7. Conclusions and Future Work

This paper presents an approach to social research of speakers of a less resourced language based on applying Natural Language Processing and topic detection techniques to a large dataset of tweets written in Basque. The collected dataset consists of more than six million tweets from almost 8000 users. This demonstrates that Twitter is a valuable source of spontaneously generated textual data for research on language use and for the discovering of latent social interactions.

Furthermore, we present a comparison of five different approaches to classify Twitter users by life stage, namely, whether they are adult or young people. As age-related information was not present in the collected data, we decided to focus instead on classifying tweets by writing style, assuming that most of the informal tweets are written by young people and vice versa. In order to develop our

classifiers, we manually annotated 1000 tweets with the labels formal and informal, and experimented with modern feature-based and deep learning techniques for text classification based on vector-based word representations (word embeddings), clustering features and neural networks. We believe that we provide one of the most competitive approaches for age (or life stage) detection. In our view, our approach is robust enough to be usable across tasks, languages and datasets.

Using the best classifier we automatically tagged the personal tweets contained in the six million corpus, classifying a user's timeline as young if more than 45% of the tweets were classified as informal, and as adult otherwise. This allowed us to classify almost 8000 Basque Twitter users by their life stage, allowing to obtain meaningful insights with respect to the most important topics per life stage. Moreover, by using the information provided by the users' retweets, it also facilitated the uncovering of the most important relationships formed within the social network, helping to understand better the differences in social behaviour between young and adult Basque users of Twitter.

More specifically, it was found out that most Basque young users use Twitter to communicate with people close to them about everyday aspects of life, whereas for adults the most important topics were those related with politics and social issues. Furthermore, we were able to characterize quite clearly the different communities that get implicitly formed within the network. Thus, although there are certain similarities, one clear difference was the fact that young people form communities around sport celebrities (footballers and so on) and media.

Future work will include the use of cross-lingual embeddings to improve results with the neural network system, researching methods to obtain the relationship graphs (semi)-automatically and add a temporal dimension to the analysis perform in this paper, so we can see how relations and topics change over time.

We publicly distribute the data collected as well as the best classifiers developed to help promoting both sociological and linguistic processing research in a less resourced high-inflected language such as Basque and about Basque speakers. We believe that the work presented in this paper is applicable to other languages around the world.

Author Contributions: conceptualization, J.F.d.L., R.A. and I.A.; methodology, J.F.d.L., R.A. and I.A.; software, J.F.d.L., R.A. and I.A.; formal analysis, J.F.d.L., R.A., I.A.; data curation, J.F.d.L. and R.A.; writing—original draft preparation, R.A. and J.F.d.L.; writing—review and editing, R.A.; supervision, R.A. and I.A.

Funding: The second author is funded by the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE), under the project CROSSTEXT (TIN2015-72646-EXP) and the Ramon y Cajal Fellowship RYC-2017-23647. He also acknowledges the support of the BBVA Big Data 2018 "BigKnowledge for Text Mining (BigKnowledge)" project.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cunliffe, D. Minority Languages and Social Media. In *The Palgrave Handbook of Minority Languages and Communities*; Springer: Berlin, Germany, 2019; pp. 451–480.
2. Leivada, E.; D'Alessandro, R.; Grohmann, K.K. Eliciting big data from small, young, or non-standard languages: 10 experimental challenges. *Front. Psychol.* **2019**, *10*, 313. [CrossRef] [PubMed]
3. Nguyen, D.; Doğruöz, A.S.; Rosé, C.P.; de Jong, F. Computational sociolinguistics: A survey. *Comput. Linguist.* **2016**, *42*, 537–593. [CrossRef]
4. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 502–518.
5. Baldwin, T.; de Marneffe, M.C.; Han, B.; Kim, Y.B.; Ritter, A.; Xu, W. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In Proceedings of the Workshop on Noisy User-generated Text, Beijing, China, 31 July 2015; pp. 126–135.

6. Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; Cherry, C. SemEval-2016 task 6: Detecting stance in Tweets. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, San Diego, CA, USA, 16–17 June 2016; pp. 31–41.
7. Derczynski, L.; Bontcheva, K.; Liakata, M.; Procter, R.; Hoi, G.W.S.; Zubiaga, A. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 69–76.
8. Bauman, Z. *Liquid Modernity*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
9. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual string embeddings for sequence labeling. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1638–1649.
10. Morgan-Lopez, A.A.; Kim, A.E.; Chew, R.F.; Ruddle, P. Predicting age groups of Twitter users based on language and metadata features. *PLoS ONE* **2017**, *12*, e0183537. [CrossRef]
11. Hu, Y.; John, A.; Wang, F.; Kambhampati, S. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, ON, Canada, 22–26 July 2012.
12. Zubiaga, A.; Spina, D.; Martínez, R.; Fresno, V. Real-time classification of twitter trends. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 462–473. [CrossRef]
13. Zhao, W.X.; Jiang, J.; Weng, J.; He, J.; Lim, E.P.; Yan, H.; Li, X. Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*; Springer: Berlin, Germany, 2011; pp. 338–349.
14. Hong, L.; Davison, B.D. Empirical study of topic modeling in twitter. In Proceedings of the First Workshop on Social Media Analytics, Washington, DC, USA, 25–28 July 2010; pp. 80–88.
15. Conover, M.D.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; Flammini, A. Political polarization on twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011.
16. Pennacchiotti, M.; Popescu, A.M. Democrats, republicans and starbucks afficionados: User classification in twitter. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 430–438.
17. Zubiaga, A.; Wang, B.; Liakata, M.; Procter, R. Stance classification of social media users in independence movements. *Catalonia* **2017**, *2*, 10–960.
18. Villena Román, J.; Lana Serrano, S.; Martínez Cámara, E.; González Cristóbal, J.C. *Tass-Workshop on Sentiment Analysis at SEPLN*; The Spanish Society for Natural Language Processing: Jaén, Spain, 2013.
19. Ritter, A.; Clark, S.; Etzioni, O. Named entity recognition in tweets: An experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 1524–1534.
20. Alegria, I.; Aranberri, N.; Comas, P.R.; Fresno, V.; Gamallo, P.; Padró, L.; San Vicente, I.; Turmo, J.; Zubiaga, A. TweetNorm: A benchmark for lexical normalization of Spanish tweets. *Lang. Resour. Eval.* **2015**, *49*, 883–905. [CrossRef]
21. Zubiaga, A.; San Vicente, I.; Gamallo, P.; Pichel, J.R.; Alegria, I.; Aranberri, N.; Ezeiza, A.; Fresno, V. Tweetlid: A benchmark for tweet language identification. *Lang. Resour. Eval.* **2016**, *50*, 729–766. [CrossRef]
22. Rao, D.; Yarowsky, D.; Shreevats, A.; Gupta, M. Classifying latent user attributes in twitter. In Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, Toronto, ON, Canada, 26–30 October 2010; pp. 37–44.
23. Al Zamal, F.; Liu, W.; Ruths, D. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. *ICWSM* **2012**, *270*, 2012.
24. Nguyen, D.; Gravel, R.; Trieschnigg, D.; Meder, T. "How old do you think I am?" A study of language and age in Twitter. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media ICWSM, Cambridge, MA, USA, 8–11 July 2013.
25. Marquardt, J.; Farnadi, G.; Vasudevan, G.; Moens, M.F.; Davalos, S.; Teredesai, A.; De Cock, M. Age and gender identification in social media. In Proceedings of the CLEF 2014 Evaluation Labs, Sheffield, UK, 15–18 September 2014; pp. 1129–1136.
26. Cesare, N.; Grant, C.; Nsoesie, E.O. Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv* **2017**, arXiv:1702.01807

27. Eckert, P. Age as a sociolinguistic variable. In *The Handbook of Sociolinguistics*; Blackwell Publishing: Hoboken, NJ, USA, 2017; pp. 151–167.
28. Rosenthal, S.; McKeown, K. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; Association for Computational Linguistics: Portland, OR, USA, 2011; Volume 1, pp. 763–772.
29. Gamallo, P.; Pichel, J.R.; Alegria, I. From language identification to language distance. *Phys. A Stat. Mech. Appl.* **2017**, *484*, 152–162. [CrossRef]
30. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
31. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2013, pp. 3111–3119.
32. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1532–1543.
33. Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; Joulin, A. Advances in Pre-Training Distributed Word Representations. In Proceedings of the 11th Language Resources and Evaluation Conference, Miyazaki, Japan, 7–12 May 2018.
34. Agerri, R.; Rigau, G. Robust multilingual Named Entity Recognition with shallow semi-supervised features. *Artif. Intell.* **2016**, *238*, 63–82. [CrossRef]
35. Agerri, R.; Rigau, G. Language independent sequence labelling for Opinion Target Extraction. *Artif. Intell.* **2019**, *268*, 85–95. [CrossRef]
36. González Bermúdez, M. An analysis of twitter corpora and the differences between formal and colloquial tweets. In Proceedings of the Tweet Translation Workshop 2015, Alicante, Spain, 5 September 2015; pp. 1–7.
37. Chen, S.F.; Goodman, J. An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.* **1999**, *13*, 359–394. [CrossRef]
38. Turian, J.; Ratinov, L.A.; Bengio, Y. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; Association for Computational Linguistics: Uppsala, Sweden, 2010; pp. 384–394.
39. Brown, P.F.; Desouza, P.V.; Mercer, R.L.; Pietra, V.J.D.; Lai, J.C. Class-based n-gram models of natural language. *Comput. Linguist.* **1992**, *18*, 467–479.
40. Clark, A. Combining distributional and morphological information for part of speech induction. In Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics, Budapest, Hungary, 12–17 April 2003; Association for Computational Linguistics: Budapest, Hungary, 2003; Volume 1, pp. 59–66.
41. Leturia, I. Evaluating different methods for automatically collecting large general corpora for Basque from the web. In Proceedings of the 24th International Conference onComputational Linguistics COLING, Mumbai, India, 8–15 December 2012; pp. 1553–1570.
42. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
43. Rehurek, R.; Sojka, P. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Citeseer, Valletta, Malta, 22 May 2010.
44. Agerri, R.; Bermudez, J.; Rigau, G. IXA pipeline: Efficient and ready to use multilingual NLP tools. *LREC* **2014**, *2014*, 3823–3828.
45. Binkley, D.; Heinz, D.; Lawrie, D.; Overfelt, J. Understanding LDA in source code analysis. In Proceedings of the 22nd International Conference on Program Comprehension, Hyderabad, India, 31 May–7 June 2014; pp. 26–36.
46. Steyvers, M.; Griffiths, T. *Probabilistic Topic Models in Latent Semantic Analysis: A Road to Meaning*; Landauer, T., Mc Namara, D., Dennis, S., Kintsch, W., Eds.; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 2007.

47. Sievert, C.; Shirley, K. LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, MD, USA, 27 June 2014; pp. 63–70.
48. Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An open source software for exploring and manipulating networks. *ICWSM* **2009**, *8*, 361–362.
49. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
50. Jacomy, M.; Venturini, T.; Heymann, S.; Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE* **2014**, *9*, e98679. [CrossRef] [PubMed]

A.2 Fernandez de Landa and Agerri (2021b)

JOURNAL OF MULTILINGUAL AND MULTICULTURAL DEVELOPMENT
https://doi.org/10.1080/01434632.2021.1962331

Social analysis of young Basque-speaking communities in twitter

Joseba Fernandez de Landa and Rodrigo Agerri

HiTZ Center – Ixa, University of the Basque Country UPV/EHU, Bilbao, Spain

ABSTRACT

In this paper, we take into account both social and linguistic aspects to perform demographic analysis by processing a large amount of tweets in Basque language. The study of demographic characteristics and social relationships are approached by applying machine learning and modern deep-learning Natural Language Processing (NLP) techniques, combining social sciences with automatic text processing. More specifically, our main objective is to combine demographic inference and social analysis in order to detect young Basque Twitter users and to identify the communities that arise from their relationships or shared content. This social and demographic analysis will be entirely based on the automatically collected tweets using NLP to convert unstructured textual information into interpretable knowledge.

ARTICLE HISTORY
Received 20 November 2020
Accepted 8 July 2021

KEYWORDS
Computational social science; cultural analytics; natural language processing; basque language; demographic analysis; social media

Introduction

Basque is a low-resourced language, spoken by 28.4% and understood by 44.8% of the population of the Basque Country (Eusko Jaurlaritza, Gobernua, and Office Public de la Langue Basque 2016). Thanks to its official status, it is a language with the presence in the regional public administration, education system and in some news media. Thus, in EiTB (Euskal Irrati Telebista), the Basque public radio and television broadcaster, it is possible to find radio and television channels in which all content is entirely broadcasted in Basque. Furthermore, there are other independent media such as Berria (newspaper), Argia (weekly magazine) and HamaikaTB (a television channel), in which Basque is the vehicular language. Still, the presence of Basque in traditional television and news media remains quite low, particularly when compared with those available for Spanish.

In this context, the increasingly used social networks such as Twitter are of particular importance for a low-resourced language such as Basque. Thus, it is possible to find a strong and active community of Basque speakers in Twitter which generates, for a low-resourced language, a large amount of textual content written on Basque. Furthermore, as users create both explicit and implicit relations and communities, this data is useful to do social research using methods that may complement those traditionally used in sociology (Baldwin et al. 2015; Nguyen et al. 2016; Rosenthal, Farra, and Nakov 2017). Following this, a promising and relatively new avenue of research in social and demographic analysis combines the study of social structures created in media such as Twitter with the automatic analysis of texts via NLP. For example, previous work has focused on Twitter to study the spread of rumours (Derczynski et al. 2017), the detection of political stance (Mohammad et al. 2016) or hate speech (Basile et al. 2019).

In this paper, we will take into account both social and linguistic aspects in order to perform demographic analysis by processing a large amount of tweets in Basque language. The study of demographic characteristics and social relationships will be approached by applying machine

CONTACT Joseba Fernandez de Landa joseba.fernandezdelanda@ehu.eus; Rodrigo Agerri rodrigo.agerri@ehu.eus
HiTZ Center – Ixa, University of the Basque Country UPV/EHU, Bilbao, Spain

learning and modern deep-learning NLP techniques, combining social sciences with automatic text processing.

More specifically, our main objective is to combine demographic inference and social analysis in order to detect young Basque Twitter users and to identify the communities that arise from their relationships or shared content. By 'Basque Twitter users', we refer to those that write at least 20% of their tweets in Basque. This social and demographic analysis will be entirely based on the automatically collected tweets using NLP to convert unstructured textual information into interpretable knowledge.

Current work substantially improves and extends the preliminary experimental work presented in Fernandez de Landa, Agerri, and Alegria (2019). These improvements have led to a number of contributions. First, and taking as a starting point the Heldugazte-corpus containing 6M tweets in Basque language (Fernandez de Landa, Agerri, and Alegria 2019), we devise a whole new methodology to classify users by life-stage (young/adult). This new method generates a new dataset, *Heldugazte-Age*, containing 80K tweets semi-automatically annotated at young/adult level. Second, we explore the application of modern pre-trained large multilingual and monolingual models (Devlin et al. 2019; Agerri et al. 2020) in order to identify young and adult users. Third, we perform a qualitative analysis comparing human performance vs life-stage classifiers for classifying Basque users into the young or adult categories. Four, we use recently developed deep learning techniques for community detection, achieving better detection and visualisation of the communities, as well as providing information of the relations among them. We believe that the methodology presented in this paper might be of interest for other NLP tasks and other types of social and demographic studies. Finally, we publicly distribute every resource (software and data) to facilitate further research for low-resourced languages such as Basque.[1]

The rest of the paper is structured as follows. In the next section, we describe related work in computational sociolinguistics and NLP. In Section 3, we present our method to build the *Heldugazte-Age* annotated dataset to train classifiers for life stage detection. Section 4 presents systems used to train classifiers for life-stage detection. These classifiers are evaluated in Section 5 and applied to perform social network analysis in Section 6. We finish with some concluding remarks and future work.

Context and related work

Social media offers the opportunity to express beliefs, sentiments or opinions in a variety of formats, including text, image, audio and video. Social media publications express conscious and/or subconscious manifestations of our social, emotional and rational condition.

Previous work in sociolinguistics argues that our writing style can even be a reflection of demographic characteristics (Nguyen et al. 2016). Considering the fact that language is a social phenomenon and thanks to the ever-growing capacity in the NLP field to collect and process large-scale amounts of texts, computational sociolinguistics is becoming increasingly popular. The widespread use of Twitter has in fact benefited such approaches as it is possible now to mine large amounts of texts also for less resourced languages.

Twitter is widely used in NLP for tasks such as mining opinions about specific products or topics (Villena et al. 2013; Rosenthal, Farra, and Nakov 2017), detecting political stance (Mohammad et al. 2016; Derczynski et al. 2017) and hate speech (Basile et al. 2019) or for basic tasks such as POS tagging (Ritter, Clark, and Etzioni 2011), named entity recognition (Baldwin et al. 2015), normalisation (Alegria et al. 2015) and language identification (Zubiaga et al. 2016).

NLP techniques specifically adapted for Twitter have also been used to infer demographic characteristics such as gender, age or location (Cesare, Grant, and Nsoesie 2017; Morgan-Lopez et al. 2017). Moreover, relationships, style shifting and community dynamics can also be inferred from language analysis (Nguyen et al. 2016). Of particular interest to us is the body of work performed with the objective of age or life-stage detection for Twitter users. Previous works usually

generate their own manually annotated datasets, covering languages such as Dutch, English or Spanish (Rao et al. 2010; Al Zamal, Liu, and Ruths 2012; Nguyen et al. 2013; Marquardt et al. 2014; Morgan-Lopez et al. 2017; Zaghouani and Charfi 2018) for a user range between 300 and 3000. The best-performing systems are those that model life-stage classification as a binary (Rao et al. 2010; Al Zamal, Liu, and Ruths 2012) or ternary (Nguyen et al. 2013; Morgan-Lopez et al. 2017) task.

In relation to the study of the social relationships that are generated within the network, closer to us are those studies that have aimed to identify communities of users based on their retweets. Among these, one can find studies about political polarisation (Conover et al. 2011), political affiliation detection (Pennacchiotti and Popescu 2011) or even studies about identifying communities in movements for independence (Zubiaga et al. 2017).

Finally, there are different research works investigating the use of low resourced languages within social networks. An investigation about Welsh (*Cymraeg*) speakers and Twitter, shows that speakers of this language are also active in social media (Jones, Cunliffe, and Honeycutt 2013). Additionally, there is another work that extracts and analyses more than 80k tweets in Irish (*Gaeilge*) to do content, sentiment and network analysis (Mhichil, Lynn, and Rosati 2018). It is also interesting a study combining Welsh, Irish and Frisian (*Frysk*) to investigate the use of hashtags across 3000 different tweets (McMonagle et al. 2019). All these works show the potentiality of Twitter to provide text data even for low-resourced languages, giving the chance to find and study a huge variety of languages and cultures.

Heldugazte-age: a new dataset for life-stage classification

In this section, we propose a new methodology to semi-automatically obtain labelled data to develop life-stage classifiers. The result is a new dataset for to train classifiers for Life-Stage Detection, namely, the *Heldugazte-Age* corpus.

The first step to identify online communities of young Basque speakers is collecting the data. As a starting point we will use the Basque corpus *Heldugazte* (Fernandez de Landa, Agerri, and Alegria 2019), which consists of 6M Basque tweets from 8000 users, collected in May 2018.[2] In this collection, the last 3200 tweets from each user were retrieved (if available), including personal tweets and retweets.

The *Heldugazte* corpus will be used to semi-automatically generate a labelled subset of the corpus, 80K tweets, to train classifiers to detect young/adult users. The obtained classifiers will then be applied to the rest of the *Heldugazte* corpus to obtain a large number of tweets written by young users. This data will be used to detect the communities between young users.

In order to obtain a young/adult classifier, we need some labelled data for training and evaluation. However, labelling users' tweets by life stage is a difficult task, due to two main reasons: (i) users age hardly ever appear in the tweets metadata and, (ii) manually annotating tweets by life stage is far from being trivial. Examples (1–3) illustrate the difficulty of manually labelling individual tweets by life stage and without any additional context.

(1) 'Zarauzko triatloian izena ematea lortu gabe, motibazioa falta' *I have not managed to sign up for the Zarautz triathlon, I am unmotivated*
(2) 'A zer nolako eguraldi kaxkarra ez al du gelditu behar edo' *What a bad weather, shouldn't stop or what.*
(3) '5 mila euro, bideo kamera eta telefono mugikor bat eroan dituzte lapurrek' *5,000 euros, a video camera and a cell phone were taken away by the burglars.*

In order to overcome this problem, previous sociolinguistic work has argued that writing style could be associated to author's life stage, assuming that young people's style is more informal than

that of adults (Rao et al. 2010; Al Zamal, Liu, and Ruths 2012; Nguyen et al. 2013; Morgan-Lopez et al. 2017).

Based on these previous works, Fernandez de Landa, Agerri, and Alegria (2019) trained various classifiers to distinguish between formal and informal writing style in tweets. Every tweet for every user in the *Heldugazte* corpus was automatically tagged, projecting from formal/informal to young/adult classification depending on the concentration of formal/informal tweets in each user's timeline. The problem with this procedure was to objectively define a threshold for the proportion of formal/informal tweets required to classify a user as young or adult. The proposed ad-hoc solution, establishing that if 45% of the tweets were labelled informal then the timeline was to be classified as young (adult otherwise) was far from ideal.

In this paper, we propose a new method to objectively and semi-automatically obtain the labelled data required to train young/adult classifiers. The procedure is illustrated in Figure 1. First, we automatically tagged the 6M tweets in the *Heldugazte* corpus using the formal/informal classifiers developed by Fernandez de Landa, Agerri, and Alegria (2019). Second, we ranked users according to the proportion of informal tweets in their timeline. The top users would contain mostly informal tweets, whereas the users at the bottom of the rank would consist mostly of formal tweets. Third, a manual inspection of 100 timelines (50 young and 50 adult) established that it was feasible to manually annotate users at both ends of the ranking as young/adult. In this step, it was particularly helpful to perform the annotation at user-level because a full timeline provides more contextual information to characterise a specific user. Fourth, we took the 500 users at the top of the ranking to be young users and the 500 at the bottom to be adult. As a result, following this new method we obtain a set of 1000 users (out of the original 8K users) classified as adult/young based on the initial formal/informal manual categorisation of 1000 tweets provided by Fernandez de Landa, Agerri, and Alegria (2019).

The final step consisted of randomly sampling a number of tweets per user. The idea was to vary the number of tweets and topics available per user providing a sample general enough to train robust young/adult classifiers. With this objective in mind, we picked 100 random tweets per user (if a user's timeline did not contain at least 100 tweets then we used the full timeline) assigned to each of them the label attributed to the user (young/adult). As it is shown in Table 1, the final labelled set, *Heldugazte-Age*, contains 80K tweets equally distributed into the young and adult classes. The data was splitted for experimentation into a training (60%), development (20%) and test (20%) set, resulting, for each class, in 24K tweets for training and 8K for development and test, respectively.

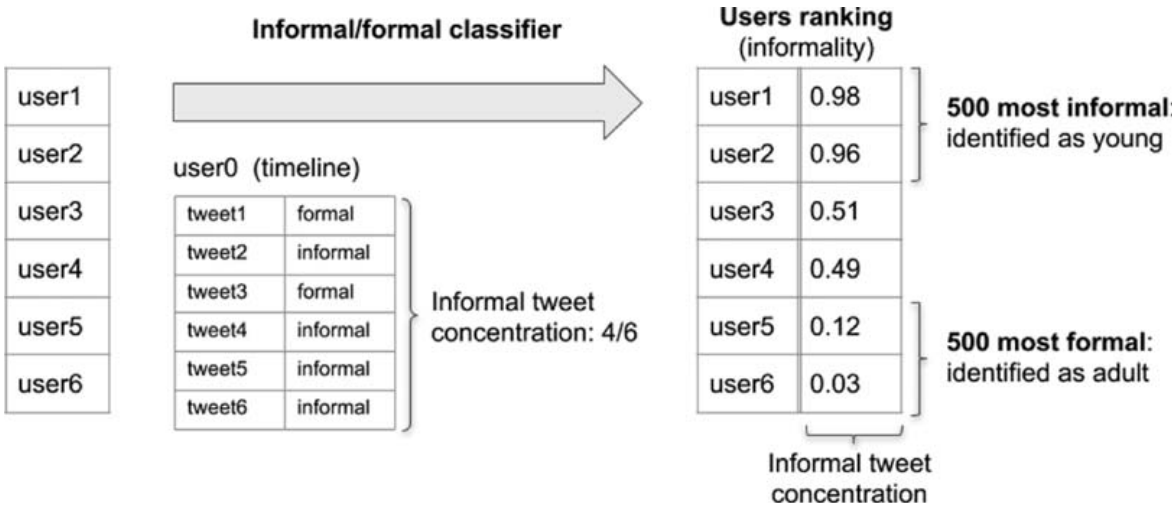


Figure 1. Ranking users by proportion of formal and informal tweets.

Table 1. Annotated corpus for life-stage detection at user level.

	Young	Adult	Total
users	500	500	1000
tweets	40,000	40,000	80,000

Life-stage classification systems

Here we present the two main systems used for life-stage detection: (i) an off-the-shelf system based on linear classification and clustering features (Agerri and Rigau 2016), and (ii) a deep-learning approach based on learning contextual, vector-based word representations and the Transformer architecture (Devlin et al. 2019).

Previous approaches address life-stage detection as a supervised text classification task (Rao et al. 2010; Al Zamal, Liu, and Ruths 2012; Nguyen et al. 2013; Morgan-Lopez et al. 2017). This means that classifiers will learn, from annotated data, that a given tweet is written by a young or an adult person. An example of the dataset annotations used for training can be seen in Table 2. The *Heldugazte-Age* dataset developed in the previous section will therefore be used to train three different text classifiers: (i) IXA pipes (Agerri, Bermudez, and Rigau 2014) (ii) multilingual BERT (Devlin et al. 2019) and (iii) BERTeus (Agerri et al. 2020).

IXA pipes

IXA pipes is a set of tools with a multilingual approach across NLP tasks. This system has been successfully used in several sequence labelling tasks for various languages, including Named Entity Recognition (Agerri and Rigau 2016), and Opinion Target Extraction (Agerri and Rigau 2019).

The general objective of IXA pipes is to provide a general semi-supervised approach that performs well across languages and tasks. This approach consists of two different components. In the first one, a set of linguistically shallow features are extracted from the local context; these features are based on orthographic and ngrams and character-based information to capture multiword patterns and prefixes and suffixes of words, which has proven useful to work with an agglutinative language such as Basque (Agerri and Rigau 2016). The second, semi-supervised, component injects external knowledge previously obtained from the unsupervised induction of clustering models over large amounts of texts. This component provides several benefits. First, it generates denser document representations, given that a document is represented with respect to the number of dimensions (clusters) specified in the obtained clustering model. Second, by training the clustering models on source data from different domains and text genres it is possible to inject domain-specific knowledge into the system. Finally, IXA pipes includes the possibility of including features from three types of clustering models (Brown et al. 1992; Clark 2003; Mikolov et al. 2013), which helps to represent domain-specific information via complementary semantically induced knowledge. More details can be found in Agerri and Rigau (2016) and Agerri and Rigau (2019).

Table 2. Examples taken from the *Heldugazte-Age* dataset.

Label	Content (tweet)
Adult	Taldeak mikel laboaren lanean oinarritu du bere hurrengo diskoa. *The band has based their next album on the work of Mikel Laboa.*
adult	Gure herriko ateak zabalik dituzu. *The doors of our town are opened.*
young	Buaa q follaa eun guztia eon zea ikasi ordez jolasateeenn jajaja. *How lucky! You have been all day playing instead of studying hahaha.*
young	Batzutan ze gutxi aguantatze zaituten. *Sometimes I can't stand you.*

For this particular work, we train the IXA pipes document classifier using the same experimental setup used in Fernandez de Landa, Agerri, and Alegria (2019).

Transformer models

As for many other NLP tasks, current best-performing systems for text classification are large pre-trained language models which allow to build rich representations of text based on contextual word embeddings. Deep learning methods in NLP represent words as continuous vectors on a low-dimensional space, called word embeddings. The first approaches generated static word embeddings (Mikolov et al. 2013; Bojanowski et al. 2017), namely, they provided a unique vector-based representation for a given word independently of the context in which the word occurs. This means that polysemy cannot be represented. Thus, if we consider the word 'bank', static word embedding approaches will generate only one vector representation even though such word may have different senses, namely, 'financial institution','bench', etc.

In order to address this problem, contextual word embeddings were proposed. The idea is to be able to generate word representations according to the context in which the word occurs. Currently, there are many approaches to generate such contextual word representations, but we will focus on those that have had a direct impact in text classification, namely, the models based on the Transformer architecture (Vaswani et al. 2017) and of which BERT is perhaps the most popular example (Devlin et al. 2019).

There are several multilingual versions of these models. Thus, the multilingual version of BERT (Devlin et al. 2019) was trained for 104 languages. More recently, XLM-RoBERTa (Conneau et al. 2019) distributes a multilingual model which contains 100 languages. Both include Basque among the languages.

These multilingual models perform very well in tasks involving high-resourced languages such as English or Spanish, but their performance drops when applied to low-resourced languages (Agerri et al. 2020). Although this is still an open issue, a number of reasons can be found in the literature. First, each language has to share the quota of substrings and parameters with the rest of the languages represented in the pre-trained multilingual model. As the quota of substrings partially depends on corpus size, this means that larger languages such as English or Spanish are better represented than lower resourced languages such as Basque. Moreover, multilingual models also seem to behave better for structurally similar languages (Karthikeyan et al. 2020).

BERTeus (Agerri et al. 2020) is a language model trained in Basque language following BERT's architecture (Devlin et al. 2019). They show that training a monolingual Basque BERT model obtains much better results than the multilingual versions. In this paper, we will compare the performance of multilingual BERT and BERTeus for life-stage detection using the same hyperparameters as in Agerri et al. (2020).

Life-stage detection

In this section, we will use the *Heldugazte-Age* corpus to train the classifiers previously described. The best classifier will then be applied to the whole *Heldugazte* dataset in order to obtain a young/adult classification of the 8K Basque tweet users contained in the corpus. Additionally, an analysis of the results is performed to better understand the quality of the semi-automatically obtained annotations.

Experimental results

It should be noted that, in contrast to our previous work (Fernandez de Landa, Agerri, and Alegria 2019), the *Heldugazte-Age* corpus allows us to directly classify users as young/adult, without having to perform the formal/informal step.

We perform minimal pre-processing on the tweets; we remove URLs, hashtags and usernames, leaving label-tweet pairs such as the examples shown in Table 2. This procedure has proven to be useful in previous text classification works with tweets (Agerri et al. 2020; Zotova, Agerri, and Rigau 2021).

Table 3 reports the results obtained using the three systems described in Section 4. The high scores show that our semi-automatic method to obtain young/adult training data produces good quality annotations. Furthermore, the differences between the systems are not that large, although BERTeus is consistently the best scoring model.

In order to further test the robustness of our semi-automatic method, described in Section 3, we decided to manually annotate 200 randomly selected tweets. Two human annotators labelled the 200 tweets and we calculated an agreement between the annotators of 0.78 and a Kappa score of 0.55, showing a moderate agreement between them. Furthermore, the accuracy of the two annotators are 0.795 and 0.775, respectively. When comparing these scores with the results reported in Table 3, it is clear that manually annotating young/adult at tweet level is a very difficult task. These results also show the effectiveness of our method to obtain the *Heldugazte-Age* corpus.

In the rest of this paper, we will use the BERTeus fine-tuned model to automatically annotate the whole *Heldugazte* corpus. It should be noted that the classifier works at tweet level (as shown by Table 2). This means that once every tweet is automatically annotated, we still need to decide whether each of the user timelines corresponds to a young or adult user based on the number of individual tweets classified as young/adult.

Labeling the large corpus

Once the tweet classifier is ready to use, we apply the following strategy to automatically annotate the tweets in the *Heldugazte* corpus. First we assign a discrete *young* or *adult* label to each tweet. We then obtain a single score by averaging the number of the young/adult classified tweets of each user's timeline.

The last step is to decide whether a given timeline corresponds to a young or adult user based on the score obtained from the classification of the individual tweets. In order to avoid establishing an ad-hoc value as a threshold, we introduce a third class for classification. In other words, a new synthetic category, 'underdetermined', is created thus transforming a binary task into a ternary one.

Based on the new ternary task, two thresholds are used instead of one, located at 60% and 40% of the number of tweets annotated as young in each timeline. Thus, if the proportion of labels or the average probability is over 60%, the user will be defined as a young user. On the other hand, if those values are lower than 40%, the timeline will be considered to be from an adult user. Finally, if the value is between 40% and 60%, we will consider the timeline to be 'underdetermined', meaning that we do not have enough evidence to decide the life stage of the user. Adding the *underdetermined* class has the benefit of avoiding to commit ourselves to classify difficult cases as young/adult.

We are also interested in comparing the distribution of young/adult users obtained using the described procedure with those that are obtained using our previous method (Fernandez de Landa, Agerri, and Alegria 2019). As a reminder, in our previous work, each tweet is classified as formal/informal and then, based on the number of informal tweets, we decide whether the user is young or adult. However, for a fair comparison, we will adapt it to use two thresholds (60/40 for young/adult) and three classes, as it has been described above.

Table 3. Evaluation results of young/adult classifier models on the *Heldugazte-Age* test set.

System	Accuracy	Precision	Recall	F1 Score
IXA pipes	0.956	0.977	0.935	0.955
mBERT	0.955	0.972	0.936	0.954
BERTeus	**0.963**	0.968	0.958	**0.963**

Table 4 shows the number of timelines classified as young/adult or underdetermined using our new and old methods. It can be seen that the main difference corresponds to the quantity of young users obtained by each of the methods. In the next section, we further look into this issue.

Comparison of methods

In this section, we look at those variations in the automatic annotations assigned by the old method (Fernandez de Landa, Agerri, and Alegria 2019) with respect to the one presented in this paper. Table 5 shows the differences of classifying the timelines using the old (based on formal/informal classification of tweets) with respect to the new method (based on young/adult classification). After a superficial look to the variations, it can be seen that 21.79% of the labels were differently labelled from previous to new system, a substantial difference. Besides, one of the most significant variations is the increase in the amount of users labelled as young.

Two of the three most important variations, marked with an asterisk, refer to the transfer of timelines to the young class. It is also important for the transfer from *adult* to *underdetermined*. Taking a deeper look into these specific cases, we manually inspected some randomly chosen timelines to see if these transfers are actually true positives or whether they are misclassifications. The objective of this comparison was to study the transfer of classifications across categories (from adult to young, for example) when using the new classification method. More specifically, we analysed a random sample of 10% of the cases for each variation.

Below we can see example tweets from three different users. With respect to @user2 and @user3, they show that our new method, as opposed to the old one, actually classifies correctly their timelines. Thus, by looking at their tweets, it seems that the users are indeed young, based on the writing style but also because the tweets talk about exams, an activity usually related to young people. The case of @user1 is more contentious, as it seems too difficult to establish the life stage of the user based on those examples, which is why the *underdetermined* classification does not seem misplaced.

- Adult to underdetermined variation example, *@user1*:

  - *tweet_1a*: A zer nolako eguraldi kaxkarra ez al du gelditu behar edo. *What a bad weather, shouldn't stop or what.*

  - *tweet_1b*: Gu erakusteko prest, etorri daitezela lasai eskuzabalik hartuko ditugu eta. *We are ready to show it, we will wait for them with open arms.*
- Adult to young variation example, *@user2*:

  - *tweet_2a*: Horrelakoekin gustua ta guzti hartzen zaio ikasteari. *With this, you take pleasure in learning.*

  - *tweet_2b*: Buenobueno ba ikasiko dut gehio jaja ta ikusikozu gainditukо dutt jaja. *Weeeell weeeell, I'll learn more haha and you'll see if I can pass the exam haha.*
- Underdetermined to young variation example, *@user3*:

  - *tweet_3a*: Ze txupi txatxi no me da la nota. *Awesome I don't get to pass…*

  - *tweet_3b*: Ai naiz rayatzen pixkat asko con la mierda de la uni. *Oh I'm going crazy a little bit with university shit.*

Figure 2 depicts the variations in the classification from the using the old method (left) with respect to the new one (right). The manual inspection performed would indicate that such variations are in fact correct. In other words, it would seem that the method introduced in Sections

Table 4. Classifying users in terms of age stage (young/adult).

	Adult	Underdetermined	Young
Previous system	5213	911	962
New system	4472	980	1635

Table 5. Variations between previous and new systems classifications.

	Previous to New
Adult to adult	4325
Adult to und*	679*
Adult to young*	209*
Und to adult	133
Und to und	285
Und to young*	493*
Young to adult	14
Young to und	16
Young to young	933

4 and 4.2 to develop classifiers to automatically annotate users in the *Heldugazte* corpus as young/adult/underdetermined produce better results.

Relationship network

In this section, we will study the relations that appear between Basque young Twitter users. The starting point will be the retweets of messages written in Basque by the 1635 users classified as *young* in the previous section. We select the retweets because they are the type of interactions between users that can show correlations better than other interactions such as mentions (Conover et al. 2011). Specifically, from the 418,903 retweets of the 1635 young users, we extracted 24,837 nodes and 148,304 edges or connections. The nodes correspond to the users doing the retweets (our sample of 1635 users) but also different users receiving them (from our sample or not). On the other hand, the edges represent if a source user has retweeted one or more times another target user, representing the connections in the graph.

Once the retweets are gathered we proceed to transform the unstructured data into a readable graph. First, we created a giant graph using the data (retweets) from each user. To build the graph, two features extracted from each retweet were used: (i) the retweeter and (ii) the user retweeted. After extracting the data, the visualisation of the graph was created using the *gephi* program (Bastian, Heymann, and Jacomy 2009). Second, we gave the network a spatial structure by using the *ForceAtlas2* algorithm (Jacomy et al. 2014), ordering the nodes according to the established relations. This algorithm displays a spatialisation process, giving a readable shape to a network with the aim of transforming the network into a map. This technique simulates a physical system in order to spatialise a network. As a result of this process, those nodes that are unrelated repulse each other, while related ones will attract each other. The algorithm can turn structural proximities into visual proximities, allowing the analysis of this particular type of data based on interactions. Thus, the relations can be displayed in a (huge) graph.

Figure 2. Previous system to new system.

After creating the graph, we focused on two different aspects. First we identified the most important nodes of the network, to establish which users are the most influential. In a second step, we uncovered the implicit communities of Basque users, splitting the huge graph into more readable subgroups that allowed us to infer the communities of young people.

Basque influencers among young users

The most retweeted users of the graph can reveal important characteristics of the investigated sample. The most important nodes show which users are the leaders for our sample. Thus, in Table 6, we can see the top 15 most retweeted users, based on two different classifications. On the one hand, there are those users with most overall retweets (Table 6(a)). On the other hand, we have the users that have been retweeted by different young users, focusing on how many different users have retweeted these users (Table 6(b)). These two rankings illustrate which users are actually the most influential between young Basque Twitter users.

By looking at the obtained rankings, we can see that at the top there are accounts related to Basque media: @berria (newspaper), @argia (weekly magazine), @naiz_info (newspaper), @topatu_eus (digital media related to young people), @HamaikaTb (a television channel), @euskaltelebista (Basque public television broadcaster) and @LeakoHitza (local newspaper); and Basque journalists: @larbelaitz, @axierL, @boligorria (the three of them journalists from Argia) and @iBROKI (sports journalist in the Basque Television). We attribute this to the fact that those people perform important roles in the creation and distribution of Basque language content in the Web.

Table 6. Most retweeted accounts by young users.

(a) Total RTs done to users.

User	Times retweeted
@berria	8671
@argia	5646
@ernaigazte	4553
@topatu_eus	4236
@enekogara	3274
@naiz_info	3262
@ZuriHidalgo	2568
@AskeGunea	2561
@RealSociedadEUS	2531
@larbelaitz	2471
@ArnaldoOtegi	2188
@iBROKI	2031
@LeakoHitza	1893
@athletic_eus	1818
@euskaltelebista	1744

(b) Young users retweeted accounts.

User	Users retweeted
@berria	998
@argia	844
@naiz_info	710
@larbelaitz	585
@topatu_eus	531
@ArnaldoOtegi	518
@ernaigazte	478
@enekogara	454
@HamaikaTb	442
@jpermach	427
@axierL	413
@ielortza	407
@MaddalenIriarte	404
@boligorria	398
@GureEskuDago	394

After a manual analysis of the obtained influencers, it has to be said that only two of them are accounts related to young people: @ernaigazte and @topatu_eus are both accounts related to organisations formed by young people. On the one hand, @ernaigazte is the account of the Basque nationalist left youth organisation, named Ernai. On the other hand, @topatu_eus is a digital media account related to young people, very related to the Basque nationalist left. The lack of influencers among young users, could be related to the characteristics of Twitter, which is mostly structured around to political issues.

Basque-speaking communities for young users

Once the main network graph was created, we split it into subgroups to analyse how the sub-communities or subgroups inside each one are formed. We divided each graph into subgroups using the node2vec algorithm (Grover and Leskovec 2016), which allows us to obtain consistent subgroups. The node2vec algorithm can freely explore network neighbourhoods which is useful to discover homophilic communities. Unlike modularity-based algorithms (Blondel et al. 2008), used in a previous analysis of Basque communities (Fernandez de Landa, Agerri, and Alegria 2019), node2vec gives the opportunity to choose the exact number of communities to be extracted. Besides, this algorithm can be tuned in order to give more importance to homophilia or to structural equivalence. Thus, Figure 3 shows that node2vec generates clearly distinguishable sub-communities, which in turn makes them more interpretable thereby facilitating the understanding of the existing relations between them.

After splitting the graph into four communities, we had to infer the main characteristics of each subgroup. For this process, we focus again on the most important nodes which are the ones used to define the community itself.

Each of the subgroups displays a common characteristic, namely, all of them have a direct relation with topics or issues related to the Basque Country. Those topics are different in each of the subgroups in the graph, showing the characteristics or differences of each community. Thus, it can be seen that Basque language interactions are used to talk about various Basque current affairs (news) and politics (Nationalist left). Also, it can be seen that music and sports are also widely commented by young people. In other words, it seems that the main function of Twitter

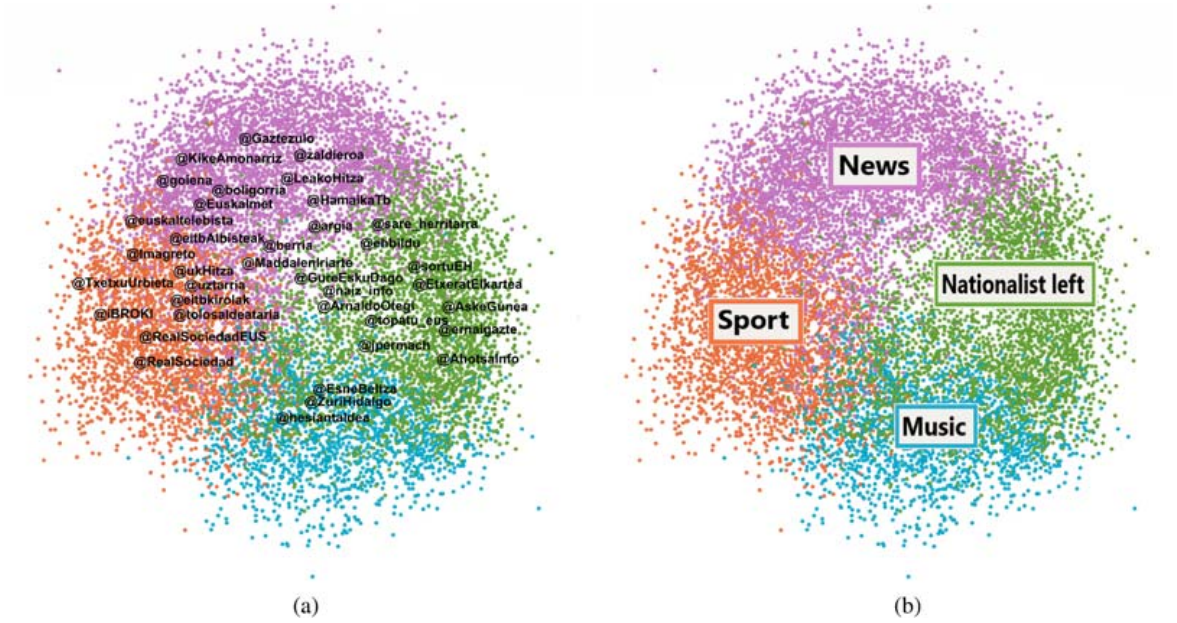


Figure 3. Young users graph divided in communities. (a) Young users communities, some important nodes and (b) Young users communities, topic.

interaction is to spread content about politics and social issues but with a clear focus on the Basque community and language. In the following, we describe the main characteristics of each of the four subgroups contained in the graph.

- **News** (29.96%): In this community, the nodes found at the top of the ranking are related to news media from the Basque Country (@berria, @argia, @HamaikaTb, @eitbAlbisteak, @euskaltelebista, @zuzeu, @euskadi_irratia, @Gaztezulo, @Sustatu, @eitbeus…), specific Basque journalists (@MaddalenIriarte, @boligorria, @urtziurkizu, @zaldieroa, @bzarrabeitia, @AneIrazabal…) and also to other very active users that write about the most noteworthy news in Basque (@ielortza, @kalaportu, @KikeAmonarriz, @maia_jon…).
- **Nationalist left** (26.98%): The composition of this particular subgroup is characterised by nodes related, in different ways, with the nationalist/independentist Basque left. The nodes can refer to news media (@naiz_info, @topatu_eus, @info7irratia, @AhotsaInfo…), political and social organisations (@ernaigazte, @GureEskuDago, @AskeGunea, @ehbildu, @sortuEH, @EtxeratElkartea…) and politicians from the main parties in this political movement (@ArnaldoOtegi, @jpermach…).
- **Sports** (22.58%): In the Sports subgroup, the most important nodes are actually journalists (@iBROKI, @XabierEuzkitze, @Imagreto, @TxetxuUrbieta, @jontolest, @unaizubeldia…) and news media (@eitbkirolak, @ukHitza, @3ErregeenMahaia…) specifically specialised in the sports domain. Thus, for this specific group, the top accounts also refer to newspapers and television broadcasters. Other important nodes here are those related to sport teams, such as football teams (@RealSociedad, @RealSociedadEUS, @SDEibar, @AthleticClub…) and Basque ball clubs (@ASPEpelota…) or their players, which are mostly footballers from professional teams (@InigoMartinez, @mikelsanjo6, @ilarra4…) or even well known cyclists (@AmetsTxurruka, @mikelastarloza, @Markelirizar…).
- **Music** (20.49%): In the Music subgroup appear in prominent places music bands or singers which sing in Basque (@ZuriHidalgo, @vendettaska, @hesiantaldea, EsneBeltza, @gatibu, @ZeEsatek…), although other accounts related to music seem to be also very active (@GustokoMusika, @euskalkantak5, @KantuBatGara…).

Figures 3(a,b) show that young Basque users generally interact with users related to social issues (politics and news) as well as with those related to leisure (music and sport). Due to the new method applied for community detection, we are able to map the communities in a more consistent way, showing in a clear way where each community is located. The position of each community on the graph and the closeness between communities show how related the topics are between them. In this way, we can see that communities related to social issues are next to each other, while the same occurs with the leisure-related communities. The community related to politics is close to news and music, illustrating both the relation with current news and the political stance of some Basque music bands. Besides, in three of the four communities (News, Nationalist left and Sports), media and journalists are referential, proving again that media is important at disseminating Basque content among young people, in spite of the main topic of the community.

In this section, we show that combining the community detection algorithm and the visualisation of the spacial representation of the graph, humans can easily interpret the meaning and characteristics of the displayed data. Thus, any information based on user interactions could be displayed and interpreted using these techniques, helping us to transform unstructured information into knowledge.

Concluding remarks

In this paper, we have presented a new methodology to perform demographic analysis by processing a large amount of tweets in Basque language. We have applied machine learning and deep-learning approaches to NLP to extract structured knowledge from unstructured data.

Our experimental results have shown that our new method produces good quality labelled data for training young/adult classifiers. This allows us to generate a new dataset of 80K tweets annotated at the user level, namely, *Heldugazte-Age*. The analysis of the classifiers performance has shown that, when compared with manual annotations at tweet level, the annotations of our semi-automatically generated *Heldugazte-Age* dataset benefit from taking into account user-level information. Furthermore, we have experimented with modern deep-learning techniques for NLP and for the detection and visualisation of communities in Twitter. The use of these technologies has allowed us to get more consistent and readable results than in our previous approach (Fernandez de Landa, Agerri, and Alegria 2019), apart from a better understanding of communities and their interactions.

As a result of our new methodology, we have seen that the young Basque users can be grouped in four main communities. Furthermore, we have also seen that the most influential accounts among young users are related with Basque media, revealing the importance of this actor at disseminating content in Basque among the youngest. A general conclusion has been that Basque is mostly used in Twitter to speak about Basque-related topics, being that news, politics, sport or music.

We believe that the methodology presented in this paper might be of interest for other NLP tasks and other types of social and demographic studies. Finally, we publicly distribute every resource (software and data) to facilitate further research for low-resourced languages such as Basque.[3]

Notes

1. https://github.com/ixa-ehu/heldugazte-corpus
2. http://ixa2.si.ehu.es/heldugazte-corpus/heldugazte-osoa.tar.gz
3. https://github.com/ixa-ehu/heldugazte-corpus

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work has been partially funded by the Spanish Ministry of Science and Innovation (DeepReading RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE), Ayudas Fundación BBVA a Equipos de Investigación Científica 2018 (BigKnowledge), and DeepText (KK-2020/00088), funded by the Basque Government. Rodrigo Agerri is also funded by the RYC-2017-23647 fellowship and acknowledges the donation of a Titan V GPU by the NVIDIA Corporation.

ORCID

Joseba Fernandez de Landa http://orcid.org/0000-0001-6067-3571
Rodrigo Agerri http://orcid.org/0000-0002-7303-7598

References

Agerri, Rodrigo, Josu Bermudez, and German Rigau. 2014. "IXA Pipeline: Efficient and Ready to Use Multilingual NLP tools." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Vol. 2014, 3823–3828.

Agerri, Rodrigo, and German Rigau. 2016. "Robust Multilingual Named Entity Recognition with Shallow Semi-supervised Features." *Artificial Intelligence* 238 (2): 63–82.

Agerri, Rodrigo, and German Rigau. 2019. "Language Independent Sequence Labelling for Opinion Target Extraction." *Artificial Intelligence* 268: 85–95.

Agerri, Rodrigo, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. "Give Your Text Representation Models some Love: The Case for Basque." In *Proceedings of The 12th Language Resources and Evaluation Conference*, 4781–4788.

Alegria, Iñaki, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2015. "TweetNorm: a Benchmark for Lexical Normalization of Spanish Tweets." *Language Resources and Evaluation* 49 (4): 883–905.

Al Zamal, Faiyaz, Wendy Liu, and Derek Ruths. 2012. "Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors." In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 270, 2012.

Baldwin, Timothy, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. "Shared Tasks of the 2015 Workshop on Noisy User-Generated Text: Twitter Lexical Normalization and Named Entity Recognition." In *Proceedings of the Workshop on Noisy User-generated Text*, 126–135.

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter." In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, 54–63.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks." *Proceedings of the International AAAI Conference on Web and Social Media* 8 (2009): 361–362.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics* 5: 135–146.

Brown, Peter F., Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. "Class-based N-gram Models of Natural Language." *Computational Linguistics* 18 (4): 467–479.

Cesare, Nina, Christan Grant, and Elaine O. Nsoesie. 2017. "Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices." arXiv preprint arXiv:1702.01807.

Clark, Alexander. 2003. "Combining Distributional and Morphological Information for Part of Speech Induction." In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Unsupervised Cross-Lingual Representation Learning at Scale." arXiv:1911.02116.

Conover, Michael D., Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. "Political Polarization on Twitter." In *Proceedings of the International AAAI Conference on Web and Social Media*.

Derczynski, Leon, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. "SemEval-2017 Task 8: RumourEval: Determining Rumour Veracity and Support for Rumours." In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 69–76.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186.

Eusko Jaurlaritza, E. J. G. V., and Nafarroako Gobernua, and Office Public de la Langue Basque. 2016. VI. Inkesta Soziolinguistikoa. *irekia.euskadi.eus*.

Fernandez de Landa, Joseba, Rodrigo Agerri, and Iñaki Alegria. 2019. "Large Scale Linguistic Processing of Tweets to Understand Social Interactions Among Speakers of Less Resourced Languages: The Basque Case." *Information* 10 (6): 212–00.

Grover, Aditya, and Jure Leskovec. 2016. "Node2vec: Scalable Feature Learning for Networks." In *Association for Computing Machinery*, 855–864.

Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software." *PloS One* 9 (6): e98679.

Jones, Rhys James, Daniel Cunliffe, and Zoe R. Honeycutt. 2013. "Twitter and the Welsh Language." *Journal of Multilingual and Multicultural Development* 34 (7): 653–671.

Karthikeyan, K., Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. "Cross-Lingual Ability of Multilingual BERT: An Empirical Study." In *International Conference on Learning Representations*.

Marquardt, James, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. 2014. "Age and Gender Identification in Social Media." In *Proceedings of CLEF 2014 Evaluation Labs*, 1129–1136.

McMonagle, Sarah, Daniel Cunliffe, Lysbeth Jongbloed-Faber, and Paul Jarvis. 2019. "What Can Hashtags Tell Us About Minority Languages on Twitter? A Comparison of #cymraeg, #frysk, and #gaeilge." *Journal of Multilingual and Multicultural Development* 40 (1): 32–49.

Mhichíl, Mairéad Nic Giolla, Theo Lynn, and Pierangelo Rosati. 2018. "Twitter and the Irish Language, #Gaeilge – Agents and Activities: Exploring a Data Set with Micro-implementers in Social Media." *Journal of Multilingual and Multicultural Development* 39 (10): 868–881.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." In *Advances in Neural Information Processing Systems*, 3111–3119.

Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016, June. "SemEval-2016 Task 6: Detecting Stance in Tweets." In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41, San Diego, CA: Association for Computational Linguistics.

Morgan-Lopez, Antonio A., Annice E. Kim, Robert F. Chew, and Paul Ruddle. 2017. "Predicting Age Groups of Twitter Users Based on Language and Metadata Features." *PloS One* 12 (8): e0183537.

Nguyen, Dong, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "'How Old Do You Think I Am?' A Study of Language and Age in Twitter." In *Proceedings of the International AAAI Conference on Web and Social Media*.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. "Computational Sociolinguistics: A Survey." *Computational Linguistics* 42 (3): 537–593.

Pennacchiotti, Marco, and Ana-Maria Popescu. 2011. "Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter." In *Association for Computing Machinery*, 430–438. ACM.

Rao, Delip, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. "Classifying Latent User Attributes in Twitter." In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, 37–44. ACM.

Ritter, A., S. Clark, and O. Etzioni. 2011. "Named Entity Recognition in Tweets: An Experimental Study." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1524–1534.

Rosenthal, Sara, Noura Farra, and Preslav Nakov. 2017. "SemEval-2017 Task 4: Sentiment Analysis in Twitter." In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 502–518.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you Need." In *Advances in Neural Information Processing Systems*, 5998–6008.

Villena, Julio, Sara Lana, Eugenio Martínez, and José Carlos González. 2013. "TASS-Workshop on Sentiment Analysis at SEPLN." *Sociedad Española para el Procesamiento del Lenguaje Natural*.

Zaghouani, Wajdi, and Anis Charfi. 2018. "Arap-Tweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Zotova, Elena, Rodrigo Agerri, and German Rigau. 2021. "Semi-Automatic Generation of Multilingual Datasets for Stance Detection in Twitter." *Expert Systems with Applications* 170: 114547.

Zubiaga, Arkaitz, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2016. "TweetLID: A Benchmark for Tweet Language Identification." *Language Resources and Evaluation* 50 (4): 729–766.

Zubiaga, Arkaitz, Bo Wang, Maria Liakata, and Rob Procter. 2017. "Stance Classification of Social Media Users in Independence Movements." *Catalonia* 2 (8, 599): 10–960.

A.3 Agerri *et al.* (2021)

Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 173-181 recibido 30-06-2021 revisado 07-07-2021 aceptado 11-07-2021

VaxxStance@IberLEF 2021: Overview of the Task on Going Beyond Text in Cross-Lingual Stance Detection

VaxxStance@IberLEF 2021: Descripción de la tarea de detección de actitudes basada en el uso de información más allá del texto

Rodrigo Agerri[1], Roberto Centeno[2], María Espinosa[2], Joseba Fernandez de Landa[1], Álvaro Rodrigo[2]
[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[2]NLP&IR group at Universidad Nacional de Educación a Distancia (UNED)
rodrigo.agerri@ehu.eus, rcenteno@lsi.uned.es, mespinosa@lsi.uned.es,
joseba.fernandezdelanda@ehu.eus, alvarory@lsi.uned.es

Abstract: This paper describes the VaxxStance task at IberLEF 2021. The task proposes to detect stance in Tweets referring to vaccines, a relevant and controversial topic in the current pandemia. The task is proposed in a multilingual setting, providing data for Basque and Spanish languages. The objective is to explore crosslingual approaches which also complement textual information with contextual features obtained from the social network. The results demonstrate that contextual information is crucial to obtain competitive results, especially across languages.
Keywords: Stance Detection, Multilingualism, Computational Social Science, Information Extraction.

Resumen: En este artículo se describe la tarea VaxxStance celebrada en el marco de IberLEF 2021. La tarea propone detectar la actitud de un conjunto de tweets relativos a las vacunas, a un tema muy actual y polémico en estos tiempos de pandemia. La tarea se ha propuesto en un marco multilingüe, euskera y español. Además del texto de cada tweet, se ha proporcionado además información relacionada con la red social de los usuarios autores de los tweets. Los resultados de los participantes han corroborado que el uso de información de la red social permite mejorar el rendimiento en esta tarea, particularmente en un entorno crosslingüe.
Palabras clave: Detección de Actitudes, Multilingüismo, Ciencias Sociales Computacionales, Extracción de Información.

1 Introduction

Stance detection is one of the tasks within the universe of Fake News detection and as such is related to other tasks such as Hyperpartisanism (Kiesel et al., 2019), Hate Speech Detection (Basile et al., 2019), Fact-checking and Claim Verification (Thorne et al., 2018), among others. The most popular formulations are perhaps those proposed in 2016 by the SemEval-2016 Task 6: Detecting Stance in Tweets (Mohammad et al., 2016) and by the Fake News Challenge (Stage 1)[1]. In the first, stance is defined as establishing whether a given tweet expresses a FAVOR, AGAINST or NEUTRAL (NONE) attitude with respect to a given, pre-defined topic. In the second formulation, provided by the Fake News Challenge, stance has to be inferred between a claim and a text commenting on the claim. In this case the stance category can be one of Agrees, Disagrees, Discusses and Unrelated.

Other subsequent contributions have focused mostly on the static variant of stance detection (with respect to a pre-defined topic) rather than on the dynamic one (classifying stance with respect to previous message), although there are some exceptions, notably the RumourEval tasks (Derczynski et al., 2017; Gorrell et al., 2019).

Furthermore, as it is usually the case in the Natural Language Processing (NLP) field, most works have experimented on En-

[1] `http://www.fakenewschallenge.org/`

ISSN 1135-5948. DOI 10.26342/2021-67-15

glish only, with some exceptions. An Arabic corpus integrated the tasks of fact-checking and stance detection (Baly et al., 2018), a dataset from comments of news was developed for Czech language (Hercig et al., 2017), and there also works for French (Evrard et al., 2020) and Russian (Vychegzhanin, 2019). Finally, an interesting new dataset for Italian was released in 2020 as part of the SardiStance@Evalita 2020 shared task (Cignarella et al., 2020), which included not only the texts of the tweets labeled with stance, but also social network information relative to the authors of the tweets. This social network information includes retweets, user accounts profile, friends and followers, among others.

Other interesting works have tried to address stance detection from a multilingual point of view. The IberEval 2017 and 2018 shared tasks (Taulé et al., 2018) provided a dataset in Catalan and Spanish to classify stance with respect to the Independence of Catalonia, while Lai et al. (2020) provided datasets in French and Italian. However, these multilingual efforts are hindered by the extremely skewed class distribution in the Catalan IberEval data, or by the fact that the data for each language was not collected on the same timeframe and addressed different topics. This makes it very difficult to investigate multilingual and crosslingual approaches to stance detection. While Zotova, Agerri, and Rigau (2021) propose a method to address these shortcomings by providing a semi-automatically generated multilingual stance detection corpus, they do not include social network features in their dataset.

In this context, we propose the VaxxStance shared task at IberLEF 2021 (Montes et al., 2021), with the aim of detecting stance in social media on vaccines in general. The task provides data in two languages, Basque and Spanish, and its objective is to promote crosslingual research on stance detection using both the text and the information provided by the Twitter social network. Thus, and unlike previous approaches, we provide, for a given topic, multilingual coetanous data of gold-standard quality in a corpus which allows to experiment using both social and textual features in multilingual and crosslingual settings.

2 Multilingual Dataset

Following the formulation of stance provided by Mohammad et al. (2016), the VaxxStance task consists of determining whether a given tweet expresses an AGAINST, FAVOR or NEUTRAL (NONE) stance towards vaccines. Additionally, and inspired by the SardiStance 2020 shared task (Cignarella et al., 2020), the dataset includes two different types of data: Textual and Contextual (retweets, friends and user data), for two language, Basque and Spanish. The dataset is publicly available in the task website[2].

2.1 Collection and Annotation

In a first attempt we tried to do the data collection and annotation for both languages in the same manner. However, as it will be explained below, due to the idiosyncrasies of Basque it was necessary to devise an alternative, more viable, method for that language, especially to obtain the required textual data.

In any case, we did specify a number of criteria that both languages needed to comply with. First, the datasets a required to have a balanced distribution in the ratio users/tweets to avoid that a large number of tweets belonged to a very few users. Second, the tweets in the training set had to be written by different users from those contained in the test set. This is to avoid obtaining artificially high results due to the existence of user-based information in both the training and test sets. As such, the general idea is that both the textual and user-based (or contextual) knowledge would help each other in order to better classify stance. Finally, we use the annotation guidelines from the SemEval 2016 task (Mohammad et al., 2016).

2.1.1 Basque

Basque is spoken by roughly the 30% of the population in the Basque Country, and understood by around 50%. Due to the fact that Basque is a co-oficial language, it does have presence in the regional public administration, as well as in the education system and some news media, including a public television broadcaster. Still, the presence of Basque in mass media is extremely low, especially when compared to Spanish, the 4th most spoken language in the world.

In this context, the increasing popularity of Twitter among Basque speakers is of

[2] https://vaxxstance.github.io/

particular importance for a low resource language, as a relatively large amount of textual content written in Basque is generated in that social network. This provides a valuable resource to study new NLP tasks such as stance detection not only for large and popular languages, but also for low resourced ones. Still, the collection process of enough tweets relevant to the VaxxStance task was rather challenging.

At first we experimented with a keyword extraction method using the following specific keywords: *"txertoa"* (vaccine) and *"txertaketa"* (vaccination), *"negazionista"* (negationist), *#pfizer*, *#moderna*, *#astrazeneca* and their respective inflections. However, it was surprising to find that the traffic of Basque tweets relative to those topics were relatively low.

We therefore decided to try an alternative, more brute-force, method. First, we collected all the available timelines of users that are identified to write mostly in Basque (around 10k users). The content of these timelines amount to around 8M tweets. Second, relevant tweets were selected following a simple keyword search using the same keywords listed for the previous attempt. Third, a first annotator manually labeled a set of around $1,400$ tweets. Finally, those same $1,400$ tweets, belonging to 210 users, were blindly annotated by a second annotator. The final composition of the textual part of the dataset can be seen in Table 1.

		Train	Test
Tweets		1,072	312
	Favor	327	85
	Neutral	524	135
	Against	219	92
Users		149	61

Table 1: Textual data in the Basque dataset.

We would like to note that the most difficult part in the process was finding enough users that explicitly expressed a stance AGAINST vaccines.

2.1.2 Spanish

Around $2,700$ tweets written in Spanish stating an opinion about "vaccines" were collected and annotated, as shown by Table 2. In order to avoid a potential bias derived from the current COVID-19 pandemic situation, the tweets were collected from the beginnings of Twitter until current time. They were also restricted to the peninsular variant of the Spanish language in order to avoid problems derived from the use of different terms in other variants such as Colombian, Peruvian, etc. To guide this process we used the Google tool *"Google Trends"*[3] which allowed us to locate temporal spaces where events related to vaccines had occurred, identifying the type of event and the date on which it happened. Some examples are the peaks in traffic for and against the vaccination against measles, which was a consequence of some measles outbreaks that happened in Spain during 2019. By using keywords related to the event and restricting the dates obtained, we managed to introduce tweets related to events other than the COVID-19 vaccination process.

		Train	Test
Tweets		2,003	694
	Favor	937	359
	Neutral	591	195
	Against	475	140
Users		1,261	414

Table 2: Textual data in the Spanish dataset.

In addition to the tweets collected through the events identified in Google Trends, for the rest of the tweets collected we followed the following process. First, we used a set of keywords such as "vaccine", "vaccination", as well as terms related to diseases whose vaccines have generated some controversy in society and in anti-vaccine movements, e.g., "chickenpox", "autism", "MMR", etc. After a first manual analysis, we observed that the vast majority of the tweets collected did not express a stance. In order to solve this problem, we then extracted the hashtags most commonly used in these tweets and manually analysed those that were used to express a position in favour and/or against vaccines. Some examples of these hashtags are *#YoMeVacuno*, *#VaccinesWork*, *#COVID19*, *#vacuna*, *#yomevacuno*, *#VacunaCOVID19*, *#YoNoMeVacuno*, *#gripe*, *#Plandemia*, *#yosimevacuno*, etc.

By using these hashtags, we managed to increase the number of tweets to start with the manual labeling. The labelling was performed manually by two annotators, using

[3] `https://trends.google.es/trends/?geo=ES`

a third annotator to resolve disagreements. For this we used the web platform created by Cignarella et al. (2020), to whom we would like to thank for their help using it.

Once the manual annotation was completed, the set of AGAINST tweets was much smaller than those expressing a FAVOR or NEUTRAL stance. To address this issue, we identified several accounts of users that may potentially be identified as supporters of anti-vaccine movements and manually collected tweets from these users expressing an AGAINST stance. This step was performed taking care in complying with the general criteria of not including more than 10 tweets per user in the final corpus, as well as not overlapping users between the training and evaluation set. In this final process we managed to increase by about 200-250 tweets the AGAINST class.

2.2 Social Media Information

The main objective of this task is studying the usefulness of the context provided by social media information to classify stance in a crosslingual setting. With this objective in mind, we collected contextual information relative to the *friends* of the authors of the tweets as well as their *retweets*. The context provided by *friends* and *retweets* can be leveraged to generate relation graphs that in turn may be used to improve the classifiers.

Table 3 shows the social media data gathered with respect to the tweets in the train and test partitions for each of the languages. In addition to the retweets of the tweets included in the datasets, for Basque we also decided to collect the all the retweets made by the users, namely, by extracting the retweets from the users' timelines (TL). This strategy was applied in order to alleviate the small number of retweets obtained from the tweets in the train and test partitions.

		Train	Test
Basque	Friends	119,977	53,029
	Retweets	203	0
	Retweets (TL)	130,369	61,438
Spanish	Friends	1,708,396	438,586
	Retweets	6,832	2,148

Table 3: Social Media Information by language.

Finally, apart from social media information, the dataset also includes the meta information of each annotated tweet as well as the information related to each user.

2.3 Final Dataset

Table 4 shows the composition of the VaxxStance dataset, including both textual and contextual information. Regarding the textual information, it can bee seen that the Spanish set is roughly double in size with respect to the Basque one, although the distribution of classes across the train and test set, as shown by Tables 1 and 2, is quite similar.

		Train	Test
Basque	Tweets	1,072	312
	Users	149	61
	Friends	119,977	53,029
	Retweets	203	0
	Retweets (TL)	130,369	61,438
Spanish	Tweets	2,003	694
	Users	1,261	414
	Friends	1,708,396	438,586
	Retweets	6,832	2,148

Table 4: Composition of the VaxxStance 2021 dataset.

With respect to the contextual information we can see that for Basque there are very few users, around 10% of the number of users for Spanish. This is a reflection of the much smaller community of Twitter users that write in Basque. In this sense, the *friends* graph also reflects the same ratio, as the number of *friends* relations is around 10% of those obtained for Spanish. If we look at the *retweets*, however, we can see that for Basque we only managed to obtain very few of them. That is why we decided to also provide the retweets for each user in the train and test sets (*retweets TL*).

In summary, the VaxxStance dataset provides an interesting benchmark to investigate crosslingual approaches to stance detection based on both textual and contextual features. While the Basque set is slightly smaller than some previous approaches (Taulé et al., 2018; Cignarella et al., 2020; Zotova, Agerri, and Rigau, 2021) it is still larger than the data provided for any of the topics in the SemEval 2016 dataset, which is perhaps the most popular benchmark for stance detection (Mohammad et al., 2016).

3 Task Definition

In this task we aimed to promote research on multilingual and crosslingual approaches to stance detection in Twitter. Ideally, this type of research requires annotated datasets on a common topic for more than one language and obtained on the same dates (coetaneous data). However, while previous work mentioned in the Introduction includes datasets in several languages, they do not provide an adequate evaluation setting for multilingual and crosslingual studies to stance detection.

3.1 Tracks

As the task contains tweets in two different languages, we proposed the following participation tracks for each language (Basque and Spanish):

- Close Track: Language-specific evaluation. Only the provided data for each of the languages is allowed. There are two evaluation settings:
  - Textual: Only the provided tweets in the target language can be used for development. No data augmentation is allowed.
  - Contextual: Text plus given Twitter-related information will be used by the participants. Contextual information refers to features related with user-based Twitter information: friends, retweets, etc. described in Section 2.2.
- Open Track: Participants can use any kind of data, including additional tweets obtained by the participants. The main objective consists of exploring data augmentation and knowledge transfer techniques for cross-lingual stance detection.
- Zero-shot Track: Texts (tweets) of the target language cannot be used for training. The main objective is to explore how to develop systems that do not have access to text in the target language, especially using Twitter-related information.

Participants could submit their systems to any of the tracks, but it was compulsory to participate in both languages for the chosen track.

3.2 Evaluation

Following previous work, we evaluate the systems with the metric provided by the SemEval 2016 task on Stance Detection (Mohammad et al., 2016) which reports F1 macro-average score of two classes, FAVOR and AGAINST, although the NONE class is also represented in the test data:

$$F1_{avg} = \frac{F1_{favor} + F1_{against}}{2} \quad (1)$$

The official evaluation script is distributed together with the dataset in the task website[4].

3.3 Baselines

We provide two baselines, one using only textual information and a second one using just social or contextual features:

- Textual: The textual baseline is based on a SVM classifier with RBF kernel function. The text of the tweets is vectorized using a TF-IDF vectorizer and then feed to the classifier. Both *C* and *Gamma* hyperparameters are tuned by means of grid search and 5 fold CV on the training data. The best configuration is used to evaluate on the test.
- Social: This classifier uses the metadata related to each user and tweet to obtain a number of features (friends count, status count, emojis in bio, etc.) which are then used to train a XGBoost classifier. Before feeding the classifier, each class data is weighted in order to create a balanced sample.

		Against	Favor	Average
Basque	Textual	51.80	57.01	54.41
	Social	5.23	48.53	26.88
Spanish	Textual	71.38	81.68	76.53
	Social	73.14	73.73	73.43

Table 5: Baseline results on Test set.

The results obtained by the baselines show that both tracks are harder for Basque. With respect to the Textual track, stance in Spanish seems to be expressed more explicitly. Regarding the social baseline, the low results were probably caused by the low number of

[4] https://vaxxstance.github.io/

Basque users from which to obtain the features.

4 Participants and Results

Twenty groups registered for the task and downloaded the datasets. However, only three groups finally submitted runs. Table 6 shows the information of the participant groups and the reference to their reports.

Team	Report
MultiAztertest	(Gonzalez-Dios and Bengoetxea, 2021)
SQYQP	(Calleja and Méndez, 2021)
WordUp	(Lai et al., 2021)

Table 6: Participants.

In total, the participants submitted 28 runs, 14 per language. While all the three teams participated in both Textual and Contextual settings of the Close Track, only one team, WordUp, participated in the Zero-shot and Open Tracks. Thus, any comparisons between the participant systems will be performed on the Close Track.

4.1 Close Track

Table 7 shows the results for the Close Track, which received 20 submission runs. We report results for each language and evaluation setting (Textual and Contextual). For all the four rankings, the best results are always obtained by the WordUp team.

As it was the case with our baseline results, the participant systems score systematically higher for Spanish. The best results for Spanish are over an 80 F1 score. These results seem to confirm that the Spanish set is easier than the Basque one.

For each language, the results improve by using contextual information, except for the MultiAztertest Basque submissions. Still, results confirm the effectiveness of employing both textual and social information.

Regarding the results of baselines, the textual baseline obtains competitive results in both languages, being only outperformed by the WordUp team. In contrast, the contextual baseline's results are improved by all the teams in Basque (except one run from WordUp which obtains a very low score) and at least one run per group, except SQYQP, in Spanish. These results suggest that, despite their simplicity and use of linear classifiers, the approaches followed by both baselines represent an adequate starting point.

Regarding the techniques followed by the participants, MultiAzterTest tested two different approaches for the Textual setting: a system which used pre-trained transformers-based language models (run 1) and another one based on training a classifier with a set of linguistic and stylistic features such as word frequencies, semantic overlap, etc. (run 2). The first approach performed much better than the second in the textual track. For the contextual track, they employ only the information relative to the user. More specifically, for each user they select the most common stance label, assuming that users tweets correspond coherently to one stance type. While this idea worked well in Spanish, it was detrimental in (run 1) in Basque.

The SQYQP team addressed the textual setting by training a LSTM initialized with multilingual BERT embeddings using the Flair toolkit (Akbik, Blythe, and Vollgraf, 2018). For the contextual setting, they added network information and measured the distance among users following the approach done by Espinosa et al. (2020). While their results for Basque are below the textual baseline, they improve results by adding contextual information.

The WordUp! team employed a large number of common features that have been proved useful for stance detection. Features were extracted mostly from stylistic and dependency analyses. Additionally, the also crawled tweets specifically for this task and topic for both languages. The tweets were then used to train FastText word embeddings and used to obtain several features. Moreover, they created a dictionary of lemmas referring to stance in English, and translated it to Spanish and Basque. In the contextual setting they tried several network-based measures to be added as features to the logistic regression classifier. They report a large number of experiments which resulted in the best performing team across all evaluation tracks.

In general, the results obtained by the participants show that, even when using very simple features, contextual information substantially improved the results obtainecd by using text only.

		Against	Favor	F1 Macro
Basque Textual	WordUp_01	57.69	56.99	57.34
	WordUp_02	55.03	54.27	54.65
	*BASELINE	51.80	57.01	54.41
	MultiAztertest_01	48.23	52.25	50.24
	SQYQP_01	38.81	46.31	42.56
	MultiAztertest_02	34.38	34.18	34.28
Spanish Textual	WordUp_02	78.36	83.47	80.92
	WordUp_01	75.54	82.58	79.06
	*BASELINE	71.38	81.68	76.53
	MultiAztertest_01	66.67	81.53	74.10
	SQYQP_01	57.14	77.61	67.38
	MultiAztertest_02	56.47	71.60	64.04
Basque Contextual	WordUp_02	82.95	72.46	77.71
	SQYQP_01	65.17	52.94	59.06
	MultiAztertest_02	25.40	48.03	36.72
	MultiAztertest_01	16.36	56.06	36.21
	*BASELINE	5.23	48.53	26.88
	WordUp_01	0.00	0.08	0.04
Spanish Contextual	WordUp_02	91.17	87.09	89.13
	WordUp_01	88.97	86.56	87.77
	MultiAztertest_01	78.77	79.84	79.31
	*BASELINE	73.14	73.73	73.43
	SQYQP_01	66.27	80.06	73.17
	MultiAztertest_02	63.93	77.17	70.55

Table 7: Close Track official results.

4.2 Open Track

The only participant in this track was the WordUp! team, which performed data augmentation. They generated FastText word embeddings (Bojanowski et al., 2017) from a set of tweets specifically obtained for this particular task and languages. They also augmented the contextual information by extracting the social network of each user. As it can be seen in the results reported in Table 8, their results are quite similar to those obtained in the Close Track - Contextual setting. This might be due to the fact that they also used the ad-hoc generated FastText embeddings also in the Close Track.

		Against	Favor	F1 Macro
Basque	WordUp_02	82.29	72.12	77.21
	WordUp_01	64.47	68.12	66.30
Spanish	WordUp_02	90.87	88.07	89.47
	WordUp_01	90.39	88.01	89.20

Table 8: Open Track official results.

4.3 Zero-shot Track

Table 9 shows the results obtained by the only participant in this track, in which the participants could not use the text (tweets) of the target language for training. The most surprising aspect of the results is perhaps the fact that, for Basque, the zero-shot results outperform the results of the Textual evaluation setting. This seems to indicate that contextual information is far more important than the texts themselves in order to perform stance detection.

		Against	Favor	F1 Macro
Basque	WordUp_01	64.47	68.12	66.30
	WordUp_02	55.70	39.74	47.72
Spanish	WordUp_01	88.03	46.13	67.08
	WordUp_02	18.63	62.77	40.70

Table 9: Zero-Shot Track official results.

5 Concluding Remarks

In this paper we provide an overview of the VaxxStance@IberLEF 2021 shared evaluation task, in which the objetive is to detect stance towards vaccines across two different languages: Basque and Spanish. As a novelty for stance detection in these languages, systems can use textual and contextual infor-

mation to train their systems in multilingual and crosslingual settings.

The techniques employed by the different participants showed that contextual information has a great impact across languages, even for small community of users such as Basque. In this sense, textual results are in general improved by adding social network features.

The datasets for both languages were built following the same criteria and objectives. However, further analysis is required to understand why results are systematically better for Spanish than those obtained for Basque. Finally, given that just one team participated in the Open and Zero-shot Tracks, one of the main objectives of the task, to promote research on crosslingual approaches to stance detection, has not completely been achieved. Therefore further work is required on this particular line of research.

Acknowledgments

This work has been partially supported by the European Social Fund through the Youth Employment Initiative (YEI 2019) and the Spanish Ministry of Science, Innovation and Universities (DeepReading RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE), and by the DeepText project (KK-2020/00088), funded by the Basque Government. Rodrigo Agerri is also funded by the RYC-2017-23647 fellowship.

Finally, we are grateful to Mirko Lai and Alessandra Cignarella for sharing with us their experience organizing the SardiStance 2020 shared task.

References

Akbik, A., D. Blythe, and R. Vollgraf. 2018. Contextual string embeddings for sequence. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA.

Baly, R., M. Mohtarami, J. Glass, L. Màrquez, A. Moschitti, and P. Nakov. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana, June. Association for Computational Linguistics.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Calleja, J. and A. Méndez. 2021. Sqyqp@vaxxstance: Stance detection for the antivaxxers movement. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

Cignarella, A. T., M. Lai, C. Bosco, V. Patti, and P. Rosso. 2020. SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In V. Basile, D. Croce, M. Di Maro, and L. C. Passaro, editors, *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Derczynski, L., K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, and A. Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada, August. Association for Computational Linguistics.

Espinosa, M. S., R. Agerri, Á. Rodrigo, and R. Centeno. 2020. Deepreading @ sardistance 2020: Combining textual, social and emotional features. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765

of *CEUR Workshop Proceedings*. CEUR-WS.org.

Evrard, M., R. Uro, N. Hervé, and B. Mazoyer. 2020. French Tweet Corpus for Automatic Stance Detection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6317–6322, Marseille, France, May. European Language Resources Association.

Gonzalez-Dios, I. and K. Bengoetxea. 2021. Multiaztertest@vaxxstance-iberlef 2021: Identifying stances with language models and linguistic features. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

Gorrell, G., E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Hercig, T., P. Krejzl, B. Hourová, J. Steinberger, and L. Lenc. 2017. Detecting stance in czech news commentaries. In *Proceedings of the 17th ITAT: Slovenskočeský NLP workshop (SloNLP 2017)*, volume 1885 of *CEUR Workshop Proceedings*, pages 176–180, Bratislava, Slovakia. Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, CreateSpace Independent Publishing Platform.

Kiesel, J., M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, and M. Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Lai, M., A. Cignarella, D. Hernandez Farias, C. Bosco, V. Patti, and P. Rosso. 2020. Multilingual Stance Detection in Social Media Political Debates. *Computer Speech & Language*, 02.

Lai, M., A. T. Cignarella, L. Finos, and A. Sciandra. 2021. Wordup! at vaxxstance 2021: Combining contextual information with textual and dependency-based syntactic features for stance detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

Mohammad, S., S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *SemEval-2016)*, pages 31–41.

Montes, M., P. Rosso, J. Gonzalo, E. Aragón, R. Agerri, M. Ángel Álvarez Carmona, E. Álvarez Mellado, J. C. de Albornoz, L. Chiruzzo, L. Freitas, H. G. Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. de Arco, and M. T. (eds.). 2021. Proceedings of the iberian languages evaluation forum (iberlef 2021). CEUR Workshop Proceedings.

Taulé, M., F. M. R. Pardo, M. A. Martí, and P. Rosso. 2018. Overview of the Task on Multimodal Stance Detection in Tweets on Catalan# 1oct Referendum. In *IberEval@ SEPLN*, pages 149–166.

Thorne, J., A. Vlachos, C. Christodoulopoulos, and A. Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June. Association for Computational Linguistics.

Vychegzhanin, S. V. Kotelnikov, E. V. 2019. Stance Detection Based on Ensembles of Classifiers. *Programming and Computer Software*, pages 228–240, January.

Zotova, E., R. Agerri, and G. Rigau. 2021. Semi-automatic generation of multilingual datasets for stance detection in Twitter. *Expert Systems with Applications*, 170:114547.

A.4 Fernandez de Landa and Agerri (2022)

Relational Embeddings for Language Independent Stance Detection

Joseba Fernandez de Landa and **Rodrigo Agerri**
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{joseba.fernandezdelanda,rodrigo.agerri}@ehu.eus

Abstract

The large majority of the research performed on stance detection has been focused on developing more or less sophisticated text classification systems, even when many benchmarks are based on social network data such as Twitter. This paper aims to take on the stance detection task by placing the emphasis not so much on the text itself but on the interaction data available on social networks. More specifically, we propose a new method to leverage social information such as *friends* and *retweets* by generating relational embeddings, namely, dense vector representations of interaction pairs. Our method can be applied to any language and target without any manual tuning. Our experiments on seven publicly available datasets and four different languages show that combining our relational embeddings with textual methods helps to substantially improve performance, obtaining best results for six out of seven evaluation settings, outperforming strong baselines based on large pre-trained language models.

1 Introduction

Stance detection consists of identifying the viewpoint or attitude expressed by a piece of text with respect to a given target. With the enormous popularity of social networks, users spontaneously share their opinions on social media, generating a valuable resource to investigate stance. This means that research on stance has a social impact, for example, to help addressing misinformation on vaccines, or to better understand public opinion about topics such as abortion, climate change or migration. Furthermore, stance detection is considered an important intermediate task for fact-checking (Augenstein, 2021) or fake news detection[1].

The SemEval 2016 task on stance detection in Twitter (Mohammad et al., 2016) presented a dataset with tweets expressing FAVOR, AGAINST and NEUTRAL stances with respect to five different targets, a trend followed by many other researchers (Derczynski et al., 2017; Taulé et al., 2018; Zotova et al., 2021; Hardalov et al., 2021a). However, in spite of many of them using Twitter-based datasets, the large majority address the task by considering only the textual content (tweets) (Augenstein et al., 2016; Mohammad et al., 2017; Schiller et al., 2020; Hardalov et al., 2021b; Li et al., 2021; Ghosh et al., 2019; Küçük and Can, 2020; Sobhani et al., 2017; Glandt et al., 2021a).

This shortcoming has recently been addressed by proposing new datasets (Cignarella et al., 2020; Agerri et al., 2021) including social interaction data such as *retweets* or *friends*. Although they have facilitated new approaches to stance detection considering also interaction data, most of the them employ manually engineered features tailored to each specific data type (Espinosa et al., 2020; Lai et al., 2021; Alkhalifa and Zubiaga, 2020) making it difficult to generalize over other languages and targets. Thus, further research is required to fully understand the potentiality of interaction data to perform stance detection and its relation with concepts such as political homophily, political polarization, echo chambers or demographic analysis (Conover et al., 2011; Colleoni et al., 2014; Zubiaga et al., 2019).

This paper aims to perform stance detection of tweets by placing the emphasis on the interaction data commonly available in social media. To this end, we make the following contributions: (i) a new method to work with interaction data, such as *friends* and/or *retweets* by generating relational embeddings, focusing on one-to-one relations; (ii) experimentation on seven publicly available datasets and four different languages show that our relational embeddings behave robustly across different targets and languages; (iii) combining our method with text-based classifiers helps to systematically improve their results, outperforming also ensembles of large pre-trained language models (Gior-

[1] http://www.fakenewschallenge.org

gioni et al., 2020); (iv) an exhaustive ablation and error analyses show that the method to obtain the *retweet* data and the size of the users community is crucial for state-of-the-art performance using our technique.

2 Related work

Most of the work and datasets released on stance detection in Twitter does not include interaction data. Küçük and Can (2020) lists stance-annotated datasets for 11 languages, whereas recent work on cross-domain and cross-lingual stance provide experimentation for 16 datasets and 15 languages (Hardalov et al., 2021b,a). The focus, however, remains on the tweets text. This trend has recently changed with the release of, to the best of our knowledge, two datasets which, in addition to the stance labeled tweets, include interaction data such as *retweets* and *friends*.

The winner (Espinosa et al., 2020) of the SardiStance shared task (Cignarella et al., 2020) used a weighted voting ensemble that combined two inputs: (a) psychological, sentiment and *friends* distances as features used to learn a XGBoost (Friedman, 2001) model, with (b) text classifiers based on the Transformer architecture (Devlin et al., 2019). Other systems combined textual data (emoticons, special characters and word embeddings) with 2 dimensions extracted from the interactions distance matrix using Multidimensional Scaling (MDS) (Ferraccioli et al., 2020), or friendship-based graphs created with DeepWalk (Perozzi et al., 2014) and various types of textual embeddings (Alkhalifa and Zubiaga, 2020).

The VaxxStance shared task (Agerri et al., 2021) provided textual and interaction data (*friends* and *retweets*) to study stance detection on vaccines in Basque and Spanish. The one system that systematically outperformed the baselines (Lai et al., 2021) manually engineered a large number of features based on stylistic, tweet and user data, lexicons, dependency parsing, and network information, which were specifically developed for this dataset and languages.

Latest approaches tackling unsupervised stance detection using social media interactions as features, use the force-directed algorithm (Fruchterman and Reingold, 1991) or UMAP (McInnes et al., 2018). The algorithms are used to transform interactions frequency vectors into features, reducing huge interaction matrices into low dimensional features. Darwish et al. (2020) use both force-directed algorithm and UMAP for unsupervised stance detection on Twitter users. UMAP is also used to get interactions based features for automatically tagging Twitter users' stance on different topics (Stefanov et al., 2020) and to explore political polarization in Turkey (Rashed et al., 2021).

Other well known algorithms for building interaction-based models like DeepWalk (Perozzi et al., 2014) and node2vec (Grover and Leskovec, 2016) are based on generating Random Walks. However, those randomly generated walks create artificial interactions, that may not occur in the gathered interaction pairs. Furthermore, selecting the structure of the random walks and deciding the number of context users to be predicted have to be manually modeled and adapted to each reality.

In contrast to previous work, our method provides dense interaction-based representations of users, focusing on real interaction pairs. The training process is designed to predict a target user receiving a *retweet* or a *follow* from a source user, being each instance an item-to-item prediction instead of context-to-item (CBOW) or item-to-context (Skip-gram) prediction. Additionally, we focus on all the interaction pairs, without generating artificial random interactions to train the model or manually selecting the most salient users.

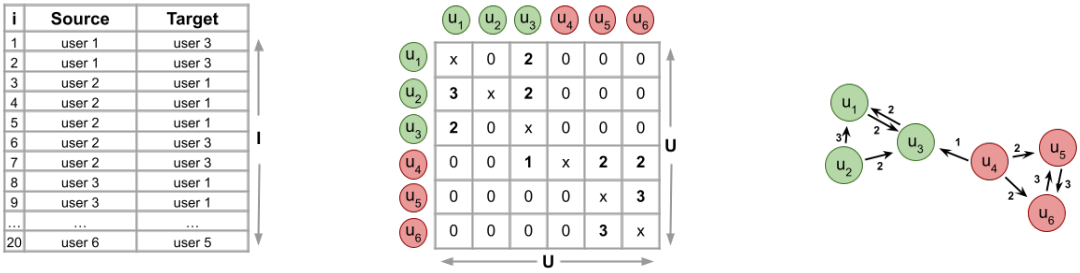
3 Method

Our new method consists of generating new vector-based representations of interactions in social networks, such as *friends* and *retweets*. We refer to these representations as *relational embeddings*. In addition to a classifier based just on the relational embeddings, we also provide two different approaches to combine existing textual classifiers with our relational embeddings.

3.1 Relational Embeddings

In this paper the interactions considered are *retweets* and/or *friends*, which are seen as relations between two users, one generating the action (source) and the other receiving it (target). Thus, users' acts of retweeting a tweet or following other users are considered interaction pairs. Generally, these interactions should help to reveal user's preferences by capturing meaningful information from their performative actions.

The first step in our method consists of gathering the interactions from the users included in



(a) Interaction pairs (b) Interaction frequency matrix (c) Interaction network

Figure 1: Different representations of the same data based on 20 interactions (I), generating a made-up directed network with 6 nodes (U) and 9 edges

the labeled data, namely, the one-to-one *retweet* and *follow* actions between the users/authors of the tweets. It should be noted that a set of *retweet* and *follow* interactions can consist of independent one-to-one actions without direct relation between them. This motivates our model to consider each interaction pair as a single instance.

Our model is trained in a unsupervised manner to predict a target user from a given user in each instance, being directly fed with interactions pairs (Figure 1a) instead of sparse interaction frequency matrices (Figure 1b) or neighbours arisen from interaction networks (Figure 1c). Furthermore, interaction based data is used without any preprocessing or modification.

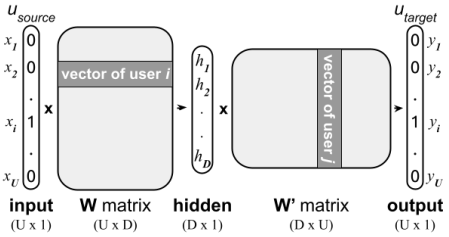


Figure 2: One hidden layer artificial neural network.

In order to obtain our relational representations, we use a single hidden-layer neural network (Figure 2). The network is used to train a dense interaction representation model using the *friends* and/or *retweet* based data. Each user is encoded as a one-hot vector of size U, where U is the number of users among interaction pairs (I) in a specific dataset. Given a one-hot vector U, the aim of the single hidden-layer feedforward neural network consists of predicting the target user. The dimensions of the hidden layer (D) determine the size of the final user relational vectors, which correspond to the number of learned features. During training, the weights W and W' are modified to minimize the loss function due to backpropagation. According to Equation 1, the summation goes over all the interaction pairs (I) in the training corpus, computing the log probability of correctly predicting the target user (u_{target}) from the source user (u_{source}) for each interaction (i). The training process is done by sub-sampling the most frequent instances and with negative sampling (Mikolov et al., 2013). Finally, the W matrix is used to retrieve the interaction vectors representing each user, generating the Relational Embedding, from which the relation vector for each user are obtained. In this model, users with similar interactions should have similar representations, turning many interaction pairs into dense relational representations of D dimensions.

$$\frac{1}{I}\sum_{i=1}^{I}\log p(u_{target}|u_{source}) \quad (1)$$

3.1.1 Relational Embeddings + SVM

Our first system consists of a linear classifier fed by the relational embeddings described earlier, without textual input. Building such a simple system will allow us to understand the performance of the generated relational embedding models on their own. Each of the tweets from a dataset will be represented by its author's (user) relations vector, which represents the interactions of its author. By doing so, we effectively project the relations of the author into tweet level, generating a link between the relational data and the stance labels. In this step it is possible that some users may be repeated among data, but their assigned stance label will be that of the corresponding tweet. It should be noted that although possible, it is quite uncommon to have a user with different labeled tweets across the data. Thus, each tweet is converted into relational

vectors, represented by the specific user's vector weights in the relational embedding model. Those users not present in the model are represented as a vector of zeros. This is usually due to the inability of retrieving user interaction data, either because the user has disappeared from Twitter or because their profiles are kept private. As shown in Figure 3, the final relational vectors for each tweet are used to train a SVM (RBF kernel) classifier without any additional input.

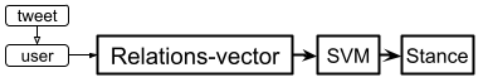


Figure 3: Relational Embeddings + SVM model architecture.

3.2 Text-based Classifiers

In order to compare relational with textual only approaches, we choose three commonly used text classification systems for stance detection (AlDayel and Magdy, 2020; Küçük and Can, 2020; Hardalov et al., 2021b; Zotova et al., 2021) to establish a baseline to compare with our relational embeddings models: (i) SVM learning algorithm with averaged word embeddings as text representations; (ii) SVM with TF-IDF vectorization and, (iii) large-pretrained multilingual Transformer-based models (Devlin et al., 2019; Conneau et al., 2019).

3.2.1 Word Embeddings + SVM

Word embeddings encode words in a low dimensional space, being able to capture semantic information. We use FastText CommonCrawl models trained using the C-BOW architecture and 300 dimensions on a vocabulary of 2M words. In order to represent OOV words, FastText word embeddings are trained with character n-grams (Grave et al., 2018). For classification, each tweet is represented as the average of its word vectors (Kenter et al., 2016). The tweet vector representation is used to train a SVM (RBF kernel) classifier.

3.2.2 TFIDF + SVM

TF-IDF (Term Frequency Inverse Document Frequency) vectorization is applied in order to reduce word vector dimensionality by lowering the impact of words that occur too frequently in the selected corpus. TF-IDF vectorizing is applied over the text of the tweets, selecting most salient features and reducing sparsity. The obtained TF-IDF vectors are used to learn a SVM (RBF kernel) model.

3.2.3 Multilingual Masked Language Models

We use multilingual BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2019)for text classification. mBERT is pre-trained with the largest 100 Wikipedias. XLM-RoBERTa was pre-trained for 100 languages on 2.5 TB of CommonCrawl text. We use both models off-the-shelf providing as input the tweet texts and their stance labels. Both models have been widely tested for stance detection with state-of-the-art performance (AlDayel and Magdy, 2020; Ghosh et al., 2019; Küçük and Can, 2020; Espinosa et al., 2020; Zotova et al., 2021).

3.3 Combining Textual and Relational Models

In order to combine textual and relational data, we use both the texts conveyed by a given user and its associated social media interactions (relational embeddings). We devise two different strategies to combine the textual classifiers and the relational embeddings introduced in the previous section. The first strategy ($\oplus$ *RelEmb*) uses relational embeddings to classify stance based on vector distances with a textual classifier acting as a back-off. The second method develops an ensemble method combining textual and relational vectors before learning (+ *RelEmb (ens)*).

3.3.1 $\oplus$ *RelEmb* by User Class Distances

Relational embeddings are used to classify stance by projecting user interaction vectors into tweet instances. Each of the tweets is only represented by its author's relations vector as in Section 3.1.1. In this case $\oplus$ *RelEmb* is trained based on the distances among users of the same stance inside the relational embeddings, backing-off to a textual classifier whenever relational data is not available.

According to Equation 2, all labeled users from the training set (U) are used to compute and predict the class (C) of each unseen item's relation vector ($\bar{b}$) in the test set. Distances (*sim*) are computed for every training item ($\bar{u}$) from each stance label (U_c), obtaining the average distance for each of the three labels. The maximum similarity value ($\arg max_c$) of one of the three classes is then selected as the predicted class, assigning the class related to the nearest community.

$$\arg max_c \left(\frac{\sum_{u \in U_c} sim(\bar{b}, \bar{u})}{|U_c|} \right) \quad (2)$$

Finally, whenever there is not relational data available, the prediction of the tweet's class is done assigning the class predicted by one of the text-based systems described in Section 3.2. By doing so, we avoid feeding the system with zero-vectors.

3.3.2 + *RelEmb* by Textual and Relational Ensemble

We develop two different methods to combine Relational and textual representations with SVM and Transformers. In Figure 4, we obtain a FastText (*FTEmb*) dense or TF-IDF sparse word vectors to represent each tweet. Then, the relational embedding of each author will be concatenated to the textual vector, adding a vector of zeros to the textual vector if no relational information is available. Finally, the concatenation of the textual and relational vectors is used to learn a SVM (RBF kernel) model.

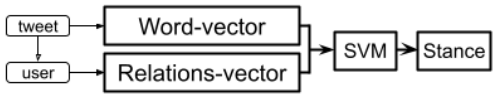


Figure 4: SVM based ensemble models architecture.

As shown in Figure 5, Transformer models and relational embeddings are combined by concatenating user vectors from the relations embeddins with the Transformer's CLS representations of the tweets. When there is no user information related to interactions, a vector of zeros is concatenated to the CLS representation. Finally, we add a linear classification layer on top of the CLS token vector concatenated with the relational vector and fine-tune the system end-to-end.

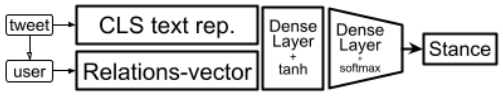


Figure 5: Transformer based ensemble models architecture.

4 Stance Detection Datasets

In order to experiment with our new relational embeddings, the datasets should include, in addition to the labeled textual data, interactions of the users that published the tweets, such as each user's *friends* and *retweets*. As far as we know, there are only two publicly available datasets with such contextual information, namely, SardiStance (Cignarella et al., 2020) and VaxxStance (Agerri et al., 2021). In order to include more data and languages, we tried to obtain user information for SemEval 2016 (Mohammad et al., 2016) and other English datasets (Lai et al., 2020; Conforti et al., 2020; Glandt et al., 2021b), without much success (less than 30% of users for SemEval or inability to retrieve tweets from the IDs). However, we did manage to retrieve relational information (over 80%) for other dataset, namely, the Catalonia Independence Corpus (Zotova et al., 2021).

The final choice consists of seven datasets on three different topics (Independence of Catalonia, antivaxxers, Sardines movement) and four languages (Basque, Catalan, Italian and Spanish). The number of labeled tweets and their distribution between train and test sets can be seen in Table 1. The choice of data offer a varied relation user-tweets (very low in SardiStance, quite high in CIC), which would also allow to test the robustness of the relational embeddings.

	Tweets			Relational Data		
	Train	Test	Total	Users	RTs	Friend
C-ca	8,038	2,010	10,048	691	10M†	24M†
C-ca*	8,056	1,992	10,048	691		
C-es	8,036	2,011	10,047	334		
C-es*	8,016	2,031	10,047	334		
S	2,132	1,110	3,242	2,827	575K	3M
S*	1,923	1,110	3,033	2,827		
V-eu	1,072	312	1,384	210	190K	170K
V-es	2,003	694	2,697	1,675	9K 552K†	2.1M

Table 1: Datasets: C (CIC), S (SardiStance), V (VaxxStance); * means no overlap of users across train and test; RT (retweets). † mark represents supplementary relational data added by us.

4.1 SardiStance

In addition to the textual data, this dataset also provides social and user information, such as the authors' friends and the retweets. As 16.30% of the users appear across the train and the test sets, we generated the alternative SardiStance* version by removing from the training set the users that also appear in the test set. The idea was to avoid

	C-ca	C-ca*	C-es	C-es*	SardiStance	SardiStance*	VaxxStance-es	VaxxStance-eu
FTEmb + SVM	61.60	62.64	57.42	59.77	56.13	54.65	71.29	47.67
⊕ RelEmb	75.81	77.13	80.54	86.11	71.06	70.94	72.38	69.01
+ RelEmb (ens)	82.17	69.20	88.55	**87.68**	74.03	71.33	89.06	73.17
TF-IDF + SVM	75.28	71.63	73.68	73.09	63.36	62.66	76.53	54.41
⊕ RelEmb	77.64	78.39	84.55	86.80	71.01	70.89	72.38	69.01
+ RelEmb (ens)	**82.23**	**80.17**	**92.48**	86.60	**74.55**	**74.01**	**90.23**	**75.30**
XLM-RoBERTa	77.63	74.55	74.21	73.88	57.17	56.40	82.52	41.17
⊕ RelEmb	78.39	78.56	84.98	87.17	71.06	70.98	72.38	69.01
+ RelEmb (ens)	78.80	75.92	76.81	77.17	60.16	55.62	81.15	51.81
mBERT	76.61	73.20	77.27	74.25	60.57	56.33	79.69	45.95
⊕ RelEmb	77.78	78.61	84.84	87.19	71.01	70.89	72.38	69.01
+ RelEmb (ens)	77.00	73.32	78.78	73.86	58.36	59.95	78.24	52.93
RelEmb + SVM	82.17	70.22	85.20	84.40	71.70	71.01	85.51	48.41
SOTA	74.68	74.87	74.72	71.84	74.45	-	89.13	77.71

Table 2: Results using relational embeddings only (RelEmb + SVM), Text-based systems (FTEmb + SVM, TF-IDF + SVM, mBERT and XLM-RoBERTa), and their combinations (⊕ RelEmb and + RelEmb (ens)). Previous SOTA: CIC (Zotova et al., 2020, 2021), SardiStance (Espinosa et al., 2020) and VaxxStance (Lai et al., 2021). Transformer results are the average of 5 randomly initialized runs.

using for training user-based features that are also available for testing. We were unable to extract any supplementary data, because both tweet and user identifiers are encrypted.

4.2 VaxxStance

The dataset was independently collected for two languages: Basque and Spanish. No user overlap across train and test set occurs in the data. Relational data is also included, such as *friends* from the users and *retweets* made to the labeled tweets. The Basque version (VaxxStance-eu) also includes retweets retrieved from the users timelines, as there are few tweets retrieved from the labeled tweets. In order to get more relational data, we extracted 552K supplementary retweets from the users timelines of the Spanish subset (VaxxStance-es), emulating the extra retweet collection as authors did for the Basque version.

4.3 Catalonia Independence Corpus

The Catalonia Independence Corpus includes coetaneous tweets in both Spanish and Catalan (Zotova et al., 2020), is multilingual, quite large (10K tweets) and reasonably balanced. In the original CIC data, 92.50% of the users in the Catalan set occur also in the test set, whereas for Spanish the proportion is even higher, namely, 99.72%. In order to avoid any possible overfit to authors' style, a second version of the dataset (Zotova et al., 2021) provides the same data collection but distributing the tweets in such a way that their authors do not appear across the training, development and test sets (CIC*).

5 Experiments

Retweets are used to share specific content from other users' publications. The reiteration of this actions may demonstrate attachment to a user or its content, actively showing the specific preferences of the source user. Furthermore, although retweet actions are more likely to gather latent information related to community or polarization (Conover et al., 2011; Zubiaga et al., 2019), we also wanted to include *friends* related data, which is a result of a *following* action. This passive action allows the source user to be aware of what is being said without sharing or promoting any content. Finally, we also combine both *retweets* and *friends* in a *mixed* representation to test whether merging passive and active interaction types in the same interaction space helps to embed social information. Therefore, for each dataset (CIC, SardiStance and VaxxStance) three different types of relational embeddings were trained, based on the data source: (i) *retweet*, (ii) *friends* and, (iii) *mixed* embeddings.

The best relational embedding for each dataset was chosen by evaluating them with the RelEmb + SVM system (which uses only relational embeddings) via 5-fold cross validation on the training data; *retweet* embeddings achieved best results for CIC and VaxxStance-eu, whereas the *mixed* embeddings were best for SardiStance and VaxxStance-es.

In this step we also learned that supplementary data extracted directly from the user's timelines, helped to improve the results.

The procedure to choose the rest of hyperparameters (relational embeddings dimensionality, etc.) are described in Appendix A. Finally, as it is customary for this task, despite training and predictions being done for the 3 classes, evaluation is performed by calculating the averaged F1-score over the AGAINST and FAVOR classes (Mohammad et al., 2016).

5.1 Evaluation Results

Table 2 reports the results obtained for every system and dataset. They show that combining our relational embeddings with the text-based classifiers systematically helps to substantially improve results. This results in SOTA performance for every language and dataset except for Basque. This is partially due to the fact that the system based only on relational embeddings (*RelEmb + SVM*) obtains high scores for most of the settings, often outperforming very strong baselines from mBERT and XLM-RoBERTa.

For SardiStance, our result is slightly better than the state-of-the-art (Espinosa et al., 2020) on this dataset. However, unlike our approach, which is based on relational embeddings and which does not require any manual tuning, they included a large number of manually engineered features based on external resources for sentiment, psychological features, as well as social network features. In any case, the results of *RelEmb + SVM* clearly improve over the best textual system published at the SardiStance shared task, which scored 68.53 in F1 score, based on an ensemble of Transformer models (Giorgioni et al., 2020).

Something similar occurs in the case of VaxxStance, where the state-of-the-art (Lai et al., 2021) for textual classifiers is of 57.34 for Basque and 80.92 for Spanish, respectively. The results reported by Lai et al. (2021) were obtained by manually implementing more than 30 different features tailored to each specific dataset and language. The features were based on stylistic information, tweet and user data, various lexicons, dependency parsing, and network information. Furthermore, they crawled 1M tweets for each language to obtain a larger word embedding model to generate also word embedding-based features. However, without any specific tuning to the dataset, or so many additional resources, we obtain new SOTA for VaxxStance-es and competitive results on Basque.

The results of text-based classifiers and their two combination types can be seen as ablation tests. In this sense, it is noticeable that ensemble methods (+ *RelEmb (ens)*) with SVM are most of the time better than the models based on user class distances ($\oplus$ *RelEmb*), while the opposite is true for Transformers. This suggests that backing-off to the already high-performing Transformer text classifier whenever no user information is available is better than generating ensembles adding vectors of zeros.

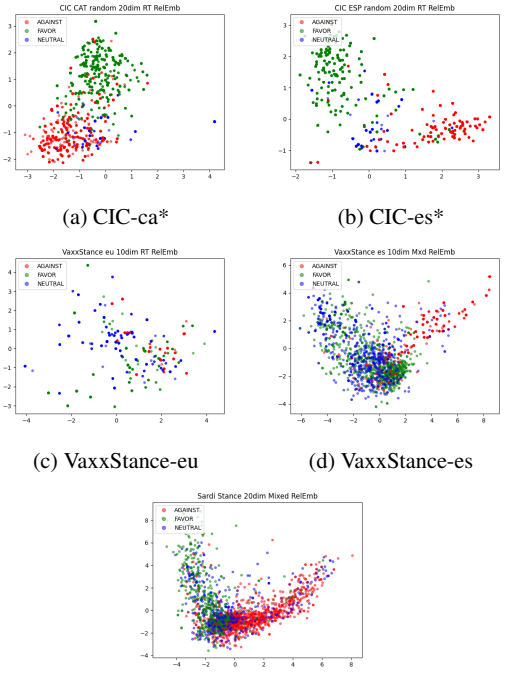
The best overall system is the one combining relational embeddings with TF-IDF + SVM, systematically outperforming any other system. While this might be a bit surprising, TF-IDF and statistical classifiers have proven to be highly competitive for stance detection in Twitter (Mohammad et al., 2016; Ghosh et al., 2019; Küçük and Can, 2020; Zotova et al., 2021), even better than text classifiers leveraging large pre-trained language models.

6 Discussion

In order to have a better understanding of the relational embeddings, we plotted the users' relational embeddings and their stance. Figure 6 shows a 2D visualization obtained by applying a PCA dimensionality reduction to each relational embedding from the training data.

The visualizations show that there is a clear relation between the readability of the relational embedding visualizations and the results obtained by the relational embedding systems (RelEmb + SVM). First, CIC embedding visualizations (Figures 6a and 6b) show very clearly defined communities of users who are in FAVOR or AGAINST of the independence of Catalonia. The relational embeddings obtained from the SardiStance relational data also correlate to well-defined communities (Figure 6e) but with more overlap of stance labels than in the CIC representations, as is the case for VaxxStance-es.

However, the visualization of the VaxxStance-eu relational embeddings in Figure 6c shows that target is quite transversal across users. This might be explained by the very small community of Basque Twitter users (see Table 1). In other words, Basque users in FAVOR or AGAINST vaccines naturally interact much more than users from the other two analyzed datasets. This might also explain why the relational embeddings classifier (*RelEmb + SVM*)

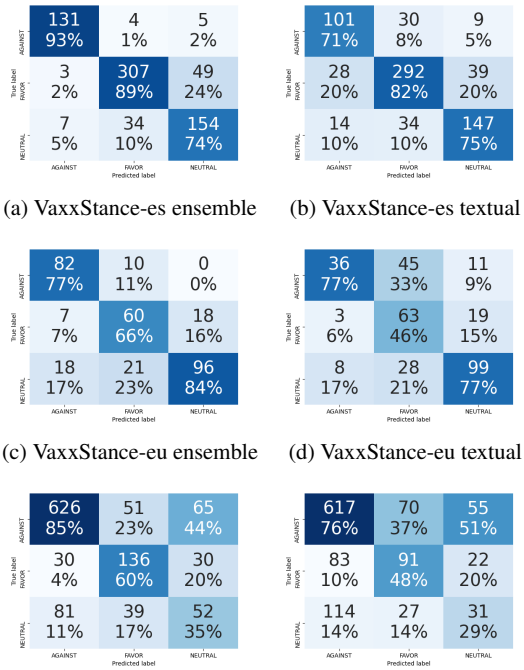


(a) CIC-ca* (b) CIC-es*

(c) VaxxStance-eu (d) VaxxStance-es

(e) SardiStance

Figure 6: Relational embedding representation of training set users (PCA dimension reduction to 2).



(a) VaxxStance-es ensemble (b) VaxxStance-es textual

(c) VaxxStance-eu ensemble (d) VaxxStance-eu textual

(e) SardiStance ensemble (f) SardiStance textual

Figure 7: Confusion matrices comparing textual and ensemble TF-IDF + SVM based models.

only fails to improve results over every textual classifier for the VaxxStance-eu data.

Finally, the graphs show that the specific nature of some targets makes them more suitable to generate our relational embeddings. Thus, topics that may reflect political homophily, such as the independence of Catalonia, seem to generate clearer FAVOR and AGAINST communities than for the other two topics (vaccinations and the Sardines movement). However, in the case of VaxxStance, the small size of the community of Basque Twitter users, may add difficulties to catch those orientation in such a small dataset.

If we combine the graphical visualizations with the confusion matrices of textual and ensemble classifiers in Figure 7, we can confirm some of the points raised above. In particular, that gains over previous SOTA results based on textual classifiers are much higher for those cases in which relational embeddings better discriminate between user communities.

7 Conclusion and Future Work

We propose a new method to perform stance detection by generating relational embeddings from social interaction data such as *retweets* and *friends*. The relational embeddings help to reduce the sparsity of interaction data by behaving like dense graphs, being able to embed information related to stance for different data sources without any manual engineering. While this technique is language independent, cheap and fast to train and to apply, the relational embeddings behave robustly across different datasets, stance targets and languages, helping to substantially and consistently improve results by combining them with text-based classifiers.

The results and analysis performed shows that we need to pay more attention to social network data, aiming to address the shortcomings discussed by further researching different strategies to leverage such interaction data. We are aware that our system is conditioned to the availability of the relational data of the user that wrote the tweet, which means that when collecting data for training and inference such data should also be gathered.

Future work may include analyzing the relational embeddings performance on zero-shot and cross-lingual settings, moving on towards a method that, by using user-based relational information, helps to drastically reduce the need of annotated data at tweet level.

References

Rodrigo Agerri, Roberto Centeno, María Espinosa, Joseba Fernandez de Landa, and Álvaro Rodrigo. 2021. Vaxxstance@iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Abeer AlDayel and Walid Magdy. 2020. Stance Detection on Social Media: State of the Art and Trends. *arXiv preprint arXiv:2006.03644*.

Rabab Alkhalifa and Arkaitz Zubiaga. 2020. QMUL-SDS@SardiStance: Leveraging Network Interactions to Boost Performance on Stance Detection using Knowledge Graphs. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR Workshop Proceedings.

Isabelle Augenstein. 2021. Towards explainable fact checking. *ArXiv*, abs/2108.10274.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. In *ACL*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv:1911.02116*.

Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Maria S. Espinosa, Rodrigo Agerri, Alvaro Rodrigo, and Roberto Centeno. 2020. DeepReading@SardiStance: Combining Textual, Social and Emotional Features. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR Workshop Proceedings.

Federico Ferraccioli, Andrea Sciandra, Mattia Da Pont, Paolo Girardi, Dario Solari, and Livio Finos. 2020. TextWiller@SardiStance, HaSpeede2: Text or Context? A smart use of social network data in predicting polarization. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR Workshop Proceedings.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Thomas M. J. Fruchterman and Edward M. Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: a comparative study. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 75–87. Springer.

Simone Giorgioni, Marcello Politi, Samir Salman, Roberto Basili, and Danilo Croce. 2020. Unitor @ sardistance2020: Combining transformer-based architectures and transfer learning for robust stance detection. In *EVALITA*.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021a. Stance detection in covid-19 tweets. In *Proceedings of the*

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, volume 1.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021b. Stance detection in covid-19 tweets. In *ACL/IJCNLP*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabel Augenstein. 2021a. Few-shot cross-lingual stance detection with sentiment-based pre-training. *ArXiv*, abs/2109.06050.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021b. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese CBOW: Optimizing word embeddings for sentence representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 941–951, Berlin, Germany. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance Detection: a Survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Mirko Lai, Alessandra Teresa Cignarella, Livio Finos, and Andrea Sciandra. 2021. Wordup! at vaxxstance 2021: Combining contextual information with textual and dependency-based syntactic features for stance detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2020. #brexit: Leave or remain? the role of user's community and diachronic evolution on stance detection. *Journal of Intelligent & Fuzzy Systems*, 31(2):2341–2352.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021. Improving stance detection with multi-dataset learning and knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6332–6345, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.*, 3:861.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3):26:1–26:23.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 701–710.

Ammar Rashed, Mucahid Kutlu, Kareem Darwish, Tamer Elsayed, and Cansin Bayrak. 2021. Embeddings-based clustering for target specific stances: The case of a polarized turkey. In *ICWSM*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2020. Stance detection benchmark: How robust is your stance detection? *ArXiv*, abs/2001.01565.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *ACL*.

M. Taulé, F. Rangel, M. A. Martí, and P. Rosso. 2018. Overview of the task on multimodal stance detection in tweets on catalan 1oct referendum. In *IberEval 2018. CEUR Workshop Proceedings. CEUR-WS.org*, pages 149–166, Sevilla, Spain.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1368–1375, Marseille, France. European Language Resources Association.

Elena Zotova, Rodrigo Agerri, and German Rigau. 2021. Semi-automatic generation of multilingual datasets for stance detection in Twitter. *Expert Systems with Applications*, 170:114547.

Arkaitz Zubiaga, Bo Wang, Maria Liakata, and Rob Procter. 2019. Political homophily in independence movements: Analyzing and classifying social media users by national identity. *IEEE Intelligent Systems*, 34:34–42.

A Appendix

The relational embedding types and dimensionality for the RelEmb + SVM model were chosen via grid search with 5-fold cross-validation. The results of cross-validation showed that low dimensional Relational Embeddings (10-20 dimensions) performed significantly better than high dimensional ones, following the idea of reducing huge interaction matrices into low dimensional features (Darwish et al., 2020). The best performing relational embeddings for CIC and SardiStance were of dimension 20, while for VaxxStance they were of dimension 10.

The same method was also used to optimize C and Gamma hyperparameters for every SVM system (RelEmb, FTEmb and TF-IDF).

For mBERT and XLM-RoBERTa, hyperparameter tuning was done by splitting the training set into a train and development sets (80/20). Results on the development set allowed to obtain the following hyperparameters: 128 maximum sequence length, 16 batch size, 2e-5 learning rate and 5 epochs.

A.5 Fernandez de Landa and Agerri (2023)

HiTZ-IXA at PoliticES-IberLEF2023: Document and Sentence Level Text Representations for Demographic Characteristics and Political Ideology Detection

Joseba Fernandez de Landa, Rodrigo Agerri

HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country UPV/EHU

Abstract

In this paper we describe our participation to the PoliticES 2023 shared task held at IberLEF 2023. The task focuses on extracting demographic and political information from tweets, and it is structured as an author profiling task. Our participation is focused on developing a multi-level textual representation that combines both the tweets text and user representations. This approach allows us to effectively capture and integrate social information, including demographic and ideological traits. Furthermore, our text-based features leverage document and sentence information, amalgamating specific and general aspects. The combination of both social and textual features results in a remarkable improvement in overall performance across the various text classification tasks proposed within the task. An additional benefit of our approach is its robustness and generalization capability, as it performs competitively using same features across all traits. Finally, we address potential memory constraints by efficiently managing extensive timelines or documents, segmenting them into individual sentences or tweets while keeping the document level information. Our technique offers promising results in effectively handling large-scale textual data in document classification tasks. Our system achieved the second highest score for the overall PoliticES task and the best score for predicting the *profession* category.

Keywords

Demographic Traits detection, Political Ideology detection, Author Profiling, Computational Social Science, Natural Language Processing

1. Introduction

This paper describes the HiTZ-IXA team participation in PoliticEs2023 shared task [1] organised in IberLEF 2023 [2], which consists of extracting demographic and political information from texts. Framed as an author profiling task, the objective is to extract Twitter user's characteristics based on 80 distinct text-based documents per author. By leveraging text-based Twitter data in Spanish language, the aim is to extract demographic traits including gender and profession, as well as, political ideology approached from both a binary and a multiclass perspective, from a given set of tweets.

There exists a significant interest in extracting demographics and ideology from Social Media,

IberLEF 2023, September 2023, Jaén, Spain

joseba.fernandezdelanda@ehu.eus (J. F. d. Landa); rodrigo.agerri@ehu.eus (R. Agerri)

0000-0001-6067-3571 (J. F. d. Landa); 0000-0002-7303-7598 (R. Agerri)

as it represents a means of gaining deeper insights into society. Twitter has become a source of spontaneously generated textual data for many human languages, and its use for doing demographic and ideological inferences increasingly common [3, 4]. As for most research topics in Natural Language Processing (NLP), recent works have experimented with Transformer-based [5] contextualized sentence embeddings for user level demographic prediction such as age or gender [6, 7]. Thus, Transformer-based approaches were the most common method in PoliticES 2022 shared task [8]. However, a limitation of those approaches is that they are usually centered on document or sentence level representations only.

In order to harness user-related data for text classification problems in social media, tasks such as stance detection have been approached from a perspective that utilizes both author and tweet characteristics to infer stance at tweet level [9, 10, 11] . In those works, authors are represented through Twitter interactions such as *friends*, *retweets*, *quotes*, or *replies*, since there is a lack of accessible data for utilizing author-level texts. These author representations are subsequently combined with textual representations derived from specific tweets, resulting in improved overall performance [11]. Therefore, taking this idea as a starting point, our approach for PoliticES focuses on the combination of user and tweet-level representations, but unlike previous aforementioned work, both derived exclusively from textual data.

Thus, in this paper we present a multi-level textual representation combining tweet and user representations in order to embed social information such as demographic and ideological traits. The contribution of the proposed text-based features lies in their ability to leverage document and sentence information, combining features derived from both sentence and author-based representations. Results demonstrate that such integration of user and tweet representation levels enhances the ability to capture meaningful information, thereby improving performance in various text classification tasks.

Our method is robust and exhibits good generalization capabilities obtaining good performance across all tasks using the same features. Furthermore, it shows the capacity to effectively manage extensive timelines or documents by dividing them into individual sentences or tweets, thereby mitigating potential memory constraints. The official results show that our system was ranked 2nd from 11 participants among the general task and in the 1st position for the *profession* category.

2. Related Work

In previous work related to this topic, we must highlight the previous task PoliticES 2022 [8], which focused on author profiling by employing text-based data in order to extract demographic and ideological traits. Most of the systems participanting in the task were based on Transformer [5] models. More specifically, there is a notable presence of monolingual models in Spanish, especially BETO [12] and MarIA [13]. The rest of the section provides an overview of the key features exhibited by the top four models presented on the 2022 task.

The first model, proposed by Carrasco and Rosillo [14], employs 512 token-blocks comprising tweets from the same author in the dataset, along with additional data, to fine-tune a combined model of BETO and MarIA. This combined model is used to predict labels at the token-block level. Subsequently, user characteristics are predicted using a majority vote strategy based on

the aforementioned token-blocks.

In the second model, Villa-Cueva et al. [15] introduce PolitiBETO, a BETO model that is pre-trained on data derived from social media and news texts. Using this specific model, predictions are made at tweet level and then aggregated through a majority vote to infer author labels.

The third model [16] employs all author's tweets to extract author features. Word and character n-grams, as well as lexical and stylistic features, are used to feed the model, manually selecting them for each of the categories.

The fourth model [17] groups tweets belonging to the same author into clusters containing 8-12 tweets, grouping more information while also accommodating memory constraints. For each category different classification techniques are presented, and manual engineering is applied accordingly. Additionally, a voting system is employed to unify the labels of tweet clusters into user labels.

In summary, three out of the four leading models are based on the Transformer architecture, with the top two specifically relying on monolingual Spanish Transformer-based language models. In addition, tweets authored by the same user tend to be grouped by different techniques, either by concatenating them at the input stage or by merging the associated labels at the output stage.

The exploration of grouping techniques for tweets authored by the same user reflects the ongoing efforts to improve the handling of sequential and contextual information in social media data in Transformer-based classifiers. The decision to concatenate tweets at the input stage or merge associated labels at the output stage implies a deliberate consideration of how to capture and leverage the inherent relationships and dependencies within user-generated content. Nonetheless, in previous works the classifiers were not given both textual and user-based information as input, a shortcoming that our approach tries to address.

3. Datasets

The dataset employed on this work is an expansion of the PoliCorpus 2020 dataset [18] and the corpus utilized for the PoliticES 2022 shared task. It encompasses information extracted from Twitter accounts belonging to politicians, political journalists, and celebrities in Spain. Political accounts were selected among members of the Spanish government, the Congress and Senate of Spain, mayors of important Spanish cities, presidents of the autonomous communities, former politicians, and collaborators affiliated with political parties. Furthermore, journalists were selected from various Spanish news media such as *ABC*, *El País*, *El Diario*, *El Mundo* or *La Razón*, among others.

The objective of creating such dataset is to facilitate techniques for the extraction of demographic characteristics and political ideology from a provided user's collection of tweets. Demographic attributes encompass elements like gender and profession, while political ideology is approached both as a binary and a multiclass problem. Users are annotated by gender (*male* or *female*), profession (*politician, journalist* or *celebrity*), and by political alignment along two axes: a binary scale (*left* or *right*) and a multi-class scale (*left, moderate left, moderate right, right*). To ensure the users privacy, they created clusters of 80 tweets each, with each user-cluster containing tweets from different users that share all the traits under evaluation. In this way

user-clusters are used instead of the users themselves with the objective of avoiding to incur in any legal and ethical issues.

In addition to users, the textual content is also anonymized. Thus, tweets that shared content from or mention news websites is filtered. Moreover, any mention to politicians on Twitter is substituted with the token *@user*, while mentions of other Twitter accounts are encoded as *@user*. Finally, References to political parties are also replaced with the token *@political_party*.

The dataset comprises approximately 2800 user-clusters, with each user-cluster containing diverse texts from different dates and topics. The user-clusters from the training and test sets are independent to prevent the possibility of identifying the authors. As shown by the train set quantitative description provided in Table 1, the train set contains 2250 user-clusters (180,000 tweets) whereas the test set includes 547 (43760 tweets).

Category	Class	Tweets	User-clusters	Distribution
Gender	male	119,440	1,493	66.36%
	female	60,560	757	33.64%
Profession	journalist	110,800	1,385	61.56%
	politician	60,160	752	33.42%
	celebrity	9,040	113	5.02%
Political Ideology: binary	left	100,400	1,255	55.78%
	right	79,600	995	44.22%
Political Ideology: multiclass	left	34,400	430	19.11%
	moderate left	66,000	825	36.67%
	moderate right	58,240	728	32.36%
	right	21,360	267	11.86%

Table 1
PoliticES 2023 train set statistics by category and class.

4. Methods

In this section we describe our two techniques to represent textual data at different levels. Firstly, we present a Transformer-based tweet-and-user representation (t&u) method, consisting of representing both levels in the same feature. Secondly, we also introduce a token-level user representation technique which we named word-to-user (w2u).Those representations are then used to feed a text classifier based on each category, but sharing the same unaltered features.

4.1. Tweet-and-user representations

First of all Transformer-based [5] language models are utilized to obtain contextual text-based representations at sentence or tweet level. These models focus on capturing context and meaning by analyzing the relationships among tokens in a text sequence. In contrast to static embedding methods like word2vec[19], which represent words with fixed vector values, Transformer-based models modify those vectors values depending on the surrounding words and their order. It should be noted that this approach is not limited to word-level representations, but rather it

can also handle sequences of text, making it suitable for text classification tasks similar to ours. For our experiments, we have selected the following models:

• mBERT [20] is the multilingual version of BERT[20] pre-trained with the largest 104 languages in Wikipedia. Rather than simply predicting the next word in the sequence, the BERT model takes into consideration all of the words in the sequence, thereby developing a more dense and rich representation of the context. BERT pre-trains bidirectional representations from unlabeled text by considering both left and right context in all layers utilising next-sentence prediction and masked-language modeling.

• DistilmBERT [21], the multilingual version of DistilBERT, a smaller and faster Transformer model distilled from BERT. It retains over 95% of BERT's performance on the GLUE benchmark while having 40% fewer parameters and 60% faster inference speed.

• XLM-RoBERTa [22] is a multilingual version of RoBERTa [23], trained with the CC100 corpus, on a large multilingual dataset spanning 100 languages. It is an optimized BERT variant that benefits from training on a dataset ten times larger than BERT, employing dynamic masking, byte-pair encoding tokenization, and omitting the next-sentence prediction objective.

• XLM-T [24] is an extension of the XLM-RoBERTa base model, further trained with 198 million multilingual tweets. This model's focus on Twitter-based data makes it particularly relevant for evaluating performance in tasks specific to this social media platform.

• mDeBERTa [25], the multilingual version of DeBERTa, utilizes the same arquitecture of DeBERTa and, as XLM-RoBERTa, it was trained on the CC100 multilingual dataset, although just for 15 languages. DeBERTa improves BERT and RoBERTa models through disentangled attention and an enhanced mask decoder, outperforming RoBERTa on the majority of natural language understanding (NLU) tasks.

• BETO [12] is a BERT model trained on a Spanish corpus comprising 3 billion tokens. It is similar in size to BERT-base and was trained using the Whole Word Masking technique.

• PolitiBETO [15] is a BERT model specifically tailored for political tasks in social media corpora. It is created through a two-stage domain adaptation process applied to the BETO model, incorporating the language structure found on Twitter and in newspapers.

• MarIA [13] or roberta-large-bne is based on the RoBERTa-large model. It has been pre-trained on a Spanish corpus totaling 570GB of clean and deduplicated text sourced from the National Library of Spain.

We use the listed Transformer models to extract the features and evaluate their performance for the specific task. To represent each tweet, we utilize the last hidden state corresponding to the start-of-sequence token as an aggregate document representation, following the approach described in [20]. This hidden state serves as a tweet vector, which acts as a textual feature representation for the tweet. In order to maintain the same representations for all the categories the Transformers models are used without fine-tuning, meaning that we are using default frozen weights.

In order to extract author representations we average local elements to generate a global representation, as done in previous approaches [26, 27, 28]. In other words, the user representation is obtained by extracting the mean vector of all the tweet vectors authored by each user. Once we get the user representations, each of the tweet representations is concatenated with its author representation vector, generating a tweet-and-user representations for each of the tweets in the dataset.

Finally, the combined tweet-and-user representations for each tweet are used to train a Logistic Regression classifier without any additional tuning. Thus, the same features, without any modification, are used to train a classifier for each of the categories in the PoliticES dataset. After predicting the labels at tweet level, a majority voting strategy is employed to infer the user label by considering the various tweet labels associated with the same author.

4.2. Word-to-user representations

The *word-to-user* model is trained in a unsupervised manner to predict a target user from a given word-token. The input of the model is the raw text data without any preprocessing or modification. In order to obtain our user representations, we use a single hidden-layer neural network. The network is used to train a dense interaction representation model using the tokens from the users' text. The aim of the single hidden-layer feedforward neural network consists of predicting the target user from a given word token that appears in the corpus. The dimensions of the hidden layer determine the size of the final user representation vectors, corresponding to the number of learned features. During training, all the word tokens present in the training and test corpus are used, computing the log probability of correctly predicting the target user from the given word token. The training process is done by sub-sampling the most frequent instances and with negative sampling [19]. A number of experiments were undertaken in order to obtain the optimal dimensionality of the *word-to-user* representations. Once all the users are embedded, the resultant vector is used to represent a given user.

The final step consists of using the extracted word-to-user representation vectors from each user to train a Logistic Regression classifier without any additional tuning. This means that the same features are used across each of the categories in the dataset.

5. Experiments on the Development Data

We experimented with the development data in order to obtain the optimal configuration for the two methods described above. With respect to *tweet-and-user*, the objective is to establish which model is best to extract the contextual representations which are used as features. For the *word-to-user* approach we want to know which dimensionality to represent word-to-user features provides the best results.

5.1. Tweet-and-user

In order to select the best configuration for tweet-and-user representations, we use different feature extraction methods using the language models described in the previous section. The obtained tweet-and-user combined vectors are then used as input to train Logistic Regression classifiers, while employing majority voting. Development results (Table 2) show that the best configuration given by the MarIA model. Therefore, the features extracted from the MarIA model will be used to train the final model.

To conduct an ablation study, we utilize the tweet vectors to train a Logistic Regression classifier independently of the user vectors, while employing majority voting. This tweet-only approach allows us to assess the influence of incorporating tweet and user level information

Category	Language Model								
	mB	dmB	mR	mRl	mRt	mD	Be	pBe	Ma
gender	0.814	**0.817**	0.782	0.800	0.795	0.621	0.768	0.782	0.806
profession	0.725	0.818	0.818	0.599	0.827	0.724	**0.851**	0.701	0.743
ideology binary	0.666	0.719	0.776	0.786	0.761	0.661	0.745	0.749	**0.842**
ideology multiclass	0.582	0.618	0.610	0.593	0.647	0.504	0.588	0.628	**0.678**
average	0.697	0.743	0.747	0.695	0.758	0.628	0.738	0.715	**0.767**

Table 2
Tweet and user level classification. F1 macro score results on development set. Algorithms used to generate the features: mB (mBERT), dmB (DistilmBERT), mR (XLM-RoBERTa-base), mRl (XLM-RoBERTa-large), mRt (XLM-T), mD (mDeBERTa), Be (BETO), pBe (PolitiBETO), Ma (MarIA). Values in bold represent best results for each category.

in extracting valuable social insights. Therefore, the train set is used to train the classifier, while the development set is used to evaluate different algorithms for feature extraction. When comparing the results of tweet-and-user combined approach (Table 2) to the results of the tweet-only approach (Table 3), it can be observed that the latter exhibits a significant loss in performance.

Category	Language Model								
	mB	dmB	mR	mRl	mRt	mD	Be	pBe	Ma
gender	0.473	0.500	0.473	0.439	0.473	0.384	0.518	**0.525**	**0.525**
profession	0.526	0.526	0.526	0.533	0.535	0.503	**0.572**	0.535	0.546
ideology binary	0.675	0.675	0.625	0.660	0.720	0.486	0.670	0.679	**0.774**
ideology multiclass	0.391	0.313	0.294	0.292	0.401	0.263	0.446	0.358	**0.515**
average	0.516	0.504	0.480	0.481	0.532	0.409	0.552	0.524	**0.590**

Table 3
Tweet level classification as ablation test. F1 macro score results on development set. Algorithms used to generate the features: mB (mBERT), dmB (DistilmBERT), mR (XLM-RoBERTa), mRl (XLM-RoBERTa-large), mRt (XLM-T), mD (mDeBERTa), Be (BETO), pBe (PolitiBETO), Ma (MarIA). Values in bold represent best results for each category.

5.2. Word-to-user

In order to select the best configuration for word-to-user representations, we trained different Logistic Regression classifiers with different dimensions. Thus, results on the development dataset (Table 4) show that the best configuration is given by obtaining the word-to-user vector representions in 200 dimensions. It is remarkable the high performance achieved with this configuration, clearly outperforming the Transformer-based tweet-and-user approach on development set.

To have a better understanding of the word-to-user (200 dimensions) user-level representations, we plotted the users' representations and the corresponding class. Figure 1 and 2 show 2 dimensional visualizations obtained by applying a t-SNE dimensionality reduction to word-to-user representations arisen from the development data; each color represents a different class. Demographic characteristics are plotted on Figure 1 while ideological binary and multiclass

Category	Dimensions					
	10	20	50	100	200	300
gender	0.690	0.789	0.830	0.798	**0.855**	0.830
profession	0.601	0.841	0.862	0.732	**0.925**	0.795
ideology binary	0.818	0.850	0.820	**0.919**	0.910	0.851
ideology multiclass	0.577	0.652	0.678	**0.696**	0.681	0.657
average	0.672	0.783	0.798	0.786	**0.843**	0.783

Table 4
F1 macro score results on development set for w2u trained over different amount of dimensions. Values in bold represent best results for each category.

representations can be seen on Figure 2.

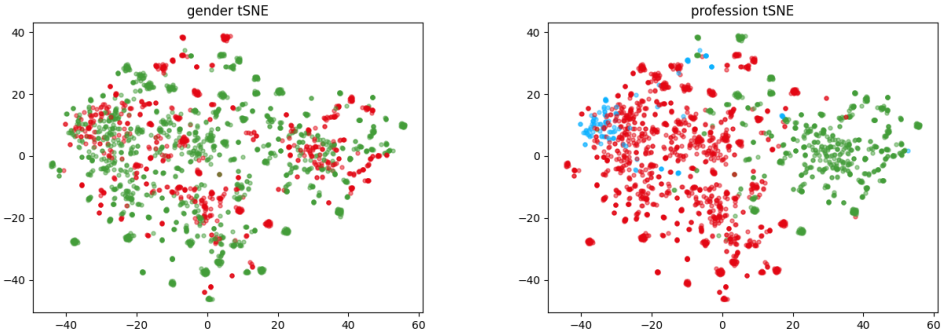


Figure 1: Visualization of t-SNE 2 dimension reduction of w2u (200 dim) user representations for gender (left) and profession (right) Demographic traits on development set.

Regarding Demographic traits, it can be seen that the classes present on profession are clearly defined (Figure 1 right), while the representations of gender (Figure 1 left) seem to be more sparse. With respect to political ideology, the binary framework (Figure 2 left) shows clearer communities than the more sparse multiclass framework (Figure 2 right). Thus, the evaluation results and the visual representations would seem to correlate, as the categories with clearer communities (profession and ideology binary) are also the categories that obtained better classification results on the development data.

6. Results on the Official Test Data

As a result of the experiments performed in the previous section, the *tweet-and-user* method will be using the MarIA model to obtain the combined vectors which are the input to Logistic Regression classifiers for each of the traits (while employing majority voting). In this setting, the classifiers are trained on the training data and evaluated on the official test set. The same procedure is applied for the *word-to-user* method, which will be based on training, for each

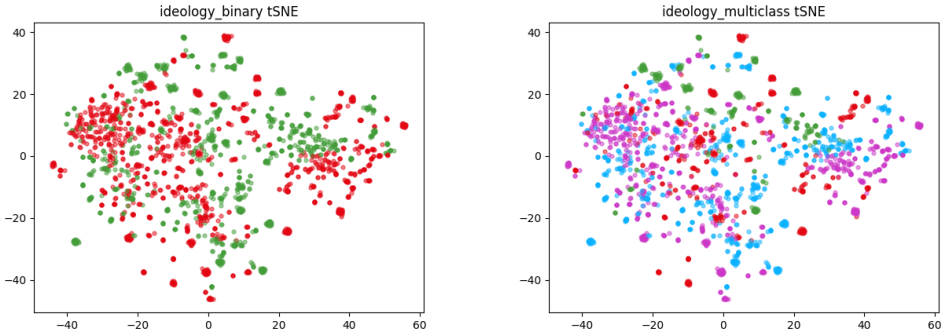


Figure 2: Visualization of t-SNE 2 dimension reduction of w2u (200 dim) user representations for binary (left) and multiclass (right) Political Ideology on development set.

of the traits, Logistic Regression classifiers which take as input features the 200 dimensional *word-to-user vectors*.

	Gen	Prof	Pib	Pim	**AVG**
baseline	66.34	60.24	79.77	54.72	65.27
w2u	75.88	78.67	87.27	63.78	76.40
t&u	73.79	85.48	85.57	59.36	76.05
t&u + w2u	**79.56**	**86.08**	**87.75**	**63.98**	**79.34**

Table 5
F1 macro average test scores for *gender* (Gen), *profession* (Prof), *ideology binary* (Pib), *ideology multiclass* (Pim) categories and overall average (AVG). Algorithms used to generate the features: word-to-user (w2u) 200d, tweet-and-user (t&u) based on MarIA and a combination between both (t&u + w2u).

In reference to the test results presented in Table 5, both the tweet-and-user (t&u) and word-to-user (w2u) approaches exhibit similar average scores, although variations can be observed among the traits. Consequently, we made the decision to merge both feature extraction techniques to assess their combined performance. In fact, it turned out that fusion of tweet-and-user and word-to-user (t&u + w2u) yielded the highest test scores across all categories and the overall average. Moreover, this combination outperformed w2u and t&u individually in all the assessed categories. These findings imply that the combination of features generated by tweet-and-user and word-to-user can result in improved performance when predicting gender, profession, and ideological aspects. Furthermore, the scores obtained provide valuable insights into the effectiveness of the algorithms and their ability to generalize across diverse categories.

7. Conclusion

This paper demonstrates the benefits of combining author and sentence level textual representations for political ideology detection and characterization of users with respect to demographic

traits. More specifically, for our participation to the PoliticES 2023 shared task we have experimented with different level Transformer-based textual features as well as with user features directly arisen from word tokens. This combination of features has allowed us to obtain the second-best overall results in the task using a general approach, namely, without performing any specific feature-engineering for any of the traits.

References

[1] J. A. García-Díaz, S. M. Jiménez-Zafra, M. T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticEs at IberLEF 2023: Political ideology detection in Spanish texts, Procesamiento del Lenguaje Natural 71 (2023).

[2] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.

[3] D. Nguyen, A. S. Doğruöz, C. P. Rosé, F. de Jong, Computational Sociolinguistics: A Survey, Computational Linguistics 42 (2016) 537–593. doi:10.1162/COLI_a_00258.

[4] N. Cesare, C. Grant, E. O. Nsoesie, Detection of user demographics on social media: A review of methods and recommendations for best practices, arXiv preprint arXiv:1702.01807 (2017).

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[6] J. Fernandez de Landa, R. Agerri, Social analysis of young basque-speaking communities in twitter, Journal of Multilingual and Multicultural Development 0 (2021) 1–15.

[7] M. Abdul-Mageed, C. Zhang, A. Rajendran, A. Elmadany, M. Przystupa, L. Ungar, Sentence-level bert and multi-task learning of age and gender in social media, arXiv preprint arXiv:1911.00637 (2019).

[8] J. A. García-Díaz, S. M. Jiménez-Zafra, M.-T. M. Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticEs 2022: Spanish author profiling for political ideology, Procesamiento del Lenguaje Natural 69 (2022) 265–272.

[9] R. Agerri, R. Centeno, M. Espinosa, J. F. de Landa, Álvaro Rodrigo, Vaxxstance@iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection, Procesamiento del Lenguaje Natural 67 (2021) 173–181.

[10] A. T. Cignarella, M. Lai, C. Bosco, V. Patti, P. Rosso, SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), CEUR-WS.org, 2020.

[11] J. Fernandez de Landa, R. Agerri, Relational embeddings for language independent stance detection, arXiv e-prints (2022) arXiv–2210.

[12] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[13] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). doi:10.26342/2022-68-3.

[14] S. S. Carrasco, R. C. Rosillo, Loscalis at PoliticEs 2022: Political author profiling using BETO and maria, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022, volume 3202 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022.

[15] E. Villa-Cueva, I. González-Franco, F. Sanchez-Vega, A. P. López-Monroy, NLP-CIMAT at PoliticEs 2022: PolitiBETO, a Domain-Adapted Transformer for Multi-class Political Author Profiling, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings, CEUR-WS, 2022.

[16] A. Mosquera, Alejandro mosquera at PoliticEs 2022: Towards robust spanish author profiling and lessons learned from adversarial attacks, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022, volume 3202 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022.

[17] E. Santibáñez-Cortés, A. Carrillo-Cabrera, Y. A. Castillo-Castillo, D. Moctezuma, V. Muñiz-Sánchez, Cimat_2021 at PoliticEs 2022: Ensemble based classification algorithms for author profiling in spanish language, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022, volume 3202 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022.

[18] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74.

[19] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[21] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, in: NeurIPS EMC2 Workshop, 2019.

[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: ACL, 2020.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019).

[24] F. Barbieri, L. Espinosa-Anke, J. Camacho-Collados, A multilingual language model toolkit for twitter, arXiv preprint arXiv:2104.12250 (2021).

[25] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021.

[26] I. R. Hallac, S. Makinist, B. Ay, G. Aydin, user2vec: Social media user representation based on distributed document embeddings, in: 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 2019, pp. 1–5. doi:10.1109/IDAP.2019.8875952.

[27] T. Kenter, A. Borisov, M. de Rijke, Siamese CBOW: Optimizing word embeddings for sentence representations, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 941–951. URL: https://aclanthology.org/P16-1089. doi:10.18653/v1/P16-1089.

[28] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

A.6 Fernandez de Landa *et al.* (2024a)

Uncovering Social Changes of the Basque Speaking Twitter Community during COVID-19 Pandemic

Joseba Fernandez de Landa[1], **Iker García-Ferrero**[1],
Ander Salaberria[1], **Jon Ander Campos**[2]
[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU, [2]Cohere
{joseba.fernandezdelanda, iker.garciaf, ander.salaberria}@ehu.eus
jonander@cohere.com

Abstract

The aim of this work is to study the impact of the COVID-19 pandemic on the Basque speaking Twitter community by applying Natural Language Processing unsupervised techniques. In order to carry out this study, we collected and publicly released the biggest dataset of Basque tweets containing up to 8M tweets from September 2019 to February 2021. To analyze the impact of the pandemic, the variability of the content over time was studied through quantitative and qualitative analysis of words and emojis. For the quantitative analysis, the shift at the frequency of the terms was calculated using linear regression over frequencies. On the other hand, for the qualitative analysis, word embeddings were used to study the changes in the meaning of the most significant words and emojis at different periods of the pandemic. Through this multifaceted approach, we discovered noteworthy alterations in the political inclinations exhibited by Basque users throughout the course of the pandemic.

Keywords: Computational Social Science, Social Networks, Basque language

1. Introduction

In this constantly connected society (Castells, 2011), we are not exempt from the effects that remote communities generate in ours. Globalized problems such as climate change, nuclear accidents, pollution, war, refugees, and even pandemics, are becoming more frequent and widespread. These global challenges often transcend traditional boundaries of protection, leaving us in a state of uncertainty (Beck et al., 1992). Furthermore, there is an observable shift towards individualism as public institutions recede, thereby integrating us into a more globalized society (Bauman, 2013). The COVID-19 pandemic serves as an example of these trends. Therefore, we highlight the importance of conducting social research to understand the multifaceted impacts of such global incidents on specific communities.

Analysing the changes generated by the COVID-19 crisis has become a topic of main interest for many researchers as it can help in better understanding the new reality brought by the pandemic. Statistical analysis of virus infection levels has been one of the most used methods for modelling the trend of the disease. However, in this work we are focusing on the social change that COVID-19 has entailed. Understanding social changes is not an easy task and specially in a worldwide community where many different realities coexist. Moreover, the infection levels and restrictions taken by governments vary depending on the country, making global analysis misleading and dominated by greater communities. Thus, we focus on the Basque speaking Twitter community as all the users have shared similar restrictions and limitations during the different phases of the pandemic.

In recent years, social networks have become a mirror of society, and their use has greatly increased as a result of proposed health measures to combat the virus (Chakraborty et al., 2020). In addition, the ability to process massive data is greater than ever before due to current advances in hardware (Micikevicius et al., 2018). Along with this, neural network-based techniques have greatly developed the ability to obtain rich representations of words known as word embeddings (Mikolov et al., 2013; Devlin et al., 2019).

Therefore, monitoring public interactions in a social network such as Twitter provides an excellent opportunity to measure society's views on different events. In addition, the importance of social networks is even greater in times of change and they have shown their usefulness in analyzing the social effects of previous phenomena and actions (Buntain et al., 2016; Wang and Zhuang, 2017).

In this work, we want to analyze the response of the Basque speaking Twitter community to the pandemic of COVID-19 through the information provided by this social network, in order to better understand the impact of the pandemic on Basque society. To carry out this study, we have collected and analyzed the tweets posted by the Basque speaking Twitter community from September 2019 to February 2021 using different Natural Language Processing (NLP) techniques. Due to the different stages that the pandemic has experienced in the Basque Country, each one with its different restrictions and COVID-19 infection levels, we have distributed the collected tweets in different groups.

SIGUL2024 Workshop, pages 363–371

This distribution enables us to analyze in much more detail the effect that the different events could have.

The main contributions of this work are the following ones: (1) We have collected and released the biggest dataset of Basque tweets ever, containing up to 8M anonymized tweets text from September 2019 to February 2021. The dataset is split over different pandemic stages enabling fine-grained and overall analysis of terms during period. [1] (2) We conducted an automatic exploration of the most representative terms during the different phases of the pandemic. Due to the combination of quantitative (frequency of use) and qualitative (meaning) analysis of those terms we are able to infer social phenomena from users' textual expressions. [2] (3) We spotted the change that the health crisis generated over people's main concerns. More specifically, we showed that general political issues have lost importance in favor of individual concerns.

2. Related Work

Since the beginning of the COVID-19 pandemic, many articles that monitor the activity of the Twitter social network have been published. Recent work has resulted in the creation of multiple datasets (Banda et al., 2021; López et al., 2020; Alqurashi et al., 2020; Chen et al., 2020a). These datasets typically contain tweets collected during the pandemic months of 2020 and 2021 and they tend to focus on the English language. Gathering English tweets enables us to collect huge datasets as the amount of English tweets is the biggest among all languages. However, as English is a worldwide spoken language, it brings difficulties when analizing social change due to all the different events that affect the English Twitter community. There are also some efforts that focus on smaller communities as the Arabic dataset presented by Alqurashi et al. (2020). All these datasets just extract tweets that contain COVID-19 related keywords as: "SaRS-CoV","COVID-19", "coronavirus"... and even if they are useful for many different tasks (Bullock et al., 2020) they do not offer information for analyzing social alterations caused by the pandemic.

In order to process unstructured text present on social networks, different NLP techniques (Chen et al., 2020b; Shahi et al., 2021) are used. To highlight the different themes treated around COVID-19, Chen et al. (2020b) use the Topic-Modeling technique by applying the LDA algorithm (Blei et al., 2003). The identified topics are visually represented through the UMAP dimension reduction technique (McInnes et al., 2018). In addition, general content analysis has also been performed on minority language scenarios such as Basque, applying Topic-Modeling (Fernandez de Landa et al., 2019) and interaction analysis (Fernandez de Landa and Agerri, 2021). Other studies use supervised techniques to analyze the content of social networks (Chen et al., 2020b; Shahi et al., 2021; Müller et al., 2020), also including Basque language (Agerri et al., 2021). However, in order to be able to train the supervised classification algorithms, previous manual work is needed, that is, an annotation expert must label different examples to be able to apply machine learning algorithms later on.

Analysis of changes in word semantics across time has been previously done by utilizing diachronic word embeddings. These embeddings have been applied for analyzing changes in culture (Hamilton et al., 2016), stereotypes (Garg et al., 2018) and political tendency (Azarbonyad et al., 2017). Similar methods were also used to model meaning change (Del Tredici et al., 2019) and to identify usage change of words across different corpora (Gonen et al., 2020). Closer to our case, Wolfe and Caliskan (2022) and Guo et al. (2021) use word embeddings in order to detect semantic changes in language on tweets related to COVID-19. Other approaches use contextual word representations (Devlin et al., 2019) to analyze the changes on the meaning of words inside specific sentences, instead of focusing on the word itself (Hu et al., 2019; Martinc et al., 2019). All those techniques are similar to ours, however, to the best of our knowledge, we are the first ones to apply this techniques into a controlled community over a specific phenomena such as the COVID-19 pandemic.

3. Data Collection

Twitter has been used as a great data source in order to analyze society and identify the latent dynamics that occur in it. This social network provides massive data for the analysis of small communities such as the Basque speaking one. Similar to any sample trying to represent social reality, ours also has a margin of error. Therefore, sample stratification problems such as age, socio-economic status or culture may occur if we extrapolate the results to the whole Basque society. Although we are able to extract information from the entire research population, our data collection is limited to Twitter users. Consequently, note that the references will center on Basque speaking Twitter community instead of Basque society. Data was gathered on February 2021 using the Twitter API.

[1] The collected data is publicly available here: https://github.com/joseba-fdl/basque_twitter_covid19_corpus

[2] Our code is publicly available here: https://github.com/ikergarcia1996/Ikergazte-Covid-Twitter-2021

As first step Basque speaking users were identified using *umap.eus* tool for Basque language monitoring in Twitter social network. This way, More than 10,000 Basque speaking users have been identified, obtaining 4M personal tweets and 4M retweets for a total of 57M tokens. Different from previous work, we consider all the tweets posted by the 10,000 Basque Twitter users and not just the COVID related ones. This decision is crucial for devising the impact of the pandemic on different aspects of society.

The collected data has been divided into five different periods or stages in order to enable a fine-grained temporal content analysis. As shown in Table 1, each division has been identified with striking moments of the pandemic that have heavily affected the Basque speaking Twitter community. In addition to that, start and end dates of each stage, as well as the distribution of tweets, retweets and word tokens are presented in the same table.

The different groups of the dataset are selected taking the following moments into account:

(0) First, a zero point has been set for the 2019 pre-pandemic era. This stage represents the moment when little or no information was known about the pandemic.

(1) This stage covers the period from the start of 2020 to the lockdown established by the Spanish government. In this period, people started getting infected with COVID-19 in the Basque Country and Spain, but no actions were taken by the authorities.

(2) The second stage consists of the duration of the lockdown order. Lockdown in Spain was defined as the obligation to stay at home, only being able to go out for essential things like buying food. After this moment, wearing a mask was compulsory.

(3) Stage 3 starts after the end of the lockdown era. This period was named as the *New Normality* and restrictions on mobility and social gathering gradually began to be lifted.

(4) Finally, the fourth stage starts when restrictive measures were again introduced due to a new increase of cases. This last stage finishes on February 2021, which was the data extraction date. In this phase important restrictions on social interactions (hospitality, gym, cultural acts...) and curfews were re-enabled in response to the increase of infections. Mobility between towns and cities was also reduced. We have named this stage as the *New Restrictions* period.

The collected data has been anonymized as the only available source is the textual one not keeping any metadata. This way, the authors of the tweets can not be tracked using our dataset, preserving the right to be forgotten. At the same time we keep user anonymity, we release a dataset based on pure text, permitting the reprodictuvity of the results as well as the use of this corpus as an informal Basque language data source.

3.1. Data Analysis

For data analysis purposes we have decided to take personal tweets and retweets into account, as these two elements are part of the content that each user makes public on their timeline. This way, this research is based on both texts of personal tweets and shared tweets (retweets). Apart from the words, that are the main component of the tweets, emojis have also been considered. These increasingly common emojis do not have an unambiguous dictionary definition, but they have their own meaning in certain contexts. Therefore, we study the frequency and meaning of different terms in order to analyze the effects of the pandemic on the Basque speaking Twitter community. We will also show how the use of terms has changed over time, while examining the impact of the pandemic on these changes.

We have carried out both quantitative and qualitative analysis using unsupervised NLP techniques grounded on the distributional hypothesis (Harris, 1954). On the one hand, we study how the frequencies of terms have changed over time, highlighting the terms that have become more and less mentioned as the pandemic has progressed. On the other hand, we have also studied the semantic changes that specific terms have undergone over time, showing the impact that the pandemic has had on the meanings of these terms.

3.1.1. Quantitative Analysis: Fluctuations in the Frequency of Terms over Time

The purpose of the quantitative study is to examine the terms with the greatest fluctuations of usage during the different pandemic stages. The quantitative study is based on the change of the frequency of the terms. We analyze the change of frequency using a linear regression over the frequency of the terms in the different dataset splits. We sort these values by the highest and lowest values to identify the terms with the biggest rise and biggest fall of usage.

First, we lemmatize all the terms using IXA pipes (Agerri et al., 2014) due to the great morphological richness of the Basque language. Suffixes and prefixes are very common and abundant in Basque and the same word can appear in very diverse forms. After lemmatization, as can be seen in Equation 1, we calculate the frequency of each

Stage	From	To	Tweets	Retweets	Word tokens
0. Before 2020	2019/09/01	2019/12/31	224,169	275,042	9M
1. Before lockdown	2020/01/01	2020/03/14	155,302	196,500	6M
2. Lockdown	2020/03/15	2020/06/21	296,627	349,368	12M
3. New normality	2020/06/22	2020/10/24	343,372	362,279	13M
4. New restrictions	2020/10/25	2021/01/31	415,388	347,533	14M

Table 1: Distribution of extracted tweets in Basque over different stages of the pandemic in the Basque Country.

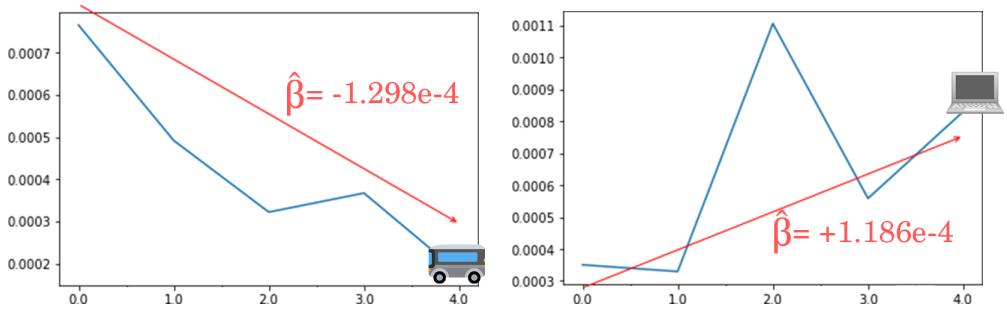


Figure 1: Laptop (💻) and bus (🚌) emoji trend. The Y-axis represents word frequency and the X-axis represents the different stages of the pandemic.

term for each dataset split that corresponds to a different moment of the pandemic. We calculate five different frequencies for each term, one for each dataset split. To calculate the trend of the term, we solve the Equation 2 linear regression system. The values $x_0..x_N$ represent the time splits and the values $y_0..y_N$ represent the frequency of each term in each time split. N is the total number of time splits. We use this linear regression to calculate the slope ($\widehat{\beta}$) of each term , which is an indicator of the trend of that term during the pandemic.

$$y = \frac{\textit{Number of tweets in which the term appears}}{\textit{Number of tweets}} \quad (1)$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \quad (2)$$

A positive slope or trend ($\widehat{\beta}$) means that the term has increased in use during the pandemic while a negative value means that the term usage has decreased. We rank all the terms described in the corpus according to their tendency. The 15 terms with the highest upward trend, and the 15 terms with the highest downward trend can be seen in Table 2. As an example, Figure 1 shows the trends of the laptop (💻) and the bus (🚌) emoji. The usage of the 💻 emoji has increased during the pandemic (especially during times when tougher restrictions were imposed) while the 🚌 emoji usage has decreased.

Terms that have increased in use can be seen in Table 2a, some of which are directly related to the pandemic like health-related terms (*covid, measure, health, pandemic, vaccine, positive, case, care, virus, #covid19*). In addition, we also have terms indirectly related to the pandemic (*online, confinement, hospitality, mask*) corresponding to some side effects such as: the increase in online communication, the reduction in hospitality and opening hours, the use of the mask in everyday life... Finally, the increase in the frequency of the word *crisis* can also be seen as a way to define the situation itself. Thus, most of the terms with the highest positive variability are directly related to pandemic issues, showing the impact of the pandemic on the Basque-speaking Twitter community.

On the other hand, Table 2b shows the terms with the most significant drop in usage. These terms are mainly related to political issues (*strike, feminist, Altsasua, pension, women, Catalonia, demonstration*) and collective initiatives (*presentation, conference, organize, lecture*). Thus, it can be confirmed that there has been a significant decline in the usage of political terms that were previously common on the social network. Feminism (*feminist, women*), economics (*strikes, pensions*) and other political issues (*Catalonia, Altsasua*) have lost their importance in the Basque community as the focus has changed to the pandemic. It also seems that terms related to political action or proclamations have lost their significance. This shows a significant loss of

Term		Trend
covid	*covid*	7.31
neurri	*restriction*	6.82
osasun	*health*	6.17
pandemia	*pandemic*	6.13
txerto	*vaccine*	5.02
positibo	*positive*	3.77
online	*online*	3.44
kasu	*case*	3.20
zaindu	*take care*	3.07
konfinamendu	*confinement*	2.80
birus	*virus*	2.79
krisi	*crisis*	2.78
ostalaritza	*hospitality*	2.75
#covid19	*#covid19*	2.70
maskara	*mask*	2.60

(a) The greatest positive variability.

Term		Trend
aurkezpen	*presentation*	-4.60
greba	*strike*	-4.43
feminista	*feminist*	-4.42
jardunaldi	*conference*	-4.23
Altsasu	*Altsasua*	-4.14
antolatu	*organize*	-3.83
pentsio	*pension*	-3.80
hitzaldi	*lecture*	-3.48
emakume	*women*	-3.22
elkartasun	*solidarity*	-3.20
Katalunia	*Catalonia*	-3.16
areto	*hall*	-3.11
aurkeztu	*presented*	-3.11
egitarau	*program*	-3.09
manifestazio	*demonstration*	-2.79

(b) The greatest negative variability.

Table 2: Variability in term usage over time.

importance of both political theory and practice, especially in Twitter, a social network with strong links to political demands and citizen protests.

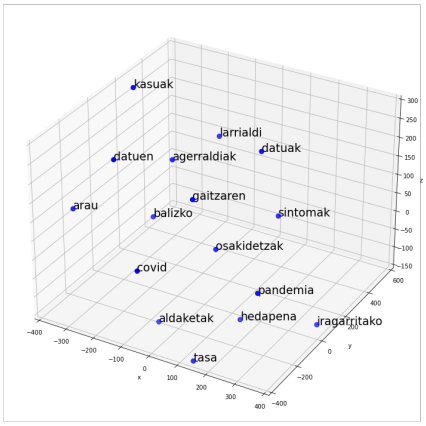
In summary, it is striking that the use of certain politically powerful concepts has decreased, while concepts such as health have gained a central place. Also, some words that have increased in frequency are related to practices that weren't common but have become everyday life, moving from abstraction to close reality. In addition, the frequency of various terms related to the restrictions or measures takes by the government has increased: the need to wear a mask, the permission to stay in bars or maximum number of people that can gather together, the way to communicate at a distance or the order to be locked up at home. It can be said that the focus has shifted to issues related to biopolitics (Foucault, 2009), that is, the regulation of human actions in everyday life. This concept alludes to measures imposed by governments or other power mechanisms that aim at regulating people's lives in their most personal and private facet. Following this reasoning, the presence of this kind of words manifests society's concerns about these restrictions, which seem to be understood as a form of control over their decision-making capacity as individuals. This way, the Basque speaking community in Twitter has shifted from focusing on general issues to focusing more on actions that affect everyday life.

3.1.2. Qualitative Analysis: Fluctuations on the Meaning of Terms over Time

The purpose of the qualitative study is to examine how the change in the meaning of terms has developed across time. Words change their meaning according to the needs of society, adapting their language to specific situations. In order to know which changes happened during the current pandemic we have used word embeddings. These word embeddings have the capability to represent semantics based on the distributional hypothesis (Harris, 1954). In this section, our intention is to generate different word embeddings for each stage and analyze whether the characteristics of terms have changed over time. We use emojis as words during the whole analysis as they are part of the usual vocabulary in social networks.

In order to represent the meaning of words and emojis, we use word2vec (Mikolov et al., 2013) and we obtain dense vector representation. Static word embeddings are used in order to capture the general meaning of the word across time. Thanks to vector representations we can get semantically similar terms, as similar terms have similar representations in the vector space. This way, vectors close to a given term can be used to identify words that are similar, that is, words that have a similar meaning. As words around each term define their meaning, we have computed word embeddings for each time period. For each stage we save the closest words of a given term and check whether there have been any changes between stages.

In order to find out how the meaning of words has changed over time, we have obtained a vector representation of words for 5 different stages. Each of these will combine the semantic features of a stage creating independent representations. To create each dense representation, we use the CBOW method with a 5-token window and 100 dimensions. Thus, we obtain 5 different instances of dense vector representations, placing terms in the corresponding vector space according to the stage and context in which the term was used. An example of the results obtained with this technique



Agerraldiak (*appearances*), aldkaetak (*changes*), arau (*rule*), balizko (*valid*), datuak (*data*), datuen (*of data*), gaitzaren (*of illness*), hedapena (*expansion*), iragarritako (*predicted*), kasuak (*cases*), larrialdi (*emergency*), Osakidetzak (*Osakidetza: Basque public healthcare system*), pandemia (*pandemic*), sintomak (*symptoms*), tasa (*rate*).

Figure 2: Closest words to the term *Covid* during the 3rd stage. Below, translations of the terms can be found.

can be seen in Figure 2, which shows the representation of the word *Covid* and the 16 semantically closest words. In this way, words related and similar to the chosen term are obtained, which will help to define the meaning of the word *Covid* in the 3rd stage.

To perform this qualitative analysis, we selected those terms that have experienced a significant increase in the frequency of use, and that experienced a clearer meaning change: positibo (*positive*), kasu (*case*) and segurtasun (*safety*). The emoji of the mask (😷) has also been chosen for the qualitative study, as it is among the emojis with the highest use frequency variation. Then, to understand each term's connotation, semantically similar words have been obtained using dense word vector representations. Similar words will define the meaning of the selected term. To illustrate how terms' connotations have changed through time, we have selected 5 similar words for each stage, as it can be seen in Table 3.

By analysing the term *positive*, it can be seen that at stages 0 and 1, it is related to many different words (*technique, difficulty, concept, h5n8, reason*..). At stages 2, 3 and 4, surrounding words have changed to terms such as *infect* and *coronavirus*, highlighting the effect of the pandemic in the meaning. During the pandemic era, this term has been used to define people who have been infected with the disease, being totally correspondent to the meaning of the term at stages 2, 3 and 4.

The term *case* at stages 0 and 1 is related to words like *affair* or *account* and also to words related to time (*moment, time, current*). On the contrary, similar words change at stages 2 to 4 showing again relations with the pandemic (*coronavirus, cases, infected*) are present 2, 3 and 4. In addition, it should be noted that the word *positive* is the closest, probably due to the appearance of the bigram *positive case*. Once again, we show that the term has now a direct relationship with the issues of the pandemic.

At stage 0 *safety* is related to words like *law*, *administration* or*system*, terms related to management. As it progresses, at stage 1 the meaning changes to words related to control (*control, to control, reduction*...) but always related to the pandemic (*coronavirus*). It should be said that from stages 2 to 4 the term has been related to words like *prevention* and *hygiene*, closely related to self-control, again showing a close relationship with the concept of biopolitics previously mentioned. In this case, the term has more relation to the regulation of daily life actions than to health status, showing a direct relationship with the impact of the pandemic on everyday life.

Regarding 😷 emoji, at stages 0 and 1, this emoji appears associated with terms related to environmental pollution (*#pollution, filter, chimney, spill, fog*...). As we move forward in time, the meaning changes again in stages 2 and 3, as they appear alongside words directly related to the pandemic (*capacity, hydroalcoholic*...) and with the need to wear the mask to avoid disease infection (*avoid, compulsory, #alwaysmask*...). Thus, the meaning of the emoji has also changed, from environmental pollution related topics to the pandemic, once again shifting to issues related to the regulation of everyday life.

Positive, *case*, *safety* and 😷 terms are excellent indicators of the situation, while they are terms directly related to pandemic issues, the changes in meaning are clearly visible. Although one might expect such changes based on common sense, we are able to demonstrate via a qualitative analysis that the previous meanings have been modified in a specific time period. Thus, this methodology is able to show the meaning of the selected term at each stage, giving the capacity to detect the moment and matter of the modification. The analysis has shown that the changes in meaning over time are closely linked to the pandemic. Those changes in the way Basque speaking Twitter users express themselves can be a sign of meaningful alterations. The modification of the written expressions is a way to show significant variations of the popular imagination of Basque users generated by the pandemic. Specifically in the terms *safety* and 😷, the changes in meaning are again closely linked to biopolitics, as they focus on concepts related to regulation of

Term		Related words on each stage
positibo (*positive*)	0.	teknika (*technique*), zailtasun (*difficulty*), kontzeptu (*concept*), ikusmen (*vision*), gertakizun (*event*)
	1.	h5n8, arrazoia (*reason*), egoiliarri (*resident*), aktiboko (*active*), ontzat (*okay*)
	2.	kutsatu (*infect*), koronabirus (*coronavirus*), kasu (*case*), PCR, infektatu (*infect*)
	3.	koronabirus (*coronavirus*), negatibo (*negative*), kutsatu (*infect*), positiboen (*positive*), PCR
	4.	kutsatu (*infected*), koronabirus (*coronavirus*), ospitaleratze (*hospitalization*), biztanleko (*per capita*), atzemandako (*detected*)
kasu (*case*)	0.	afera (*affair*), galdera (*question*), une (*moment*), kontu (*account*), zentzu (*sense*)
	1.	oraingo (*current*), garai (*time*), mota (*type*), legegintzaldi (*legislature*), afera (*affair*)
	2.	positibo (*positive*), koronabirus (*coronavirus*), kasuak (*cases*), PCR, kutsatu (*infect*)
	3.	positibo (*positive*), koronabirus (*coronavirus*), proba (*test*), kutsatu (*infected*), test
	4.	positibo (*positive*), kasuak (*cases*), test, hildako (*dead*), kutsatu (*infected*)
segurtasun (*safety*)	0.	sistemak (*systems*), hondakinen (*waste*), murrizteko (*reduction*), administrazio (*administration*), legearen (*law*)
	1.	prebentzio (*prevention*), kontrol (*control*), murrizteko (*reduction*), koronabirusak (*coronavirus*), kontrolatzeko (*to control*)
	2.	prebentzio (*prevention*), distantzia (*distance*), higiene (*hygiene*), errespetatu (*respect*), beharrezko (*necessary*)
	3.	prebentzio (*prevention*), higiene (*hygiene*), zorrotz (*strict*), neurriekin (*measures*), protokolo (*protocol*)
	4.	prebentzio (*prevention*), higiene (*hygiene*), malgutu (*adjust*), ezarritako (*established*), mugikortasun (*movility*)
😷	0.	#kutsadura (*#pollution*), albistegitan (*in the news*), #nipenanigloria (*#neitherpitynorglory*), #bizitzaerdigunera (*#lifeinthecenter*), Margaret
	1.	isurketa (*spill*), filtro (*filter*), argindar (*electricity*), tximinia (*chimney*), laino (*fog*)
	2.	saihesteko (*avoid*), besteekiko (*others*), musukoa (*mask*), maskara (*mask*), derrigorrezkoa (*compulsory*)
	3.	#maskarabeti (*#alwayswearmask*), aforo (*capacity*), #euskotrenmetrobilbao (*#train&underground*), edukiera (*capacity*), hidroalkoholikoa (*hydroalcoholic*)
	4.	bidalketa (*submission*), #htxonline, #getxo, #udalsarea2030, #amasavillabona

Table 3: Selected terms and related words over time.

everyday life (*control, to control, reduction, prevention, hygiene, avoid, compulsory, #alwaysmask*...).

4. Conclusions

This work examines the impacts of the COVID-19 pandemic on the Basque-speaking Twitter community, identifying significant changes in the ways of expression reflected in the textual data. The results generated may not fully represent the social reality, since the analyzed sample, despite being a large sample, is conditioned to the use of Twitter social network. While the results are not totally transferable from our selected sample into the entire Basque society, it can be said that they show some symptoms that affect many sectors of the general public.

With the intention of uncovering those variations, we carried out a massive collection of the available data from each of the Basque speaking community users that we identified. Our dataset generation strategy involved data collection and curation of tweets in the Basque language, resulting in the creation of the largest datasets in this minority language. This resource not only facilitates further research but also serves to amplify the visibility of the Basque language within the academic community.

Employing unsupervised Natural Language Processing (NLP) techniques allowed us to uncover significant transformations in language usage. Through a combination of quantitative analysis, tracking term frequency variations over time, and qualitative examination, utilizing dense word vectors to elucidate shifting word and emoji meanings, we are able to detect linguistic variations.

Fluctuations in word usage frequency and semantic meanings underscore the influence of the pandemic, showing how certain terms and symbols have significantly evolved. Moreover, the shift from discussions centered on general political matters to a focus on individual freedoms reflects a broader societal adaptation towards personal concerns, away from traditional political discourse. Nevertheless, these phenomena may be temporary, specific to the circumstances of the pandemic. Investigating

the long-term effects of these occurrences presents an interesting avenue for future research.

5. Limitations

Our research is constrained by its use of static word embeddings and frequency variations. While we acknowledge the existence of more sophisticated algorithms for learning unsupervised word representations, our technique demonstrates the capability to detect changes in word usage reflective of broader social shifts. The simplicity of our approach enables easy replication of experiments across various languages and contexts.

One limitation is our focus solely on a single small language and community. Although this choice facilitated analysis within a geographically confined community, our findings would hold greater significance if conducted across multiple small global communities.

In any case, the results that we have shown were reached due to our selected techniques, as evidenced by the linguistic shift observed among Basque users influenced by the pandemic.

6. Acknowledgments

This work has been partially supported by several MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and by FEDER, EU; (ii) Disargue (TED2021-130810B-C21) and European Union NextGeneration EU/PRTR; (iii) AWARE (TED2021-131617B-I00) and European Union NextGeneration EU/PRTR. (iv) DeepR3 (TED2021-130295B-C31) and European Union NextGeneration EU/PRTR. This work has also been partially funded by the LUMINOUS project (HORIZON- CL4-2023-HUMAN-01-21-101135724).

7. Bibliographical References

Rodrigo Agerri, Josu Bermudez, and German Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, volume 2014, pages 3823–3828.

Rodrigo Agerri, Roberto Centeno, María Espinosa, Joseba Fernandez de Landa, and Álvaro Rodrigo. 2021. Vaxxstance@iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*.

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1509–1518, New York, NY, USA. Association for Computing Machinery.

Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324.

Zygmunt Bauman. 2013. *Liquid modernity*. John Wiley & Sons.

Ulrich Beck, Scott Lash, and Brian Wynne. 1992. *Risk society: Towards a new modernity*, volume 17. sage.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Joseph Bullock, Alexandra Luccioni, Katherine Hoffman Pham, Cynthia Sin Nga Lam, and Miguel Luengo-Oroz. 2020. Mapping the landscape of artificial intelligence applications against covid-19. *Journal of Artificial Intelligence Research*, 69:807–845.

Cody Buntain, Jennifer Golbeck, Brooke Liu, and Gary LaFree. 2016. Evaluating public response to the Boston Marathon bombing and other acts of terrorism through Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10.

Manuel Castells. 2011. *The rise of the network society*, volume 12. John wiley & sons.

Tanusree Chakraborty, Anup Kumar, Parijat Upadhyay, and Yogesh K Dwivedi. 2020. Link between social distancing, cognitive dissonance, and social networking site usage intensity: a country-level study during the COVID-19 outbreak. *Internet Research*.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020a. COVID-19: the first public coronavirus twitter dataset. *CoRR*, abs/2003.07372.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020b. Tracking social media discourse about

the COVID-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Joseba Fernandez de Landa and Rodrigo Agerri. 2021. Social analysis of young Basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, 0(0):1–15.

Joseba Fernandez de Landa, Rodrigo Agerri, and Iñaki Alegria. 2019. Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, 10(6):212.

Michel Foucault. 2009. *Nacimiento de la biopolítica: curso del Collège de France (1978-1979)*, volume 283. Ediciones Akal.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Yanzhu Guo, Christos Xypolopoulos, and Michalis Vazirgiannis. 2021. How covid-19 is changing our language: Detecting semantic shift in twitter word embeddings. *arXiv preprint arXiv:2102.07836*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Christian E. López, Malolan Vasu, and Caleb Gallemore. 2020. Understanding the perception of COVID-19 policies by mining a multilanguage twitter dataset. *CoRR*, abs/2003.10359.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2019. Leveraging contextual embeddings for detecting diachronic semantic shift. *arXiv preprint arXiv:1912.01072*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2018. Mixed Precision Training. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of COVID-19 misinformation on Twitter. *Online social networks and media*, page 100104.

Bairong Wang and Jun Zhuang. 2017. Crisis information distribution on Twitter: a content analysis of tweets during Hurricane Sandy. *Natural hazards*, 89(1):161–181.

Robert Wolfe and Aylin Caliskan. 2022. Detecting emerging associations and behaviors with regional and diachronic word embeddings. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 91–98.

A.7 Fernandez de Landa and Agerri (2024)

Political Leaning Inference through Plurinational Scenarios

Joseba Fernandez de Landa,[1] **Rodrigo Agerri**[1]
[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
joseba.fernandezdelanda@ehu.eus

Abstract: Social media users express their political preferences via interaction with other users, by spontaneous declarations or by participation in communities within the network. This makes a social network such as Twitter a valuable data source to study computational science approaches to political learning inference. In this work we focus on three diverse regions in Spain (Basque Country, Catalonia and Galicia) to explore various methods for multi-party categorization, required to analyze evolving and complex political landscapes, and compare it with binary left-right approaches. We use a two-step method involving unsupervised user representations obtained from the retweets and their subsequent use for political leaning detection. Comprehensive experimentation on a newly collected and curated dataset comprising labeled users and their interactions demonstrate the effectiveness of using Relational Embeddings as representation method for political ideology detection in both binary and multi-party frameworks, even with limited training data. Finally, data visualization illustrates the ability of the Relational Embeddings to capture intricate intra-group and inter-group political affinities.
Keywords: Political analysis, Computational Social Science, Social Media

1 Introduction

Social media users spontaneously engage in political activities, sharing their generated content and interacting with other users from similar or different ideology. As such, social media has become a valuable data source for researchers to study political communities (Conover et al., 2011b; Barberá, 2015; Garimella and Weber, 2017), understand partisanship (Leong et al., 2020), polarization (Morales, Monti, and Starnini, 2021), and other related political phenomena.

Many computational science studies of political behaviour in social media have assumed that it is feasible to reliably infer political leanings by automatic methods. However, most of the research on inferring political leaning has predominantly focused on binary orientations such as left-right (Conover et al., 2011b; Barberá and Rivero, 2015; Barberá, 2015; Garimella and Weber, 2017; Conover et al., 2011a), liberal-conservative (Barberá et al., 2015; Preotiuc-Pietro et al., 2017) and even democrat-republican (Pennacchiotti and Popescu, 2011; Hua, Ristenpart, and Naaman, 2020; Xiao et al., 2020). In fact, the limited number of works that considered a wider leaning spectrum are limited to a specific region (Boutet, Kim, and Yoneki, 2012; Makazhanov and Rafiei, 2013; Rashed et al., 2021). This binary and region-specific outlook represents the political landscape as a uniform and static entity. In reality, every social context has its unique political landscape, characterized by more than just two ideological choices, which evolve in response to societal requirements.

In other words, political leaning inference remains a challenging problem, specially for social media users which are not actively engaged in political activities or that very infrequently express their political preferences. Furthermore, traditional studies on binary orientations are not able to capture the wide spectrum of political ideology and nuances in traditional two-party political systems which are evolving into scenarios where new political actors are emerging (Lisi, 2018) as a response to new ideological conflicts (Ford and Jennings, 2020), socio-economic consequences (Kotroyannos, Tzagkarakis, and Pappas, 2018; Morlino and Raniolo, 2017) or

by suggesting innovative political proposals and novel approaches (Rama, Cordero, and Zagórski, 2021).

Furthermore, transnational integration is leading to the emergence of plurinational states, where the presence of a singular national sentiment is not readily apparent (Keating, 2001), and diverse political leanings representing distinct national sentiments are arising (McGann, Dellepiane-Avellaneda, and Bartle, 2019). Consequently, each political region develops its own political parties tailored to suit specific socio-political circumstances with the aim of obtaining support and sympathy among the population. Hence, characterizing political leaning in terms of proximity to a specific political party would offer a more accurate representation of political nuances in contrast to oversimplifying frameworks such as left-right or conservative-liberal. Moreover, a dynamic approach is essential for tailoring political leaning inference to specific times and regions, specially on complex political scenarios characterized by numerous and evolving political options. We believe that this paradigm could enrich opinion polls, political polarization studies, stance detection and also achieve more accurate social and political research.

To effectively deal with the complexities of real-world politics, our investigation delves into the multi-party framework outlined by Fernandez de Landa, Zubiaga, and Agerri (2023), where the main challenge is to establish effective and robust automatic user representation methods able to capture complex socio-political information for a number of political leaning inference tasks, including left-right or multi-party political leaning, among others. In this work users are represented via their retweet interactions (Conover et al., 2011a; Magdy et al., 2016; Darwish et al., 2020; Stefanov et al., 2020; Fernandez de Landa and Agerri, 2022).

More specifically, a number of unsupervised techniques to generate user representations are explored: Relational Embeddings (Fernandez de Landa and Agerri, 2022), ForceAtlas2 (Jacomy et al., 2014), DeepWalk (Perozzi, Al-Rfou, and Skiena, 2014) and Node2vec (Grover and Leskovec, 2016). In this paper these language independent user representation techniques are evaluated for their effectiveness in inferring binary and multi-party political leanings within three different politically complex regions in Spain, namely, Basque Country, Catalonia and Galicia.

To this end, this paper makes the following contributions on political leaning inference: (i) a novel publicly available dataset containing labeled users by political party and left-right orientation alongside their retweets from the regions of Basque Country, Catalonia and Galicia; (ii) comprehensive experimentation with multi-party (7 political parties) and binary (left-right) frameworks for the three aforementioned political regions, showing that Relational Embeddings (Fernandez de Landa and Agerri, 2022) outperform other user representation methods, especially in few-shot evaluation settings; (iii) an error analysis and data visualization illustrate the potential of our method to infer political ideology, even in dynamic and politically complex scenarios; (iv) data and code will be available upon publication.

2 Related work

Social media is presented as a source to extract data that can then be used to represent users. One of the most common approaches to transform social media interactions into these users representations is based on the force-directed algorithm (Fruchterman and Reingold, 1991). The force-directed algorithm has been used to create unsupervised user representations and subsequently perform stance detection on Twitter users (Darwish et al., 2020). ForceAtlas2 (Jacomy et al., 2014), a forced-directed algorithm, has also been widely used to represent latent user structures. For instance, to study political misinformation in the 2018 presidential election in Brazil, based on opposed hashtags (Soares and Recuero, 2021), or to identify the roles that users play in political conversations during polarized online discussions (Recuero, Zago, and Soares, 2019). Similar techniques were used to map Persian Twitter during Iran's 2017 presidential election by investigating the network structures generated by the users and their sharing practices (Kermani and Adham, 2021). Furthermore, the ForceAtlas2 method has also been used to represent latent user structures through interactions to analyze affinities towards independence movements (Zubiaga et al., 2019), young Basque communities (Fernandez de Landa, Agerri, and Alegria,

2019) or left-right alignments (Conover et al., 2011a). These methods convert large interaction matrices into two-dimensional features, significantly reducing sparsity and memory usage, albeit at the cost of losing some information.

Neural approaches for unsupervised user or node representation based on connections include DeepWalk (Perozzi, Al-Rfou, and Skiena, 2014) and node2vec (Grover and Leskovec, 2016) as the most popular and effective methods (Jusup et al., 2022; Ma et al., 2021), also to represent users via social media interactions (Fernandez de Landa and Agerri, 2021; Fernandez de Landa and Agerri, 2022; Alkhalifa and Zubiaga, 2020). The node-representation features are generated from unlabeled data, representing nodes as low-dimensional features, and they try to predict a group of neighboring users that emerge from Random Walks, based on the input of a given user (Mikolov et al., 2013). More recently, Relational Embeddings have been proposed as an alternative technique to generate user representations based on real user-to-user interaction pairs (Fernandez de Landa and Agerri, 2022).

3 Datasets

We build a dataset to classify multiparty political ideology in three different regions of Spain, namely, Basque Country, Catalonia and Galicia. First, we select the most relevant political parties for each of them. Subsequently, we extract data from Twitter, manually labeling users and collecting the interaction data required to build the user representations.

3.1 Political Regions

Given our interest in analyzing complex political contexts, we have chosen the Kingdom of Spain on 2020 summer as our case of study. Spain is a complex political context, characterized not only by its status as a plurinational country (Keating, 2001) but also because of the emergence of new political actors in the last years (Rama, Cordero, and Zagórski, 2021). Thus, new political actors (UP, Cs, and Vox) burst into Spanish political scenario ruled by traditional forces (PSOE and PP), suggesting updated political proposals and novel approaches (Rama, Cordero, and Zagórski, 2021). Moreover, our study focuses on the Basque Country, Catalonia and Galicia, which are considered stateless nations within the multinational state of Spain (Keating, 2001). These regions house a greater number of political parties than other areas, each with its own regional parliaments and distinct nationalist orientations that drive a wide range of political choices. For each of the selected regions we have picked the political parties that have (or potentially may have) political representation on the regional parliaments. In order to compare multi-party to a binary framework, we also annotate each of the political parties with left-right labels.

Basque Country (EUS): *Basque Nationalist Party* (Partido Nacionalista Vasco - PNV ●) Basque nationalist and Christian-democratic political party; *Unite* (EH Bildu ●) left-wing Basque pro-independence coalition; *Socialist Party* (Partido Socialista Obrero Español - PSOE ●) social-democratic Spanish political party; *Together We Can* (Unidos Podemos - UP ●) democratic socialist Spanish electoral alliance; *People's Party* (Partido Popular - PP ●) conservative and Christian-democratic Spanish political party; *Citizens* (Ciudadanos - Cs ●) liberal Spanish political party; *Vox* (Vox ●) conservative Spanish political party.

Galicia (GAL): *Galician Nationalist Bloc* (Bloque Nacionalista Galego - BNG ●) left-wing Galician nationalist coalition; *Galicianist Tide* (Marea Galeguista - MG ●) left-wing Galician electoral alliance. Despite some punctual differences for PSOE, UP, PP, Cs and Vox are the same as the Basque Country above, whereas representatives and party accounts collected are specific for this region.

Catalonia (CAT): *Republican Left of Catalonia* (Esquerra Republicana de Catalunya - ERC ●) social-democratic Catalan pro-independence political party; *Together for Catalonia* (Junts per Catalunya - JxC ●) progressives Catalan pro-independence political party; *Popular Unity Candidacy* (Candidatura d'Unitat Popular - CUP ●) left-wing Catalan pro-independence political party. PSOE, UP, PP and Cs are the same as for Basque Country and Galicia, although the collected representatives and party accounts are specific for Catalonia. Note that at the time of data collection the Vox party did not have yet representation.

Binary framework: In addition to the political party identification, we have incor-

porated a binary categorization to determine each party's left-right alignment. This was done to facilitate a comparison between the proposed multi-party framework, grounded in political parties, and a more straightforward binary framework. To categorize each political party, we utilized data from opinion polls (CIS, 2019; CIS, 2020a; CIS, 2020b), which gauge public perceptions of the left-right orientations of these parties.

3.2 Data collection strategy

Our data collection strategy consists of two steps, each applied to the selected regions. The initial step involves labeling users for supervised learning, while the second step focuses on gathering interaction data that will be use as input data for user representations. (1) **Manual labeling:** Our starting point relies on a user's seed list that consists of users related to the selected political parties of each region. Thus, we are selecting a sample of users in order to collect data, following the same technique as done in other works (Makazhanov and Rafiei, 2013; Barberá, 2015; Garimella and Weber, 2017; Hua, Ristenpart, and Naaman, 2020). The selected users are related to political parties of each region, such as the political organizations, elected members, candidates or even political militants. The identified user's are labeled by its political party, forming our labeled sample. Apart for the party level categorization for the multi-party framework, we add a binary categorization as left (L) or right (R) leaning for the binary framework (see Table 1). In the following steps, this user list is going to be used to extract the interactions to build user features for the labeled users.

EUS		GAL		CAT	
party	n	party	n	party	n
PNV (R)	146	BNG (L)	39	ERC (L)	18
Bildu (L)	134	MG (L)	7	JxC (R)	18
UP (L)	177	UP (L)	48	UP (L)	16
PSOE (L)	157	PSOE (L)	35	PSOE (L)	18
PP (R)	132	PP (R)	45	PP (R)	14
Cs (R)	40	Cs (R)	12	Cs (R)	14
Vox (R)	8	Vox (R)	7	CUP (L)	11
TOTAL:	794	TOTAL:	193	TOTAL:	109

Table 1: Labeled users for each region. Columns correspond to political party and number of users (n) per region. The labels between parenthesis correspond to left (L) or right (R) leaning for the binary framework.

(2) **Twitter data retrieval:** For every labeled user we first retrieve a history of their retweet interactions. The purpose of this initial retrieval is not to gather the retweets themselves but to identify all users who have interacted with the labeled users through retweets. Afterwards, we gathered all retweets accessible from the timelines the users, extracting a substantial number of interactions involving the sharing of content among users. The data obtained from the timeline extraction ensures that we have enough quantity to later represent the labeled users as well as the users interacting with them (Fernandez de Landa and Agerri, 2022). Table 2 shows the number of retweets retrieved from the labeled and interacting users during the summer of 2020.

	EUS	GAL	CAT
Labeled users	794	193	109
Interacting users	155K	50K	144K
Retweets	58M	13M	41M

Table 2: Final dataset composition for each region.

4 Method

The general idea consists of building user representation features able to capture sociopolitical information leveraging social media data, emulating the same methodology presented in Fernandez de Landa, Zubiaga, and Agerri (2023) and extending it to complex scenarios. These representations will be exclusively built from retweets, without any textual data, as research has shown the effectiveness of retweet-based interactions in user classification tasks (Conover et al., 2011a; Magdy et al., 2016; Darwish et al., 2020; Stefanov et al., 2020; Fernandez de Landa and Agerri, 2022; Cignarella et al., 2020; Agerri et al., 2021). In order to compress that information from sparse interaction matrices into dense user representations, the selected unsupervised training strategies act like feature extraction methods, representing each user in a dense vector space. These methods are used to effectively transform a large number of content sharing actions into meaningful sociopolitical user representations. Those features will be subsequently employed alongside different classification algorithms to evaluate their performance for inferring the political leaning of the labeled users.

4.1 User Representations

We have chosen four different unsupervised approaches to extract user features from interactions. Instead of leveraging huge sparse adjacency matrices, user interactions are brought into a low dimensional vector space, based on approximation-repulsion (Fruchterman and Reingold, 1991) or context embeddings (Mikolov et al., 2013). We employ the following methods for comparison:

- **ForceAtlas2** (Jacomy et al., 2014) (FA2) operates repulsing unconnected nodes and attracting connected ones in order to generate a 2 dimensional graph.
- **DeepWalk** (Perozzi, Al-Rfou, and Skiena, 2014) (DW) simulates uniform random walks among connected instances from a network to learn feature representations.
- **Node2vec** (Grover and Leskovec, 2016) (N2V) algorithm is similar to DeepWalk, but they add control parameters to control the structure of the generated graph.
- **Relational Embeddings** (Fernandez de Landa and Agerri, 2022) (RE) are based on a single hidden-layer neural network trying to predict who retweeted whom for all the gathered interaction pairs. Instead of generating random walks among nearest neighbours, in this method the interactions are based on the relations between two users.

4.2 Methodology

Separate user representations are generated with the four methods listed above, exclusively utilizing retweet data and excluding any labels. Independent models are trained for EUS, GAL and CAT regions without data mixing across them, with the intention of keeping the idiosyncrasies of each region unaltered. All the available retweet interactions per region are employed without any filtering for every method. Feature dimensions are set on 20 dims for DeepWalk, node2vec and Relational Embeddings, with the intention of generating low dimensional but meaningful features (Darwish et al., 2020; Stefanov et al., 2020; Fernandez de Landa and Agerri, 2022). The parameter settings used for node2vec and DeepWalk are the default values typically used by these algorithms: walks_per_node = 10, walk_length = 80, window or context_size = 10, and the optimization is run for a single epoch (Perozzi, Al-Rfou, and Skiena, 2014; Grover and Leskovec, 2016; Fernandez de Landa and Agerri, 2022). Specifically for node2vec, we set p=1 and q=0.5 in order to enhance network community related information (Grover and Leskovec, 2016). Default parameters were set for ForceAtlas2.

5 Experimental setup

In order to identify the political leaning of the labeled users among EUS, GAL and CAT regions, we represent each user with the vectors arisen from the different user representations. To evaluate the performance of these generated user representations, we will conduct two separate sets of experiments with the aim of inferring political leaning. The first experiment will compare the different user representation methods when inferring both the left-right orientation (binary) or political leaning (multi-party) of each user within a strongly supervised scenario. The second experiment will involve a more challenging weakly supervised scenario, where a classifier is provided with very limited training data.

5.1 Strongly supervised scenario

The main goal of this experiment is to compare the performance of various user representation methods when applying the same classifiers. Furthermore, we aim to compare our approach, which defines political leaning within a dynamic multi-party framework, with the conventional approach of treating it as a binary categorization. Thus, we experiment with different user representation methods while inferring binary or multi-class political leanings: (i) *Binary framework*, where only two classes are used to define users' political orientation as left or right. These same categories are used across regions, making it a generalizable and uniform approach. (ii) *Multi-party framework*, where users' political leaning is defined as the closest political party. Political parties vary by region, with seven parties in each region, as specified in Section 3.1.

For each region and framework, we conduct experiments using a leave-one-out (LOO) cross-validation approach. This means that one user is held out for testing while all the remaining users are utilized for training. Therefore, a model is trained and tested individually for every user in the dataset, a feasible task due to the low dimensionality of the representations. The user representations obtained from each of the

methods are used to train six different classification algorithms for each region and framework: Logistic Regression (LogReg), Random Forest (RF), Naive Bayes (NB) and linear, polynomial and RBF-kernel Support Vector Machines (SVM). We use the Scikit-learn implementation (Pedregosa et al., 2011) with default configurations.

5.2 Weakly supervised scenario

Next we experiment in a more challenging, weakly supervised scenario, where the classifiers are provided with limited training data for each region in two different settings: (i) *One-shot*, where only one item per class is selected for training. The selected item for each class is manually selected, being the item representing each of the political parties. (ii) *Three-shot*, a few-shot setting where three items per class are selected for training. In this occasion, for each class we will select a single user corresponding to the political party, as well as two users representing the most referential candidates.

The remainder of the users are left for the test set. In this scenario, the inference will be conducted at the political party level within the multi-party framework. For this setting we only use the RE user representations and SVM-linear classifier, which achieved the best results in the multi-party framework for the strongly supervised scenario (94.0 F1 macro score, Table 4). Additionally, we provide the results obtained both with and without employing dimensionality reduction techniques. The dimensions have been reduced to 2, in accordance with prior research (Darwish et al., 2020; Stefanov et al., 2020), while the remaining hyperparameters are set to their default values. Three different dimensionality reduction techniques are used for this purpose, PCA, t-SNE (Van der Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018).

6 Evaluation

We evaluate the performance of the generated user representations with the aim of inferring political leaning among EUS, GAL and CAT regions in two distinct scenarios, which are defined on the amount of data used for training: strongly and weakly supervised scenarios.

6.1 Strongly supervised scenario

We compare the performance of the diverse user representations combined with different classifiers in a strongly supervised scenario (leave-one-out cross-validation). Besides, results are also compared between binary (Table 3) and multi-party (Table 4) frameworks, empirically showing the challenges that involves shifting from binary to multi-class inference.

Binary framework. Looking at the results reported in Table 3, we notice that each user representation model effectively captures political orientation through the left-right categorization, yielding high-performance results that depend on the employed classifier. However, it is evident that models trained using RE representations consistently demonstrate superior performance and stability in all regions, regardless of the classifier used.

Multi-party framework. When analyzing the results presented in Table 4, it is evident that models trained using RE representations consistently demonstrate superior performance and stability across all regions. Among the models obtained with RE representations, the SVM classifiers obtains the best results. However, despite its popularity, the FA2 representations yield the lowest performance scores, indicating that they are the least suitable for this task. Both N2V and DW outperform FA2, but they are still surpassed by the models generated using RE representations.

As previously mentioned, FA2, DW and N2V can effectively capture information related to left-right orientation while they struggle when dealing with multi-class classifications. The notably superior results achieved within the binary framework compared to the multi-party framework illustrate that bipolar approaches to define political leaning lead to higher overall accuracy at the cost of essential nuances required to comprehend the specific political and social context. On the contrary, RE stands out as the most effective method for capturing finer-grained information related to more specific party-based political leanings, achieving high performance results in both the challenging multi-party context and the binary framework.

	EUS				GAL				CAT				average			
	N2V	DW	FA2	RE	N2V	DW	FA2	RE	N2V	DW	FA2	RE	N2V	DW	FA2	RE
LogReg	83.7	87.2	76.3	**96.0**	87.1	91.4	97.7	**98.2**	97.2	**99.1**	69.5	98.1	89.3	92.6	81.2	**97.4**
RF	82.7	87.3	83.0	**96.5**	91.9	97.1	**98.8**	98.2	96.2	**98.1**	89.7	96.2	90.3	94.2	90.5	**97.0**
NB	49.9	50.3	74.5	**93.0**	59.0	60.1	97.6	**98.2**	69.3	63.6	74.6	**98.1**	59.4	58.0	82.2	**96.4**
SVM - lin.	80.5	85.5	77.2	**95.4**	85.0	89.4	**98.3**	98.2	96.2	**98.1**	73.7	97.2	87.2	91.0	83.1	**96.9**
SVM - pol.	57.4	59.7	82.9	**93.9**	41.7	43.3	**98.8**	96.4	88.0	87.1	74.6	**98.1**	62.4	63.4	85.4	**96.1**
SVM - rbf	69.5	73.9	83.5	**95.1**	64.4	76.1	**98.3**	98.2	84.0	86.8	74.6	**98.1**	72.6	78.9	85.5	**97.1**
average	70.6	74.0	79.6	**95.0**	71.5	76.2	**98.3**	97.9	88.5	88.8	76.1	**97.6**	76.9	79.7	84.6	**96.8**

Table 3: BINARY FRAMEWORK (left-right). F1 macro score results for strongly supervised scenario (LOO CV) on EUS, GAL and CAT datasets. Values in **bold** represent best results for each classifier, while underlined values represent best overall results for each dataset.

	EUS				GAL				CAT				average			
	N2V	DW	FA2	RE	N2V	DW	FA2	RE	N2V	DW	FA2	RE	N2V	DW	FA2	RE
LogReg	63.7	66.1	33.3	**94.0**	42.2	56.4	46.6	**92.6**	86.7	88.0	33.6	**95.1**	64.2	70.2	37.8	**93.9**
RF	59.8	70.1	50.6	**92.8**	62.2	71.3	63.2	**90.1**	77.2	80.3	47.8	**91.6**	66.4	73.9	53.9	**91.5**
NB	38.5	41.5	42.6	**86.7**	46.4	46.8	50.3	**87.9**	39.5	53.9	63.1	**91.8**	41.5	47.4	52.0	**88.8**
SVM - lin.	61.3	64.6	33.8	**93.1**	38.6	53.9	47.0	**92.8**	82.5	86.4	21.0	**96.2**	60.8	68.3	33.9	**94.0**
SVM - pol.	36.1	37.1	42.2	**87.7**	06.2	08.4	47.3	**89.9**	59.4	71.1	07.1	**93.7**	33.9	38.9	32.2	**90.4**
SVM - rbf	39.6	42.1	48.6	**92.3**	20.5	28.1	47.4	**91.7**	52.2	59.2	18.7	**94.2**	37.4	43.1	38.2	**92.7**
average	49.8	53.6	41.9	**91.1**	36.0	44.2	50.3	**90.8**	66.3	73.2	31.9	**93.8**	50.7	57.0	41.3	**91.9**

Table 4: MULTI-PARTY FRAMEWORK (7 political parties). F1 macro score results for strongly supervised scenario (LOO CV) on EUS, GAL and CAT datasets. Values in **bold** represent best results for each classifier, while underlined values represent best overall results for each dataset.

	EUS				GAL				CAT			
Dim. Red.	LOO	3-shot	1-shot	*avg*	LOO	3-shot	1-shot	*avg*	LOO	3-shot	1-shot	*avg*
none	**93.1**	*76.5	55.4	66.0	**92.8**	*81.4	***88.5**	85.0	**96.2**	***94.2**	*89.2	91.7
UMAP 2d	89.5	***90.0**	***81.6**	**85.8**	80.2	*83.5	*82.4	83.0	95.4	***94.2**	***95.1**	**94.7**
t-SNE 2d	90.5	*77.7	*72.9	75.3	85.1	***85.5**	*86.8	**86.2**	94.3	*93.3	***95.1**	94.2
PCA 2d	49.9	28.2	26.6	27.4	59.3	45.5	40.1	42.8	79.9	61.8	66.6	64.2

Table 5: F1 macro score results for SVM-linear classifier on multi-party framework fed by RE features on strongly (LOO CV) and weakly (3- and 1-shot) supervised scenarios. *avg* column represents average values for weakly supervised scenario. Values with * represent when the results of REs with 1- or 3-shot training are higher than the best overall result of non RE methods for multi-party framework on strongly supervised scenario: DW with RF for EUS (*70.1*) and GAL (*71.3*); DW with LogReg for CAT (*88.0*).

6.2 Weakly supervised scenario

As a second step, the RE model is evaluated on the weakly supervised scenario to see the performance when considerably reducing training data. Table 5 demonstrates that 2 dimensional representations obtained from REs through t-SNE and UMAP dimension reduction techniques yield superior results compared to representations generated by REs without any dimensionality reduction for both one-shot and three-shot settings. PCA may not be the most appropriate method for dimension reduction in this context, as it is unable to retain information specific to each community and yields inferior results compared to using the full-dimensional representations. In contrast, UMAP and t-SNE reductions outperform full-dimensional representations as they may preserve community related information due to their architecture based on nearest-neighbours.

Furthermore, the results obtained by combining REs with UMAP and t-SNE dimension reduction techniques in the weakly supervised scenario (Table 5) outperform any other model from the strongly supervised scenario (Table 4). Specifically, when REs are combined with UMAP, only one data point is necessary for training (one-shot) to outperform any other model of the strongly supervised scenario in more than 10 points at EUS and GAL and 7 points in CAT. We confirm that compressing RE representations into 2 dimensional features with UMAP or t-SNE is a good strategy to handle situations with very few annotated data, obtaining similar scores to those of RE on the strongly supervised scenario. Moreover, the results demonstrate consistency across all three regions, indicating that RE representations reach competitive and robust results even where only the user belonging to the political party is

annotated.

7 *Discussion*

In this section, we will delve into a deeper analysis of the reported results by conducting an error analysis and generating visualizations of the user representations for the best method.

Error analysis. In order to understand the considerable differences in performance among user representation methods, we conducted a detailed comparison of the best and second best methods. The confusion matrices presented in Figures 4, 5 and 6 report the errors performed by the Logistic Regression classifier using RE and DW user representation models for multi-party framework for the strongly supervised scenario.

In relation to the EUS dataset, the RE (Figure 1 left) model shows classification errors that occur among classes such as Bildu-UP, PP-Cs and PSOE or PNV as UP. These errors take place among parties that are ideologically close, showing that the model fails between simmilar classes. On the other hand, the DW (Figure 1 right) model makes a considerable number of errors across UP, PSOE and PNV, failing to classify users arround these orientations. Moreover, DW generally fails to classify right-wing unionist party members, grouping them all as PP users.

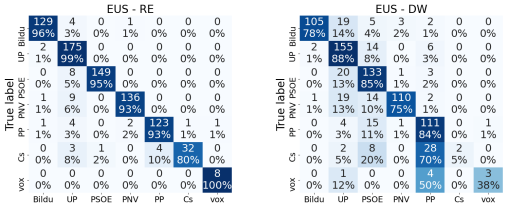


Figure 1: Confusion matrices for Logistic Regresion classifier using RE and DW user representation models in the strongly supervised scenario on the EUS dataset.

When considering GAL dataset, DW (Figure 2 right) misclassifies BNG, MG and PSOE users as UP members, grouping many left-wing representatives under the UP political orientation. Furthermore, some users from Vox and Cs are classified as PP users, being that party the main representative of the right-wing. The DW model seems to underrepresent political options in a left-right dichotomy, simplifying political complexity. On the other hand, the RE (Figure 2 left) model has very few classification errors occurring among ideologically similar orientations, demonstrating the capacity of REs to represent parties as well as political orientations.

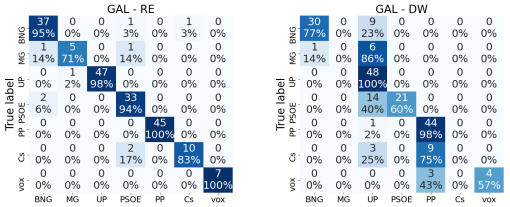


Figure 2: Confusion matrices for Logistic Regresion classifier using RE and DW user representation models in the strongly supervised scenario on the GAL dataset.

Finally, with respect to CAT dataset, we can see that RE (Figure 3 left) gets better results than DW (Figure 3 right). However, both models achieve high performance despite minor deviations. This can be attributed to factors such as the smaller dataset size or balanced classes contributing to their outstanding performance.

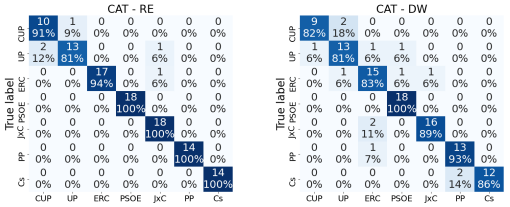


Figure 3: Confusion matrices for Logistic Regresion classifier using RE and DW user representation models in the strongly supervised scenario on the CAT dataset.

Summarizing, classification errors usually take place among parties that are ideologically close, showing that retweet-based DW and RE models can capture general ideological tendencies. This conclusion aligns with the high results obtained by DW for the binary framework, confirming these models' ability to capture general ideological traits. However, the RE models show a high accuracy with a very low error rate, demonstrating that this model can also capture previously specified political parties besides the general ideological alignments.

Data visualization. In order to better understand the effectiveness of the RE user representation techniques, we visualize the RE

representations of labeled users for the three regions, namely, EUS, GAL and CAT, by performing PCA, UMAP and t-SNE dimensionality reduction (as it was done in weakly supervised scenario). This visualization allows us to undertake a qualitative evaluation by correlating common-sense political knowledge to the representations obtained by our representation model.

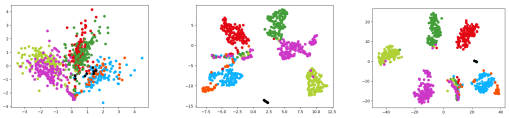


Figure 4: Visualization of PCA, UMAP and t-SNE 2 dimension reduction for EUS Relational Embedding representation.

EUS *(Figure 4):* In the EUS dataset UMAP (Figure 4 center) and t-SNE (Figure 4 right) visualizations have similar groupings with clear clusters for specific political organizations, which are arranged based on their political proximity. Thus, the centered positions of the graph are taken by the parties in the Basque government, formed by PNV (●) and PSOE (●). The leftist positions are represented by UP (●) and Bildu (●), located on one side. Whereas right-winged positions represented by PP (●) and Cs (●), are pictured on the other side, while alt-right Vox (●) is represented as an outlier. PCA visualization (Figure 4 left) is fuzzier but able to group users on 3 different groups corresponding to the previous global views; PNV and PSOE representing centered positions; Bildu and UP representing the left and progressives; PP, Cs and Vox representing the right and the conservatives.

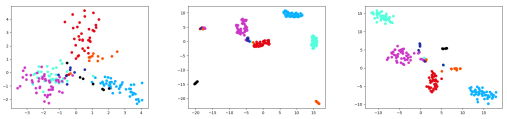


Figure 5: Visualization of PCA, UMAP and t-SNE 2 dimension reduction for GAL Relational Embedding representation.

GAL *(Figure 5):* In this dataset t-SNE visualizations (Figure 5 right) show clear clusters for the political parties, drawing them on a singular axis. On one end we have BNG (●) representing a pro-independence left-wing followed by UP (●) and MG (●) representing the left; PSOE (●) and Cs (●) represent centered positions next to PP (●) representing the right and the conservatives on the oposite end. PCA visualization (Figure 5 left) also groups the users on 3 different groups mixing political parties into a more simple layout; UP, BNG and MG as left and progressives; PSOE and Cs grouped on centered positions; PP (●) and Vox (●) grouped as right and the conservatives.

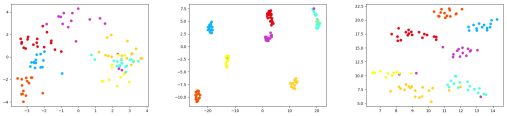


Figure 6: Visualization of PCA, UMAP and t-SNE 2 dimension reduction for CAT Relational Embedding representation.

CAT *(Figure 6):* In the CAT dataset PCA (Figure 6 left) and t-SNE (Figure 6 right) visualizations provide symbolic representations of the political reality, wherein political parties are positioned close to each other based on their stance regarding the independence process. On one side, parties like Cs (●) and PP (●), which strongly advocate for unity with the Spanish kingdom, are clustered in the left (PCA) or top (t-SNE) portions of the plot. Conversely, pro-independence parties such as CUP (●), JxC (●), and ERC (●) are clustered on the right (PCA) or bottom (t-SNE) side. In more central positions, UP (●) and PSOE (●) are situated between both groupings, serving as a connecting links bridging the divide between the two sides.

It seems that the dataset size has an impact on the visualizations, with representations becoming fuzzier as the dataset size increases. Being the biggest dataset, EUS exhibits the largest degree of fuzziness, while CAT is the smallest dataset and has the most defined communities. Additionally, the clarity of the communities depending on the dimensionality reduction technique is in accordance with the results achieved by these dimension reduction techniques on the weakly supervised scenario. On the one hand, the PCA dimension reduction technique does not clearly discriminate the communities related to the specific political parties. However, it can effectively display information related to more general political orientations by clustering ideologically similar parties closely to-

gether in the same euclidean space. On the other hand, in the graphs arisen from UMAP and t-SNE dimension reduction techniques, it can be seen that the communities are clearly defined and situated depending on their ideological similarities. Regardless of the dimension reduction technique or the dataset size, REs can effectively represent the political communities as well as the ideological similarities and disparities among them. These findings suggest that RE user representations have the capacity to embed knowledge about the socio-political environment leaning on retweet based user interactions.

8 Conclusion and Future Work

In this work we have explored the ability to dynamically infer the political leaning of social media users across left-right and multi party-based frameworks which, to the best of our knowledge, has not been studied before. In order to compare binary and multi-party frameworks, we compile a dataset labeling users according to both their affiliation with a political party and their left-right orientation. We collect data from three politically complex areas in Spain, characterized by plurinationality and the emergence of new political actors. To conduct dynamic political leaning inference we propose a two-step approach, starting with different unsupervised user representations through retweets, followed by political party classification. When assessing the performances of user representation methods among different frameworks, it is observed that while performances are high at the binary framework, there is a significant decrease for baseline methods in the multi-party framework. We believe that this illustrates the challenges associated with inferring political leanings in multi-party frameworks.

Nonetheless, results indicate that Relational Embeddings, when combined with any classifier, yield high-performance results across all regions and evaluation scenarios. Thus, even when only a single user belonging to each political party is annotated, REs in combination with SVM-linear classifier outperforms any other baseline model from the strongly supervised multi-party framework. Furthermore, error analysis and visual representations reveal that REs can effectively capture political affinities within and across political leanings. These considerations underscore the adaptability of REs and its potential to capture socio-political information in extremely diverse and dynamic real-world situations.

Considering this, we have a keen interest in applying interaction-based user representation techniques to other tasks such as propaganda or misinformation detection. Additionally, we want to include additional data sources for user representation, such as textual data, in order to extend the user representation idea to social media platforms where retweet interactions are not available.

References

[Agerri et al.2021] Agerri, R., R. Centeno, M. Espinosa, J. Fernandez de Landa, and A. Rodrigo. 2021. VaxxStance@IberLEF 2021: Overview of the Task on Going Beyond Text in Cross-Lingual Stance Detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

[Alkhalifa and Zubiaga2020] Alkhalifa, R. and A. Zubiaga. 2020. QMUL-SDS@SardiStance: Leveraging Network Inter-actions to Boost Performance on Stance Detection using Knowledge Graphs. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR Workshop Proceedings.

[Barberá2015] Barberá, P. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23:76 – 91.

[Barberá et al.2015] Barberá, P., J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542.

[Barberá and Rivero2015] Barberá, P. and G. Rivero. 2015. Understanding the political representativeness of twitter users. *Social Science Computer Review*, 33:712 – 729.

[Boutet, Kim, and Yoneki2012] Boutet, A., H. Kim, and E. Yoneki. 2012. What's in your tweets? i know who you supported in the uk 2010 general election. *Proceedings of the International AAAI Conference on Web and Social Media*, 6:411–414.

[Cignarella et al.2020] Cignarella, A. T., M. Lai, C. Bosco, V. Patti, and P. Rosso. 2020. SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In V. Basile, D. Croce, M. Di Maro, and L. C. Passaro, editors, *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

[CIS2019] CIS. 2019. *Barómetro de diciembre 2019. Postelectoral Elecciones Generales 2019*, volume 3269. Centro de Investigaciones Sociológicas, 11.

[CIS2020a] CIS. 2020a. *Preelectoral de Galicia. Elecciones Autonómicas julio 2020*, volume 3287. Centro de Investigaciones Sociológicas, 06.

[CIS2020b] CIS. 2020b. *Preelectoral del País Vasco. Elecciones Autonómicas julio 2020*, volume 3286. Centro de Investigaciones Sociológicas, 06.

[Conover et al.2011a] Conover, M. D., B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. 2011a. Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 192–199. IEEE.

[Conover et al.2011b] Conover, M. D., J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. 2011b. Political Polarization on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[Darwish et al.2020] Darwish, K., P. Stefanov, M. Aupetit, and P. Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152.

[Fernandez de Landa and Agerri2021] Fernandez de Landa, J. and R. Agerri. 2021. Social analysis of young basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, 0(0):1–15.

[Fernandez de Landa and Agerri2022] Fernandez de Landa, J. and R. Agerri. 2022. Relational embeddings for language independent stance detection. *arXiv e-prints*, pages arXiv–2210.

[Fernandez de Landa, Agerri, and Alegria2019] Fernandez de Landa, J., R. Agerri, and I. Alegria. 2019. Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, 10(6):212.

[Fernandez de Landa, Zubiaga, and Agerri2023] Fernandez de Landa, J., A. Zubiaga, and R. Agerri. 2023. Generalizing political leaning inference to multi-party systems: Insights from the uk political landscape. *ArXiv*, abs/2312.01738.

[Ford and Jennings2020] Ford, R. and W. Jennings. 2020. The changing cleavage politics of western europe. *Annual review of political science*, 23:295–314.

[Fruchterman and Reingold1991] Fruchterman, T. M. J. and E. M. Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21.

[Garimella and Weber2017] Garimella, V. R. K. and I. Weber. 2017. A long-term analysis of polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

[Grover and Leskovec2016] Grover, A. and J. Leskovec. 2016. node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[Hua, Ristenpart, and Naaman2020] Hua, Y., T. Ristenpart, and M. Naaman. 2020. Towards measuring adversarial twitter interactions against candidates in the us midterm elections. In *ICWSM*.

[Jacomy et al.2014] Jacomy, M., T. Venturini, S. Heymann, and M. Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6):e98679.

[Jusup et al.2022] Jusup, M., P. Holme, K. Kanazawa, M. Takayasu, I. Romić, Z. Wang, S. Geček, T. Lipić, B. Podobnik,

L. Wang, W. Luo, T. Klanjšček, J. Fan, S. Boccaletti, and M. Perc. 2022. Social physics. *Physics Reports*, 948:1–148. Social physics.

[Keating2001] Keating, M. 2001. *Plurinational Democracy: Stateless Nations in a Post-Sovereignty Era*. Oxford University Press, 11.

[Kermani and Adham2021] Kermani, H. and M. Adham. 2021. Mapping persian twitter: Networks and mechanism of political communication in iranian 2017 presidential election. *Big Data & Society*, 8(1):20539517211025568.

[Kotroyannos, Tzagkarakis, and Pappas2018] Kotroyannos, D., S. I. Tzagkarakis, and I. Pappas. 2018. South european populism as a consequence of the multidimensional crisis? the cases of syriza, podemos and m5s. *European Quarterly of Political Attitudes and Mentalities*, 7(4):1–18.

[Leong et al.2020] Leong, Y. C., J. Chen, R. Willer, and J. Zaki. 2020. Conservative and liberal attitudes drive polarized neural responses to political content. *Proceedings of the National Academy of Sciences*, 117:27731 – 27739.

[Lisi2018] Lisi, M. 2018. *Party system change, the European crisis and the state of democracy*. Routledge.

[Ma et al.2021] Ma, X., J. Wu, S. Xue, J. Yang, C. Zhou, Q. Z. Sheng, H. Xiong, and L. Akoglu. 2021. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

[Magdy et al.2016] Magdy, W., K. Darwish, N. Abokhodair, A. Rahimi, and T. Baldwin. 2016. #isisisnotislam or #deportallmuslims? predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, page 95–106, New York, NY, USA. Association for Computing Machinery.

[Makazhanov and Rafiei2013] Makazhanov, A. and D. Rafiei. 2013. Predicting political preference of twitter users. *Social Network Analysis and Mining*, 4:1–15.

[McGann, Dellepiane-Avellaneda, and Bartle2019] McGann, A., S. Dellepiane-Avellaneda, and J. Bartle. 2019. Parallel lines? policy mood in a plurinational democracy. *Electoral Studies*, 58:48–57.

[McInnes et al.2018] McInnes, L., J. Healy, N. Saul, and L. Großberger. 2018. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.*, 3:861.

[Mikolov et al.2013] Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Morales, Monti, and Starnini2021] Morales, G. D. F., C. Monti, and M. Starnini. 2021. No echo in the chambers of political interactions on reddit. *Scientific Reports*, 11.

[Morlino and Raniolo2017] Morlino, L. and F. Raniolo. 2017. *The impact of the economic crisis on South European democracies*. Springer.

[Pedregosa et al.2011] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pennacchiotti and Popescu2011] Pennacchiotti, M. and A. M. Popescu. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *KDD*.

[Perozzi, Al-Rfou, and Skiena2014] Perozzi, B., R. Al-Rfou, and S. Skiena. 2014. Deepwalk: Online learning of social representations. KDD '14, page 701–710. Association for Computing Machinery.

[Preotiuc-Pietro et al.2017] Preotiuc-Pietro, D., Y. Liu, D. J. Hopkins, and L. H. Ungar. 2017. Beyond binary labels: Political ideology prediction of twitter users. In *ACL*.

[Rama, Cordero, and Zagórski2021] Rama, J., G. Cordero, and P. Zagórski. 2021. Three is a crowd? podemos, ciudadanos, and vox: The end of bipartisanship in spain. *Frontiers in Political Science*, 3.

[Rashed et al.2021] Rashed, A., M. Kutlu, K. Darwish, T. Elsayed, and C. Bayrak. 2021. Embeddings-based clustering for target specific stances: The case of a polarized turkey. In *ICWSM*.

[Recuero, Zago, and Soares2019] Recuero, R., G. Zago, and F. Soares. 2019. Using social network analysis and social capital to identify user roles on polarized political conversations on twitter. *Social Media+ Society*, 5(2):2056305119848745.

[Soares and Recuero2021] Soares, F. B. and R. Recuero. 2021. Hashtag wars: Political disinformation and discursive struggles on twitter conversations during the 2018 brazilian presidential campaign. *Social Media+ Society*, 7(2):20563051211009073.

[Stefanov et al.2020] Stefanov, P., K. Darwish, A. Atanasov, and P. Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *ACL*.

[Van der Maaten and Hinton2008] Van der Maaten, L. and G. Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[Xiao et al.2020] Xiao, Z., W. Song, H. Xu, Z. Ren, and Y. Sun. 2020. Timme: Twitter ideology-detection via multi-task multi-relational embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 2258–2268, New York, NY, USA. Association for Computing Machinery.

[Zubiaga et al.2019] Zubiaga, A., B. Wang, M. Liakata, and R. Procter. 2019. Political homophily in independence movements: Analyzing and classifying social media users by national identity. *IEEE Intelligent Systems*, 34:34–42.

A.8 Fernandez de Landa *et al.* (2023)

Generalizing Political Leaning Inference to Multi-Party Systems: Insights from the UK Political Landscape

Joseba Fernandez de Landa,[1] **Arkaitz Zubiaga,** [2] **Rodrigo Agerri** [1]

[1] HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[2] Queen Mary University of London
joseba.fernandezdelanda@ehu.eus, a.zubiaga@qmul.ac.uk, rodrigo.agerri@ehu.eus

Abstract

An ability to infer the political leaning of social media users can help in gathering opinion polls thereby leading to a better understanding of public opinion. While there has been a body of research attempting to infer the political leaning of social media users, this has been typically simplified as a binary classification problem (e.g. left vs right) and has been limited to a single location, leading to a dearth of investigation into more complex, multiclass classification and its generalizability to different locations, particularly those with multi-party systems. Our work performs the first such effort by studying political leaning inference in three of the UK's nations (Scotland, Wales and Northern Ireland), each of which has a different political landscape composed of multiple parties. To do so, we collect and release a dataset comprising users labelled by their political leaning as well as interactions with one another. We investigate the ability to predict the political leaning of users by leveraging these interactions in challenging scenarios such as few-shot learning, where training data is scarce, as well as assessing the applicability to users with different levels of political engagement. We show that interactions in the form of retweets between users can be a very powerful feature to enable political leaning inference, leading to consistent and robust results across different regions with multi-party systems. However, we also see that there is room for improvement in predicting the political leaning of users who are less engaged in politics.

Introduction

Ideology is a set of people's beliefs that can be understood as ways of thinking and acting in society. Those beliefs can generally be represented by political parties, acting like social hubs of coordinated thoughts and actions. However, ideology is often presented and simplified into binary frameworks based on individuals stance over left/right or conservative/liberal orientations. Political leaning inference is proposed as a way of representing individual actors by the closest political party, analyzing ideology from a richer perspective. To better understand society, social researchers can benefit from the development of efficient methods capable of generalizing political leaning inference across different regions (Imhoff et al. 2022). We address this challenge by investigating new tools and techniques for conducting deeper and more accurate social and political research with the aim of improving opinion polls, political polarization studies, stance and propaganda detection or disinformation analysis among others.

The vast majority of research on political leaning inference has been limited to binary classification between the two prevailing parties or stances (Conover et al. 2011b,a; Barberá and Rivero 2015; Barberá 2015; Garimella and Weber 2017; Barberá et al. 2015; Pennacchiotti and Popescu 2011; Hua, Ristenpart, and Naaman 2020; Xiao et al. 2020). The few works that conducted multiclass classification (Boutet, Kim, and Yoneki 2012; Makazhanov and Rafiei 2013; Rashed et al. 2021) were constrained to a single scenario or region. This however limits both the applicability of such methods and the insights learned from such studies. Indeed, each social context has its own political reality which is typically reflected in more than two ideological options.

In order to broaden the study on the ability to infer the political leaning of social media users, we identify four key limitations in previous work. First, the widely used binary frameworks can be limiting and imprecise as individuals hold a wider range of political beliefs, which instead calls for multiclass classification. Second, political ideologies and beliefs can vary substantially across different regions since each community has its own idiosyncrasy, which calls for the study of applicability across regions. Third, to achieve generalizability, it is crucial to include scenarios where the labeled data available for training is limited, which makes critical the study of few-shot learning approaches. Fourth, it is important to consider that not every social media user is as engaged in politics and/or is vocal about their beliefs, which posits the importance of assessing the ability to predict the political leaning of all kinds of users regardless of their level of engagement.

By addressing the above limitations, we aim to further research in political leaning inference by providing the first study that focuses on generalizing a multiclass political leaning inference model across different scenarios or regions. With this goal in mind, we propose and evaluate a range of techniques for data extraction and user representation for multiclass political leaning inference. The aim is to independently represent Twitter users by leveraging their interactions, effectively transforming content sharing actions on Twitter into vector spaces. By accomplishing this, we seek to achieve adaptability across diverse situations, in turn opening an avenue for further exploration in other tasks.

Thus, retweet interactions are selected to represent users, known for their effectiveness in achieving user classification (Conover et al. 2011a; Magdy et al. 2016; Darwish et al. 2020; Stefanov et al. 2020; Fernandez de Landa and Agerri 2022). We conducted experiments using four distinct unsupervised techniques, namely ForceAtlas2 (Jacomy et al. 2014), DeepWalk (Perozzi, Al-Rfou, and Skiena 2014), Node2vec (Grover and Leskovec 2016) and Relational Embeddings (Fernandez de Landa and Agerri 2022). Those user representations are evaluated in three different regions, each with different political parties, including also the first study on the ability of those techniques to rely on scarce training data as well as to determine the political leaning of users with different levels of engagement.

To the best of our knowledge, our work is the first to study the political leaning inference task across diverse multiclass political realities, which in turn leads to new insights into tackling this challenging setting. More specifically, this paper makes the following novel contributions:

- we devise and experiment with a pluralistic framework that includes multiple political leanings, which proves adaptable to different regions since it is grounded on localized political actors;
- we propose and evaluate a range of methodologies to make the most of retweet interactions among social media users to infer their political leaning, showing that Relational Embedding based approach is effective even in weakly-supervised and realistic scenarios;
- we perform a comprehensive error analysis and feature visualizations in order to show the ability of the proposed methodology to capture socio-political information coherently and in alignment with the specific political context.
- data resources such as labeled and interaction-based data and code will be made publicly available to facilitate reproducibility of results and enable further research.

Through experiments on three of the United Kingdom's regions (Wales, Scotland and Northern Ireland), we find that the use of interaction-pairs leads to competitive performance on political leaning inference when compared to random-walk based approaches. In addition to improved performance, our analysis shows that the resulting visualizations and errors obtained from this approach can provide valuable insights into the political context in the UK.

Related Work

In this section we discuss prior work on political leaning inference and stance detection across different countries, and different methods to extract features from interaction-based data such as retweets between users.

Political leaning inference

Political leaning of elected representatives has been approached using roll call votes on parliaments of bipartite systems such as US (Akoglu 2014) or even multy-party systems like Brazil (Vaz de Melo 2015). Whereas elected people's political leaning can be inferred based on their public votes, the extension of these studies is limited to the specific parliaments and can not be extended to other populations as can be done with social media data.

Political leaning inference from social media was pioneered by Conover et al. (2011b), who framed it as a left-right dichotomy in the US context by using retweets. Subsequently, others have followed a similar categorization of users into left or right, such as Barberá and Rivero (2015) and Barberá (2015) in the context of elections, Barberá et al. (2015) classifying users as liberal or conservative, or Garimella and Weber (2017) focused on political accounts and media outlets. In addition, Twitter users from the USA have been classified as Democrats or Republicans (Pennacchiotti and Popescu 2011; Hua, Ristenpart, and Naaman 2020) and, although some other work has gone beyond the strictly binary classification by proposing a seven-point scale to place users in a conservative-liberal spectrum (Preotiuc-Pietro et al. 2017), they still adhere to the same dichotomy.

Multiclass classification in more diverse political landscapes, the most common situation for a large number of countries, has barely been studied. This limits the ability to evaluate existing methods in those scenarios as well as the capacity to learn new insights from social media for those specific contexts.

Political ideology can also be approached as, or associated with, the arguably more popular stance detection task (Mohammad et al. 2016; Hardalov et al. 2021). Still, much of the stance detection research has focused on topics relevant to politics but not explicitly on political leaning inference. This is the case, for example, of recent shared tasks and datasets such as SardiStance (Cignarella et al. 2020) and VaxxStance (Agerri et al. 2021), which proposed stance detection tasks associated with the sardines social movement and the stance towards vaccination. The datasets released by these shared tasks allowed researchers to leverage not only textual content but also social interactions (Ferraccioli et al. 2020; Alkhalifa and Zubiaga 2020; Lai et al. 2021; Fernandez de Landa and Agerri 2022), highlighting the importance of interactions in determining the stance of users who form homophilic connections with one another.

Stance detection in social media has also been addressed as an unsupervised task by relying on interactions between users. Previous works following this approach have applied a force-directed algorithm (Fruchterman and Reingold 1991) or other methods such as UMAP (McInnes et al. 2018). For example, Darwish et al. (2020) used both force-directed algorithm and UMAP for unsupervised stance detection on Twitter users using retweets. Furthermore, UMAP has also been used to get interaction-based retweet features for automatically tagging Twitter users' stance on different topics (Stefanov et al. 2020) and to study political polarization in Turkey (Rashed et al. 2021). A shortcoming of these approaches is that, in order to be able to handle the huge interaction networks, the features are based only on a set of users manually picked as being more salient.

Approaches to modeling user interactions

User interactions in social media have been used to study political disinformation in the 2018 presidential election in Brazil, based on opposed hashtags (Soares and Recuero 2021), or to identify the roles that users play in political conversations during polarized online discussions (Recuero, Zago, and Soares 2019). Besides, interaction features have also been useful to map Persian Twitter during Iran's 2017 presidential election by investigating the network structures generated by the users and their sharing practices (Kermani and Adham 2021). Finally, Zubiaga et al. (2019) studied stance detection to analyze independence movements in Catalonia, Basque Country and Scotland, showing that features extracted from the followers network obtained the best performance. Following the same idea, an analysis of young Basque Twitter users were also mapped based on their retweet-based data (Fernandez de Landa and Agerri 2021).

Neural approaches for learning node representations on user interaction data include DeepWalk (Perozzi, Al-Rfou, and Skiena 2014) and node2vec (Grover and Leskovec 2016) among the most popular and effective choices (Jusup et al. 2022; Ma et al. 2021). These node-representation features are created based on unlabeled data, characterizing the nodes as low-dimensional features. These methods try to predict a group of neighboring users that emerge from Random Walks, based on the input of a given user. This means that they focus more on the overall structure of the interaction network instead of trying to model pair-based interactions. An issue to consider is that random walks might create artificial interactions that may not actually occur in the interaction pairs. An alternative proposal to build interaction-based models is the Relational Embeddings method, based on real interaction-pairs applied to predict users' stance (Fernandez de Landa and Agerri 2022).

Finally, other neural approaches such as graph convolutional networks (GCN) (Kipf and Welling 2017) and graph attention networks (GAT) (Velickovic et al. 2018) require previously obtained feature representations to train end-to-end models. Therefore, these neural models require labeled data and additional features during the feature learning process. As a result, these models cannot be effectively employed to extract user representations using an unsupervised approach and are left out of our study. TIMME is a technique developed for identifying Democrat/Republican leaning on Twitter by utilizing multi-task learning and multi-relational data such as follow, retweet, reply, mention and like (Xiao et al. 2020). However, testing these three algorithms was not possible with the hardware we currently have at our disposal (see Table 2 for size of our datasets).

The Political Context in the UK

Given our interest in analyzing the socio-political context of the United Kingdom as a multi-party system (Lynch 2007), our study focuses on political parties in Scotland (5.5M citizens), Wales (3.1M) and Northern Ireland (1.8M). These regions form politically diverse contexts, each with its own devolved government and strong nationalist sentiments that foster many political options. The UK's political landscape has evolved substantially in recent decades, from being dominated by two parties in the 1950s (Conservative and Labour parties attaining over 95% of the votes) to a more diverse, multi-party landscape (75% across both parties in 2019).

Scotland (SCT): *Scottish National Party* (SNP ●) is a Scottish nationalist and social democratic political party; positioned on the center-left, pro-independence and pro-European. *Scottish Conservative & Unionist Party* (SCU ●) is a conservative party in Scotland, Nationally affiliated with the Conservative Party; centre-right and unionist. *Scottish Labour Party* (SL ●) is a Scottish social democratic political party, an autonomous section of the UK Labour Party; considered to be centre-left and unionist. *Scottish Green Party* (SGP ●) is a Scottish green political party, affiliated with the Global Greens and associated mainly with environmentalist policies; positioned on the left, pro-independence and pro-European. *Scottish Liberal Democrats* (SLD ●) is a Scottish liberal and federalist political party, part of the United Kingdom Liberal Democrats; positioned on the political centre, pro-European and unionist.

Wales (WAL): *Welsh Labour* (WL ●) is a Welsh social democratic political party, and formally part of the UK Labour Party; centre-left and unionist. *Welsh Conservatives* (WC ●) is a conservative party in Wales, a branch of the UK's Conservative Party; ideology is centre-right and unionist. *Plaid Cymru* (PC ●) is the principal Welsh nationalist political party; positioned on the left and pro-independence. *Welsh Liberal Democrats* (WLD ●) is a Welsh liberal and federalist political party, branch of the UK's Liberal Democrats; positioned on the political centre, pro-European and unionist.

Northern Ireland (NIR): *Sinn Féin* (SF ●) is an Irish republican and democratic socialist political party; considered to be left-wing, pro-unification and pro-independence. *Democratic Unionist Party* (DUP ●) is a conservative and loyalist political party in Northern Ireland; positioned on the right-wing and unionist. *Alliance Party of Northern Ireland* (APNI ●) is a liberal political party in Northern Ireland, aligned with the UK's Liberal Democrats; positioned on the political centre, pro-European and they consider themselves outside of Nationalism and Unionism. *Ulster Unionist Party* (UUP ●) is a unionist political party in Northern Ireland, a branch of the UK's Conservative Party; positioned on the center-right and unionist. *Social Democratic and Labour Party* (SDLP ●) is a social-democratic and Irish nationalist political party in Northern Ireland; centre-left and pro-Irish.

Datasets

We devise a generalizable data collection methodology, in our case tested in the UK but extensible to other regions. Once the regions of interest and the relevant political parties have been identified, our methodology consists of three steps to collect: (i) an initial seed of users (members), (ii) other users with different levels of engagement or interest (supporters and sympathizers) and, (iii) interactions and timelines pertaining to those users. Data collection was done between September and October 2022.

Step 1. Manual labeling of seed users. In line with data collection strategies followed in previous work (Makazhanov and Rafiei 2013; Barberá 2015; Garimella and Weber 2017; Hua, Ristenpart, and Naaman 2020), we start by collecting an initial seed of users. For each of the political parties in our datasets, we identify party members with Twitter accounts including members of parliament (MPs) or members of regional parlaments (MSPs, MSs and MLAs). This leads to a collection of users where each user is linked to a specific region and party (details in Table 1, column 'Members').

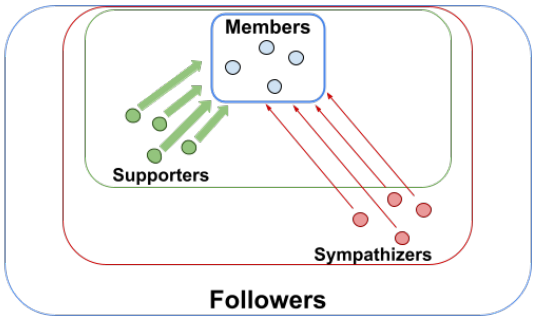


Figure 1: Creation scheme for Supporter and Sympathizer evaluation sets. Supporters: more engaged, following 5 or more member users. Sympathizers: less involved, following up to 2 member users.

Step 2. Snowball collection of friends and followers. To expand the datasets beyond direct members of the party, we look for less engaged users including supporters and sympathizers (see Figure 1). We rely on *follower* connections with member users as a proxy to collect less engaged users (Barberá 2015; Xiao et al. 2020), where the level of engagement or political interest is determined by the number of members they follow (Xiao et al. 2020). Engaged users with a strong interest in politics are referred to as 'supporters' if they follow 5 or more members of a party. On the other hand, users with less vested interest, such as those who follow 2 or fewer members, are called 'sympathizers'. The specific thresholds of 2 and 5 were empirically determined by looking at the frequencies in the data so that we could obtain a balanced number of supporter and sympathizer user groups (similarly done in previous work such as Xiao et al. (2020)). We retrieve up to 100 users per party for each of the supporter and sympathizer groups, filtering out those for which interaction data (see step 3) is not available, leading to the counts shown in the columns 'Supporter' and 'Sympathizer' of Table 1.

Step 3. Twitter timeline retrieval. For every user in the member, supporter and sympathizer groups we retrieve a history of their retweet interactions, regardless of whether these interactions are with users included in our datasets (see Figure 2). The purpose of this retrieval is to identify all users who have interacted with the labeled users through retweets, referred to as *interacting users*. Subsequently, we collected all available retweets from the timelines of both the labeled users and the interacting users, thereby extracting a substantial amount of interactions between pairs of users. Table 2 shows the final statistics of retweets retrieved and the number of total users performing those interactions.

Region	Party	Member	Supporter	Sympathizer
SCT	SNP	184	96	85
	SCU	59	97	84
	SL	52	95	86
	SGP	42	99	88
	SLD	24	98	94
	total	361	485	437
WAL	WL	55	97	88
	WC	42	98	85
	PC	42	99	85
	WLD	27	100	95
	total	166	394	353
NIR	SF	80	98	63
	DUP	65	75	67
	APNI	52	83	79
	UUP	58	73	68
	SDLP	59	76	72
	total	314	405	349

Table 1: Manually labeled Member users and automatically labeled Supporter and Sympathizer users, by region and class.

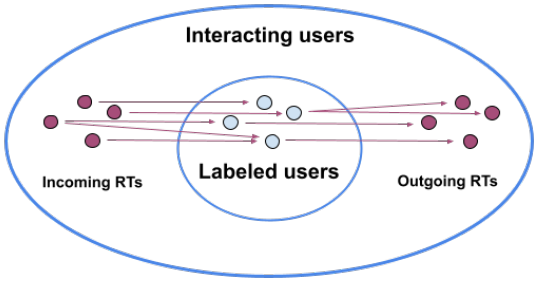


Figure 2: Interacting user's identification scheme. Central circle represents users manually or automatically labeled. External circle illustrates all the users interacting with the labeled users by retweeting them (incoming) or being retweeted by them (outgoing).

	SCT	WAL	NIR
Member users	361	166	314
Supporter users	485	394	405
Sympathizer users	437	353	349
Interacting users	87k	62k	21k
Retweets	19M	21M	4M
All users	937k	933k	426k

Table 2: Final dataset composition for each region.

Methods

We experiment with different unsupervised user representation methods in order to represent users based solely on interactions. Those features will be used along with different classification algorithms, to assess their effectiveness in embedding socio-political information. We also evaluate the impact of dimensionality reduction techniques.

User Representation Methods

We experiment with a set of user representation methods based on leveraging retweet interactions that can represent users' preferences and behavior. Thus, retweet based interactions on their own have been shown to be effective for user classification tasks (Conover et al. 2011a; Magdy et al. 2016; Darwish et al. 2020; Stefanov et al. 2020; Fernandez de Landa and Agerri 2022) and have been used to represent users as done in different approaches (Cignarella et al. 2020; Agerri et al. 2021; Fernandez de Landa and Agerri 2022; Darwish et al. 2020).

Next, we describe the proposed user representation methods which are suited to transform large and heterogeneous data sources.

ForceAtlas2 (Jacomy et al. 2014) is a continuous graph layout algorithm, that transforms a network into a 2-dimensional space to obtain a readable shape. This algorithm orders all the nodes represented in a graph according to the established relations, being a redesign of existing force-directed algorithm (Fruchterman and Reingold 1991). As a result of that approximation-repulsion process, those nodes that are unrelated repulse each other, while related ones will attract each other.

DeepWalk (Perozzi, Al-Rfou, and Skiena 2014) algorithm learns user representations by simulating uniform random walks among the connected nodes of the network. Based on Skip-gram (Mikolov et al. 2013), being given an instance the context or neighbours has to be predicted. The context for each instance is generated by the random walks among the surrounding connected data points of the instance. The length and number of walks for each instance will determine the context of such instance. Based on local information the method is able to learn representations which encodes structural regularities.

Node2vec (Grover and Leskovec 2016) algorithm is similar to DeepWalk, but adds two parameters to control the structure of the network due to a search bias while random walks happen. Those parameters are the return (p) and in-out (q) parameters. Return parameter (p) controls the probability to return to visited points in the random walks, at higher values the probability of revisiting a node decreases. The in-out parameter (q) controls the probability to explore undiscovered parts of the graphs, higher values are related to further points.

Relational Embeddings (Fernandez de Landa and Agerri 2022) are based on a single hidden-layer neural network trying to predict who retweeted whom for all the gathered interaction pairs. Instead of generating random walks among nearest neighbours, in this method the interactions are based on the relations between two users. This allows to capture social information and to build meaningful representations based on real user interactions.

The obtained retweet-based interactions are used to feed the aforementioned methods to train the unsupervised models. We use all the available users in order to embed as much information as possible. The user author of the retweet is considered the source of the interaction, while the user receiving the retweet would be the target. We generate source-target user pairs in this manner to provide the input to the user representation methods. With the intention of generating low dimensional but meaningful representations and based on previous work (Darwish et al. 2020; Stefanov et al. 2020; Fernandez de Landa and Agerri 2022), dimensions were set on 20 dims for DeepWalk, Node2vec and Relational Embeddings. The hyperparameters for node2vec and DeepWalk were set to the default values typically used by these algorithms: walks_per_node = 10, walk_length = 80, window or context_size = 10, and the optimization is executed for a single epoch (Perozzi, Al-Rfou, and Skiena 2014; Grover and Leskovec 2016). For node2vec, we set p=1 and q=0.5 in order to enhance network community related information (Grover and Leskovec 2016). Default parameters were set for ForceAtlas2 and independent models were trained for each of the regions.

Dimensionality Reduction Techniques

We want to study the performance of the best user representation method also for settings in which only limited labeled data is available, namely, in what we refer to as a weakly supervised scenario. We employ various dimension reduction techniques to compress information for weakly supervised settings with the aim of forcing our model to generalize more from the characteristics seen in the training data. We use 3 different dimensionality reduction methods:

PCA is a linear algorithm for dimensionality reduction which aims to represent the data in a low-dimensional space while preserving the original global structure of the data. As this is a linear algorithm, it will only find linear dependencies or relationships in the data, without considering the neighbours on their own.

t-SNE (Van der Maaten and Hinton 2008) is a nonlinear dimensionality reduction technique that embeds high-dimensional data into a lower dimensional space. A probability distribution is built over data point pairs, giving similar data points a high probability while dissimilar ones are assigned a lower one. The algorithm computes the probability that pairs of data points in the higher dimensional space are related, and then chooses low-dimensional embeddings which produce a similar distribution based on the Kullback–Leibler divergence.

UMAP (McInnes et al. 2018) or Uniform manifold approximation and projection is also a nonlinear dimensionality reduction technique. mThis technique is similar to t-SNE, but it assumes that the data is uniformly distributed on a locally connected Riemannian manifold. UMAP creates a fuzzy graph that reflects the topology of the high dimensional graph based on the nearest neighbours of each data point. Then the low dimensional graph is built based on the fuzzy graph. This dimension reduction technique is able to

reflect the large scale global structure, while also preserving the local structure.

The inputs are user vectors derived from the selected user representation method as presented in the previous subsection. Every dimension reduction techniques are used with their default hyper-parameters. Dimensionality is set to 2, following previous work (Darwish et al. 2020; Stefanov et al. 2020) and separate models are trained for each of the regions.

Experiment #1: Strongly Supervised Scenario

Experiment Settings. For each region, we experiment with the users in the group of members using a leave-one-out (LOO) cross-validation setting, i.e., one user is left for testing while all the others are used for training. Thus, the model is trained and tested once for each user in the dataset, which is feasible given the low dimensionality of the representations. The primary objective of this experiment is to compare the performance of user representation methods. The user representations obtained using the different methods are used to train six classification algorithms: Logistic Regression (LogReg), Random Forest (RF), Naive Bayes (NB) and linear, polynomial and RBF-kernel Support Vector Machines (SVM). We use their scikit-learn implementation (Pedregosa et al. 2011) with default configuration.

Results. Looking at the results reported in Table 3, we observe that the models trained with RE representations achieve best results for all the classifiers across every region. Among the models trained with RE representations, Logistic Regression consistently obtains the best results. On the other hand, and despite its popularity, the FA2 representations lead to the lowest performance scores, showing that it is the least suitable for this task. Both N2V and DW are clearly better than FA2, but still are clearly outperformed by the models obtained with RE representations. The RE method also behaves more robustly in multi-party scenarios and across regions. Finally, N2V, DW and FA2 show substantial variability across the different regions, while RE is the most stable method, showing robustness and adaptability.

Experiment #2: Weakly Supervised Scenario

Experiment Settings. We now experiment with a more challenging, weakly supervised scenario, where the models are provided with very limited training data in two different settings: (i) one-shot, where only one user is selected for training per class, and (ii) three-shot, a few-shot setting for which only three users are selected for training per class. The remainder of the users are left for the test set. In the interest of focus and brevity, for this setting we only use the RE user representations and Logistic Regression method, which was the best combination in Experiment #1 above. Furthermore, we also provide results obtained with and without dimensionality reduction techniques.

Results. Table 4 shows that 2 dimensional representations derived from t-SNE and UMAP dimension reduction technique get better results than RE without any dimensionality reduction for one-shot and few-shot settings. Results with PCA reduction are substantially worse across evaluation settings and regions. Compressing RE user representations into 2 dimensional representations with UMAP or t-SNE can be a good solution to handle community detection on weakly supervised scenarios as they can highlight communities due to their architecture based on nearest-neighbours. Interestingly, despite being evaluated on few-shot and one-shot settings, these methods obtain scores similar to those of RE on the strongly supervised scenario. Furthermore, results are consistent across the 3 regions, showing that RE representations reach competitive and robust results even in weakly supervised scenarios.

Experiment #3: Realistic Scenario

Experiment Settings. We define a more realistic, challenging scenario, in which we test the ability of the models to predict the political leaning of less engaged users, namely, of supporters and sympathizers. This assessment can offer insights into the model's applicability in a real-world context, where individuals may not be directly affiliated with political parties and have different levels of attachment. In order to do this, we use *members* for training and *supporters* and *sympathizers* for testing. We break down the performance of the models for each of the groups –supporters and sympathizers– to evaluate the impact of the level of political engagement of users to infer political leaning. For these experiments we use the two best overall classifiers in the strongly supervised scenario: Logistic Regression and Random Forest.

Results. Table 5 shows that the performance for this scenario is considerably lower. Overall, supporter users get better results than sympathizer users, showing that models suffer more when trying to learn political leaning of users that do not engage so much with political parties (which is only natural, in a way). N2V and DW fail to produce satisfactory results for this realistic scenario. We hypothesize that random walks may generate noisy or irrelevant paths that can negatively affect the quality of the embeddings. The FA2 method also fails to infer political leaning, showing that the use of a two-dimensional vector space for the approximation-repulsion process is insufficient to embed complex socio-political information. In any case, as it has been the case for previous experiments, the pair-based RE user representations are significantly better for every evaluation setting across every region. All the methods without exception show higher results on WAL datasets (4 classes) than on SCT and NIR datasets (5 classes). It seems that including one class more in a multiclass approach to political leaning inference has negative influence to classify users which are less engaged in politics (supporters and sympathizers).

Discussion

In this section we discuss the analysis of political leaning in light of the reported results. We also consider different ways of making our results explainable and provide an error analysis to identify possible weaknesses of our approach.

	SCT					WAL					NIR				
	Mj	N2V	DW	FA2	RE	Mj	N2V	DW	FA2	RE	Mj	N2V	DW	FA2	RE
LogReg	13.5	71.5	68.0	31.0	**99.4**	12.4	55.2	62.8	24.6	**99.2**	8.1	51.5	64.8	28.7	**97.7**
RF	13.5	81.8	78.2	63.5	**99.2**	12.4	67.5	78.2	68.9	**96.2**	8.1	66.4	80.9	76.3	**97.4**
NB	13.5	37.5	35.5	51.1	**99.5**	12.4	30.3	32.7	42.3	**98.7**	8.1	28.0	32.1	34.0	**97.7**
SVM - linear	13.5	73.0	69.0	30.6	**99.4**	12.4	33.2	54.7	37.9	**96.4**	8.1	35.4	46.9	41.2	**97.7**
SVM - poly.	13.5	39.7	43.2	30.0	**96.4**	12.4	22.2	25.9	12.6	**96.4**	8.1	08.1	10.3	10.3	**94.5**
SVM - rbf	13.5	41.3	43.1	61.6	**99.9**	12.4	26.1	28.6	38.4	**98.6**	8.1	17.8	29.7	50.6	**97.4**
average	13.5	57.5	56.2	44.6	**98.9**	12.4	39.1	47.2	37.5	**97.6**	8.1	34.5	44.1	40.2	**97.0**

Table 3: F1 macro score results leave-one-out CV on SCT, WAL and NIR member datasets. Mj refers to majority label classifier used as baseline. Algorithms used to generate the representations: N2V (Node2vec), DW (DeepWalk), FA2 (ForceAtlas2), RE (Relational Embeddings). Underlined values represent statistically significant ($p < 0.05$) highest average values.

	Dim. Red.	LOO	3-shot	1-shot
SCT	none	99.4	71.8	91.9
	UMAP 2d	**99.9**	91.2	90.8
	t-SNE 2d	99.4	**95.3**	**92.2**
	PCA 2d	87.3	71.2	67.4
WAL	none	**99.2**	96.9	**98.5**
	UMAP 2d	98.5	**97.8**	94.0
	t-SNE 2d	98.5	**97.7**	96.2
	PCA 2d	93.9	89.5	75.3
NIR	none	**97.7**	97.0	94.6
	UMAP 2d	97.3	**97.2**	94.7
	t-SNE 2d	97.0	94.8	**94.9**
	PCA 2d	69.9	65.5	49.8

Table 4: F1 macro score results on *member* datasets using logistic regression with RE user representations on strongly (leave-one-out cross validation) and weakly (3- and 1-shot) supervised scenarios.

Explainability. In order to better understand and explain the effectiveness of the different user representation techniques, we visualize RE, N2V and DW user representations for the three regions, SCT, WAL and NIR by performing t-SNE dimensionality reduction into 2 dimensions.

The first noticeable point looking at Figures 3, 4 and 5 is that, in contrast to N2V and DW, the visualizations obtained from the RE representations are clearly able to discriminate the multiparty political communities represented by the member users for each of the countries. In fact, the clear communities obtained in the visualization of the RE user representations is arguably in accordance with them outperforming other methods in the experimental evaluations reported in Tables 3 and 5.

Taking a closer look at the SCT visualization (left plot in Figure 3), we can see parties represented by clearly distinguishable communities. Thus, SNP (●) takes a big part of the figure, mainly isolated from the others. Unionist parties are forming their own cluster at the right side of the chart, separated from the other two parties. Inside the unionist community, it can be seen that SLD (●) acts like a link between SCU (●) and SL (●), showing its position as a central political actor. SLD also takes a central role in the which highlights its centrist political outlook. The representation locates SGP (●) apart from SNP and the unionists but between SNP and SL, showing a proximity to those. Furthermore, it can be seen that the pro-independence (SNP and SGP) and unionist parties (SL, SLD and SCU) are represented in different positions, showing a high polarization in the dispute across national identities.

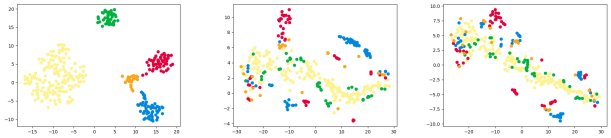


Figure 3: Visualization of t-SNE 2 dimension reduction of Relational Embeddings (left), node2vec (center) and Deep Walk (right) representations for SCT Member users.

Moving on to Figure 4 which shows the visualization obtained for WAL, it is possible to note that WLD (●) is situated in the center of the political spectrum, which is interesting as in reality they are considered to be positioned in the political center. The other 3 parties are surrounding WLD, but WL (●) is between PC (●) and WC (●), showing that the last two are the most opposed poles (left/right or pro-independence/unionist).

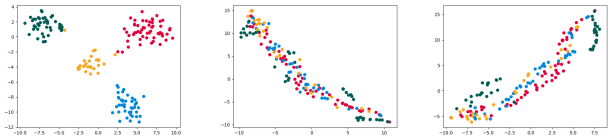


Figure 4: Visualization of t-SNE 2 dimension reduction of Relational Embeddings (left), node2vec (center) and Deep Walk (right) representations for WAL Member users.

Finally, for NIR we can see in Figure 5 that political parties are grouped in their own clusters, except for a few instances that may have been incorrectly classified. These few

		SCT					WAL					NIR				
		Mj	N2V	DW	FA2	RE	Mj	N2V	DW	FA2	RE	Mj	N2V	DW	FA2	RE
Supporter	LogReg	6.8	23.0	24.7	21.6	**90.8**	10.1	38.8	48.2	27.2	**95.3**	7.8	21.2	20.9	31.1	**75.4**
	RF	6.8	50.8	44.1	40.3	**81.4**	10.1	50.0	59.9	56.1	**93.8**	7.8	30.5	33.3	36.8	**65.4**
Sympathizer	LogReg	7.1	8.3	09.4	18.0	**63.3**	10.6	19.8	17.6	23.3	**60.6**	7.4	6.5	6.5	21.6	**41.2**
	RF	7.1	22.8	20.7	25.6	**51.7**	10.6	17.8	17.2	23.8	**60.2**	7.4	9.3	7.7	26.3	**38.1**
	avg.	7.0	26.2	24.7	26.4	**71.8**	10.4	31.6	35.7	32.6	**77.5**	7.6	16.9	17.1	28.9	**55.0**

Table 5: F1-score results on SCT, WAL and NIR Supporter and Sympathizer users datasets. Mj refers to majority label classifier used as baseline. Algorithms used to generate the user representations: N2V (Node2vec), DW (DeepWalk), FA2 (ForceAtlas2) RE (Relational Embeddings). Underlined values represent statistically significant ($p < 0.05$) highest average values.

errors seem to align with to those causing slightly lower performance, as shown in Table 3, of the RE user representations for NIR when compared to SCT and WAL. Looking at the positions of the parties, we can see that DUP (●) is located next to UUP (●) forming a conservative and unionist pole. Besides, SF (●) and SDLP (●) define the left-wing and pro-Irish pole. As a centralist political actor APNI (●) is located between both main groups, but much closer to the conservative-unionist pole forming with them a wider liberal-conservative, right-wing pole at the left of the chart. Summarizing, we believe that REs capture well multiple ideological disparities (left/right or pro-Irish/unionist) among these parties.

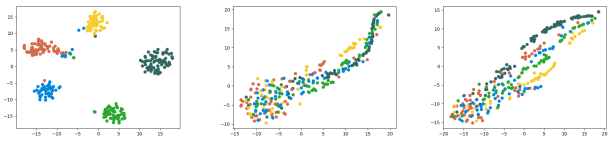


Figure 5: Visualization of t-SNE 2 dimension reduction of Relational Embeddings (left), node2vec (center) and Deep Walk (right) representations for NIR Member users.

The ability of RE user representations to depict distinct communities in all three cases is noteworthy, especially when compared to the inability of N2V and DW. Furthermore, REs are able to locate the communities following a pattern of ideological similarities and disparities. These observations lead us to infer that RE user representations have the potential to embed socio-political information within the generated features.

Data quantity. Relating the results with the number of the interactions collected in the datasets, we can see that there are some similarities among the user representation methods. The RE user representation method has better results for strongly supervised approach at SCT (19M RTs) and WAL (21M RTs) comparing to a small drop for NIR (4M RTs) which contains considerably less interactions (Table 2). The same occurs in the weakly supervised scenario, in which SCT and WAL obtain better evaluation results in comparison to NIR. This not only occurs with REs, but also with other user representation methods such as N2V and DW. The take out message is seems to be that the larger the number of interactions the better the results. The results for the realistic scenario reported in Table 4 further confirm this trend. Thus, the larger the timelines and the amount of users from which to extract the retweets, the better representations we get for all the user representation methods.

Error analysis. Considering the almost perfect scores obtained by REs on the strongly and weakly supervised scenarios, our error analysis will focus only on the realistic scenario, which consists of users less engaged politically, namely, supporters and sympathizers. The confusion matrices presented in Figures 6, 7 and 8 report the main errors performed by the classifiers based on REs on this particular scenario.

With respect to SCT, it can be observed that most misclassified instances correspond to models predicting SNP instead of the correct label. Thus, for supporter users (Figure 6 left), 22% of the errors in predicting SGP users (75% acc.) are wrongly predicted as SNP. This is even more pronounced for sympathizer users (Figure 6 right) given that the classifiers performs substantially worse in this evaluation setting. Thus, 61% of the errors in classifying SGP correspond to the model predicting SNP instead. We hypothesize that this may be explained by the fact that SNP and SGP have certain ideological similarities and have been in a cooperation agreement since 2021.

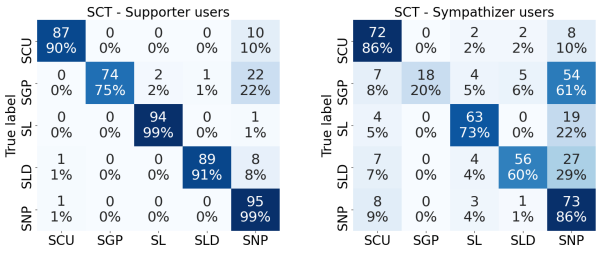


Figure 6: Confusion matrices of SCT Supporter (left) and Sympathizer (right) users of Logistic Regression trained with RE user representations.

The model trained for WAL, compared to the SCT case, has a lower error rate when predicting supporter's political

leaning (Figure 7 left). There, the few errors correspond to predicting WLD instead of the correct classes. If we look at the sympathizers errors (Figure 7 right), they substantially amplify the pattern seen for in the supporters setting. We believe that the source of errors may be caused by the central role played by WLD in Wales's politics and by the plurality in policies across the political spectrum.

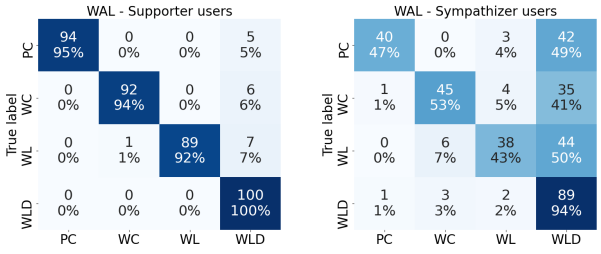


Figure 7: Confusion matrices of WAL Supporter (left) and Sympathizer (right) users of Logistic Regression trained with RE user representations.

With respect to NIR, we can see that the model has more problems discriminating between the different political options. If we look at the confusion matrix for supporters (Figure 8 left), DUP gets 32% of UUP and 39% of APNI instances. Moreover, 12% of the APNI errors correspond to UUP, which shows that the model is not able to detect well APNI users. This might be due to the centralist and liberal profile of APNI, which makes it difficult to discriminate from other political options. The same pattern can be observed for the sympathizer users (Figure 8 right) although amplified by the larger number of classification errors. It is particularly interesting the difficulties of the model in distinguishing UUP and APNI from DUP, which seems to indicate that the DUP is seen as the main reference of the right unionist space.

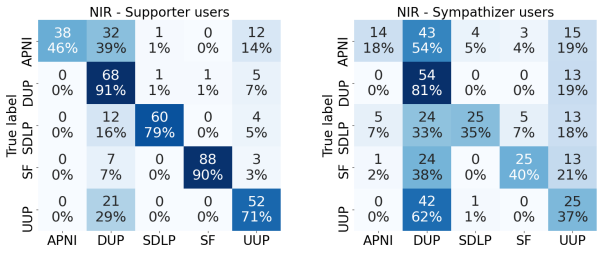


Figure 8: Confusion matrices of NIR Supporter (left) and Sympathizer (right) users of Logistic Regression trained with RE user representations.

The confusion matrices in the realistic scenario show that the best model suffers to clearly classify multi-party political leaning especially for users less engaged politically (sympathizers). It is noticeable the large amount of errors when trying to discriminate UUP and APNI from DUP, which is indicative perhaps of the dominance by DUP of the right wing agenda. In the case of WAL, the main source errors seem to be due to the centralist position of WLD with respect to other political options. Finally, in SCT we can see the phenomenon of a smaller party (SGP) cooperating with a larger one (SNP) and getting assimilated as a result of this collaboration. Summarizing, our political party-based outlook may capture sociopolitical information since errors commonly occur among ideologically adjacent classes and increase when targeting less politically engaged users. In any case, we confirm the necessity of addressing political leaning as a multipolar classification task which, despite being more difficult, would provide a more representative analysis of the social reality.

Conclusion and Future Work

In this work we look at the ability to infer the political leaning of social media users across multiple regions with multi-party systems, a challenging scenario that, to the best of our knowledge, has not been studied before. In order to do this, we collect a dataset spanning three UK regions, where users with different levels of political engagement (members, supporters, sympathizers) are labelled by the political party they align with. By conducting a set of experiments with these three datasets, we find that a model leveraging user interactions based on Relational Embeddings, in combination with a Logistic Regression classifier, achieves the best results. Unlike the other methods, REs use real user interactions without generating any artificial user connections (Fernandez de Landa and Agerri 2022). Experimental results are consistent across the three regions and different political engagement, demonstrating its robustness. However, experiments also show that predictions get particularly more challenging as the level of engagement of users decreases. Parallel to other social behaviors, less attachment means fewer performative actions that may define the political preferences of an individual, becoming more difficult to infer. Finally, visualizations and error analysis evidenced that REs are capable of capturing socio-political information.

There are other avenues for future research which were not within the scope of this work but would be worth exploring. For example, the use of other features beyond interactions could be potentially useful, as it could be the use of textual data to improve classification results for sympathizers. This would also enable the extensibility of the model to other social media, where retweet interactions do not exist. Moreover, in order to develop a richer political analysis, we are eager to apply the data extraction and user representation techniques into other tasks such as propaganda and disinformation detection.

References

Agerri, R.; Centeno, R.; Espinosa, M.; de Landa, J. F.; and Álvaro Rodrigo. 2021. VaxxStance@IberLEF 2021: Overview of the Task on Going Beyond Text in Cross-Lingual Stance Detection. *Procesamiento del Lenguaje Natural*, 67: 173–181.

Akoglu, L. 2014. Quantifying political polarity based on bipartite opinion networks. In *Proceedings of the Inter-*

national AAAI Conference on Web and Social Media, volume 8, 2–11.

Alkhalifa, R.; and Zubiaga, A. 2020. QMUL-SDS@SardiStance: Leveraging Network Inter-actions to Boost Performance on Stance Detection using Knowledge Graphs. In *Proceedings of EVALITA*. CEUR Workshop Proceedings.

Barberá, P. 2015. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23: 76 – 91.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10): 1531–1542.

Barberá, P.; and Rivero, G. 2015. Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review*, 33: 712 – 729.

Boutet, A.; Kim, H.; and Yoneki, E. 2012. What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election. In *Proceedings of ICWSM*.

Cignarella, A. T.; Lai, M.; Bosco, C.; Patti, V.; and Rosso, P. 2020. SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In Basile, V.; Croce, D.; Di Maro, M.; and Passaro, L. C., eds., *Proceedings of EVALITA*. CEUR-WS.org.

Conover, M. D.; Gonçalves, B.; Ratkiewicz, J.; Flammini, A.; and Menczer, F. 2011a. Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, 192–199. IEEE.

Conover, M. D.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011b. Political Polarization on Twitter. In *Proceedings of ICWSM*.

Darwish, K.; Stefanov, P.; Aupetit, M.; and Nakov, P. 2020. Unsupervised user stance detection on twitter. In *Proceedings of ICWSM*, volume 14, 141–152.

Fernandez de Landa, J.; and Agerri, R. 2021. Social analysis of young Basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, 0(0): 1–15.

Fernandez de Landa, J.; and Agerri, R. 2022. Relational Embeddings for Language Independent Stance Detection. *arXiv e-prints*, arXiv–2210.

Ferraccioli, F.; Sciandra, A.; Pont, M. D.; Girardi, P.; Solari, D.; and Finos, L. 2020. TextWiller@SardiStance, HaSpeede2: Text or Context? A smart use of social network data in predicting polarization. In *Proceedings of EVALITA*. CEUR Workshop Proceedings.

Fruchterman, T. M. J.; and Reingold, E. M. 1991. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21.

Garimella, V. R. K.; and Weber, I. 2017. A long-term analysis of polarization on Twitter. In *Proceedings of ICWSM*, volume 11.

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable Feature Learning for Networks. *Proceedings of KDD*.

Hardalov, M.; Arora, A.; Nakov, P.; and Augenstein, I. 2021. Cross-Domain Label-Adaptive Stance Detection. In *Proceedings of EMNLP*, 9011–9028.

Hua, Y.; Ristenpart, T.; and Naaman, M. 2020. Towards Measuring Adversarial Twitter Interactions against Candidates in the US Midterm Elections. In *Proceedings of ICWSM*.

Imhoff, R.; Zimmer, F.; Klein, O.; António, J. H.; Babinska, M.; Bangerter, A.; Bilewicz, M.; Blanuša, N.; Bovan, K.; Bužarovska, R.; et al. 2022. Conspiracy mentality and political orientation across 26 countries. *Nature human behaviour*, 6(3): 392–403.

Jacomy, M.; Venturini, T.; Heymann, S.; and Bastian, M. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6): e98679.

Jusup, M.; Holme, P.; Kanazawa, K.; Takayasu, M.; Romić, I.; Wang, Z.; Geček, S.; Lipić, T.; Podobnik, B.; Wang, L.; Luo, W.; Klanjšček, T.; Fan, J.; Boccaletti, S.; and Perc, M. 2022. Social physics. *Physics Reports*, 948: 1–148. Social physics.

Kermani, H.; and Adham, M. 2021. Mapping Persian Twitter: Networks and mechanism of political communication in Iranian 2017 presidential election. *Big Data & Society*, 8(1): 20539517211025568.

Kipf, T.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

Lai, M.; Cignarella, A. T.; Finos, L.; and Sciandra, A. 2021. WordUp! at VaxxStance 2021: Combining Contextual Information with Textual and Dependency-Based Syntactic Features for Stance Detection. In *Proceedings of IberLEF*, CEUR Workshop Proceedings.

Lynch, P. 2007. Party system change in Britain: Multi-party politics in a multi-level polity. *British Politics*, 2(3): 323–346.

Ma, X.; Wu, J.; Xue, S.; Yang, J.; Zhou, C.; Sheng, Q. Z.; Xiong, H.; and Akoglu, L. 2021. A Comprehensive Survey on Graph Anomaly Detection with Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.

Magdy, W.; Darwish, K.; Abokhodair, N.; Rahimi, A.; and Baldwin, T. 2016. #ISISisNotIslam or #DeportAllMuslims? Predicting Unspoken Views. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, 95–106. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342087.

Makazhanov, A.; and Rafiei, D. 2013. Predicting political preference of Twitter users. *Social Network Analysis and Mining*, 4: 1–15.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.*, 3: 861.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pennacchiotti, M.; and Popescu, A. M. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of KDD*.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. DeepWalk: Online Learning of Social Representations. KDD '14, 701–710. Association for Computing Machinery.

Preotiuc-Pietro, D.; Liu, Y.; Hopkins, D. J.; and Ungar, L. H. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *ACL*.

Rashed, A.; Kutlu, M.; Darwish, K.; Elsayed, T.; and Bayrak, C. 2021. Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey. In *Proceedings of ICWSM*.

Recuero, R.; Zago, G.; and Soares, F. 2019. Using social network analysis and social capital to identify user roles on polarized political conversations on Twitter. *Social Media+ Society*, 5(2): 2056305119848745.

Soares, F. B.; and Recuero, R. 2021. Hashtag Wars: Political Disinformation and Discursive Struggles on Twitter Conversations During the 2018 Brazilian Presidential Campaign. *Social Media+ Society*, 7(2): 20563051211009073.

Stefanov, P.; Darwish, K.; Atanasov, A.; and Nakov, P. 2020. Predicting the Topical Stance and Political Leaning of Media using Tweets. In *Proceedings of ACL*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaz de Melo, P. O. S. 2015. How Many Political Parties Should Brazil Have? A Data-Driven Method to Assess and Reduce Fragmentation in Multi-Party Political Systems. *PLOS ONE*, 10(10): 1–24.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio', P.; and Bengio, Y. 2018. Graph Attention Networks. *ArXiv*, abs/1710.10903.

Xiao, Z.; Song, W.; Xu, H.; Ren, Z.; and Sun, Y. 2020. TIMME: Twitter Ideology-Detection via Multi-Task Multi-Relational Embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, 2258–2268. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379984.

Zubiaga, A.; Wang, B.; Liakata, M.; and Procter, R. 2019. Political Homophily in Independence Movements: Analyzing and Classifying Social Media Users by National Identity. *IEEE Intelligent Systems*, 34: 34–42.

A.9 Fernandez de Landa *et al.* (2024b)

HTIM: Hybrid Text-Interaction Modeling for Broadening Political Leaning Inference in Social Media

Joseba Fernandez de Landa[a], Arkaitz Zubiaga[b], Rodrigo Agerri[a]

[a]*HiTZ Center - Ixa, University of the Basque Country UPV/EHU,*
[b]*Queen Mary University of London,*

Abstract

Political leaning can be defined as the inclination of an individual towards certain political orientations that align with their personal beliefs. Political leaning inference has traditionally been framed as a binary classification problem, namely, to distinguish between left vs. right or conservative vs liberal. Furthermore, although some recent work considers political leaning inference in a multi-party multi-region framework, their study is limited to the application of social interaction data. In order to address these shortcomings, in this study we propose Hybrid Text-Interaction Modeling (HTIM), a framework that enables hybrid modeling fusioning text and interactions from Social Media to accurately identify the political leaning of users in a multi-party multi-region framework. Access to textual and interaction-based data not only allows us to compare these data sources but also avoids reliance on specific data types. We show that, while state-of-the-art text-based representations on their own are not able to improve over interaction-based representations, a combination of text-based and interaction-based modeling using HTIM considerably improves the performance across the three regions, an improvement that is more prominent when we focus on the most challenging cases involving users who are less engaged in politics.

Keywords: Political Analysis, Computational Social Science, User-based Representations, Natural Language Processing

1. Introduction

Ideology is broadly understood as a set of people's beliefs formed by ways of thinking and acting in society [1]. Those beliefs can generally be represented by political parties, acting like social hubs of coordinated thoughts and actions. Thus, each region may have their own political parties, adjusted to a certain sociopolitical context in order to attract the votes and support of specific parts of the population. Understanding political leaning as the proximity to a political party, will allow to represent better ideological nuances than by reducing them to binary frameworks such as left vs right or conservative vs liberal. Therefore, our objective is to represent individual actors leveraging social media interactions and published text, allowing political leaning inference even in scenarios where no interactions are available, such as news media or political speeches. By doing so, we aim to incorporate wider and more accurate representations into the ongoing exploration of public opinions for diverse social science studies, including hate speech, disinformation or propaganda detection [2, 3].

The task of inferring the political leaning of social media users has widely been tackled through the use of user interactions as features [4, 5, 6, 7, 8, 9, 10], primarily based on the assumption that the behaviour exhibited by users through interactions with one another reflect homophilic connections and patterns of polarization between users [4, 7, 11, 12]. Others have approached the task by leveraging textual data instead of interactions through the use of Natural Language Processing techniques [13, 14, 15, 16, 17, 18, 19]. The few previous approaches that

have combined both texts and interactions [20, 21, 22], have addressed the task as a binary classification problem, in which only two political dimensions (e.g. right vs left) are considered at a time [4, 20, 5, 6, 9, 15, 17, 8, 7, 13, 10, 21, 23, 14, 16, 22]. The few studies that have gone beyond binary classification have been limited to a single region [19, 18]. This binary and region-specific standpoint represents the political context as a static and uniform reality.

In previous research [24], we showed that the use of features derived from interactions between users can lead to high performance in inferring the political leaning of social media users that are actively engaged in politics. However, a pure interaction-based approach proved to have clear limitations in making predictions for users who are less engaged in politics, hence hindering the broader applicability of the approach. Taking this research as a starting point, we propose a hybrid approach that aims to make the most of both textual and interaction-based features characteristic of social media. To achieve this, we introduce Hybrid Text-Interaction Modeling (HTIM), which integrates textual and interaction-based information into a hybrid model for improved political leaning inference across broader groups of social media users of varying degrees of political engagement. In doing so, our work presents the first attempt at addressing political leaning inference across a range of multiclass political realities fusioning text and interaction data.

Our proposed HTIM approach is flexible in that it can incorporate different encoding techniques. We test HTIM with different methods including text-based approaches like TF-IDF, Word2Vec [25], and Transformers [26], as well as interaction-based methods such as DeepWalk [27], Node2vec [28] and Relational Embeddings [29]. We evaluate the resulting HTIM-based models in three datasets pertaining to three different regions of the UK, each with different political parties. We study the performance of our models on users with different levels of engagement in politics, with a particular focus on users with lower levels of engagement in politics, whose posting of political content and interactions with content and users relevant to politics are predicted to be less frequent, which poses an additional challenge. Broadening political leaning inference with users of lower levels of engagement is important when recent studies, such as a survey conducted by the UK government [30], show decreasing levels of engagement in politics and in democracy from certain demographic groups.

Our experiments demonstrate that while interaction-based representations are superior to those based on textual content only, fusioning both types of information using HTIM lead to improved results, particularly for users with lower levels of political engagement. We therefore demonstrate the potential to broaden political leaning inference to a wider group of users not necessarily limited only to users who are highly active in politics.

Our work makes the following novel contributions:

- We propose Hybrid Text-Interaction Modeling (HTIM), as a means for hybrid modeling of social media users leveraging textual and interaction-based features, showing that the combination of both data types is required for optimal performance, especially for users with lower political engagement.

- We enquire into the proposed HTIM framework based on multiple political leanings anchored on representative political parties, ensuring adaptability to diverse regions and different levels of political engagement.

- We conduct political learning inference without relying on one specific data type. However, our results demonstrate that interaction-based Relational Embeddings outperform textual approaches, including those applying Transformer-based language models.

- All the data resources such as labeled users or their texts and interactions will be made publicly available upon publication for further study.

2. Related Work

In this section, we will discuss prior research on political leaning inference across different countries, paying special attention to the diverse data sources employed for this purpose. The analysis will commence with sources based on interaction-based data concerning retweets and user follows and subsequently advance to text-based data. Ultimately, we will consider approaches that integrate both types of data.

2.1. Interaction-based Political Leaning Inference

Well-known methodologies leverage interaction-based data obtained from social media for inferring political leaning. These methodologies are centered on user actions such as *following* and *retweeting*. For example, in one of the first research works about political leaning inference using Twitter political alignment was studied as a left-right spectrum in the US context, showing retweets as the most polarized interactions [4]. *Follow* interactions [5, 6] and even combinations of *follow* and *retweet* interactions [9] have also been leveraged to infer users left-right leaning. Similarly, retweets have also been used to infer the conservative and liberal leanings of Twitter users, quantifying such leanings [10], and even deploying them to analyze polarization and echo chambers [7]. Additionally, other works have investigated *retweet* behaviour in the context of pro- or anti-alignments among different topics [8].

2.2. Text-based Political Leaning Inference

Other approaches to political leaning inference revolve around analyzing text-based data derived from social media. Several researchers have focused on linguistic textual features as a means to measure political leaning along the liberal-conservative spectrum [13]. Sentiment analysis has also been employed to study left-right political leaning, as demonstrated in Plà and Hurtado [15]. Moreover, text-based methods based on topic vectors [14] or autoencoders trained on n-grams [16] have been applied to explore users' alignment with respect to the Democratic and Republican parties. Rashed *et al.* [19] used text-based embedding features to study political polarization in Turkey.

Alternatively, the PoliticEs 2022 shared task [17] aimed to extract political leaning from texts through an author profiling approach. In order to infer left or right alignment at user level, tweets published by the same author are grouped, either by concatenating them at the input stage or by merging the associated labels at the output stage. Best results in the PoliticEs 2022 task are achieved by methods based on Transformers [26], mainly fine-tuning pre-trained models. A similar idea based on author profiling is proposed by Fagni and Cresci [18] for inferring political leaning on Italy focusing on political parties, achieving best results with word2vec [25].

2.3. Hybrid Approaches to Political Leaning Inference

Several studies have employed a combined approach, by mixing textual content with interactions-based data. A pioneering research combined and compared text and interaction-based data in order to capture the left-right leaning of Twitter users, showing that retweets where the most representative interaction type to capture political alignment [20]. Another innovative study employed text and interactions, such as friends or followers, to infer the alignment of users with respect to the Democrat and Republican parties [21]. Hua *et al.* [23] also infer Democrat or Republican party preference of users based on hashtags, *retweets* and *followers*. Continuing within the US framework, another study aimed to estimate user ideology as liberal or conservative, achieving best results while combining textual content with a social graph derived from retweets and followers [22]. Finally, interaction and textual features were used to identify pro or anti independence Twitter users in Catalonia, Basque Country and Scotland [31].

2.4. Related Tasks

Similar to political leaning inference, but focused on a specific topic, stance detection tasks [32, 33] commonly target the political domain. Recently released datasets [34, 35] enabled researchers to apply both textual content and social interactions [36, 37, 38], emphasizing the significance of interactions in determining the stance of users via manually engineered approaches. In previous research [29], we introduced Relational Embeddings, achieving better results than DeepWalk [27] and node2vec [28] while showing that interaction and textual inputs provide best results across different datasets without any manual engineering. The present work leverages these and other methods within the HTIM framework for political leaning inference, thereby broadening their applicability to a wider spectrum of users of varying degrees of engagement in politics.

3. Datasets

For building our own datasets, we follow a generalizable methodology applicable to different political realities [24]. First, we choose the political context we want to analyze by selecting regions and their most relevant political parties. Then, we start with the data collection from Twitter by manually labeling users and collecting their associated textual and interaction-based data to build the proposed user representations.

3.1. Political Context

Given our interest in exploring multi-party and diverse scenarios, we chose the United Kingdom (UK) and examined three regions with devolved governments: Scotland, Wales, and Northern Ireland. These regions possess a diverse political landscape, characterized by strong nationalist sentiments and a wide range of political options. Our analysis focuses on the major political parties represented in each region's devolved parliament.

Scotland (SCT): The Scottish political landscape is dominated by five major parties, namely, the Scottish National Party (SNP), the Scottish Conservative & Unionist Party (SCU), the Scottish Labour Party (SL), the Scottish Green Party (SGP), and the Scottish Liberal Democrats (SLD). The SNP is a center-left party that advocates for Scottish independence and prioritizes Scotland's membership in the European Union. The SCU, on the other hand, is a center-right party that supports the union with the UK and opposes Scottish independence. The SL is a center-left party that also supports the union with the UK and opposes Scottish independence. The SGP, a left-wing party, advocates for both Scottish independence and membership in the European Union. The SLD is a center party that supports both Scotland's membership in the European Union and the union with the UK.

Wales (WAL): In Welsh politics there are four major parties that dominate the landscape: Welsh Labour (WL), Welsh Conservatives (WC), Plaid Cymru (PC), and Welsh Liberal Democrats (WLD). WL is a center-left party that supports the UK and opposes Welsh independence. The WC is a center-right party also supporting the UK and opposing Welsh independence. PC is a left-leaning party that advocates for Welsh independence and greater autonomy from the UK. The WLD is a centrist party that supports the European Union and the union with UK.

Northern Ireland (NIR): In Northern Ireland politics, there are five main political parties: Sinn Féin (SF), Democratic Unionist Party (DUP), Alliance Party of Northern Ireland (APNI), Ulster Unionist Party (UUP), and Social Democratic and Labour Party (SDLP). SF is a left-wing party that advocates for Irish unity. The DUP is a right-wing party that supports the UK and opposes Irish unity. APNI is a centrist party that supports the UK and the European Union, and also advocates for greater cross-community cooperation in Northern Ireland. UUP is a center-right party that supports the UK and opposes Irish unity. The SDLP is a

center-left party that supports Irish reunification and greater cooperation between Northern Ireland and the Republic of Ireland.

3.2. Types of Users based on Levels of Political Engagement

Political participation encompasses different levels of involvement [39], ranging from passive observers to highly engaged activists, each harboring distinct perspectives and motivations. Therefore, in line with previous research in political science studying engagement levels [40], in our work we define members, supporters, and sympathizers, labeled according to their alignment with a specific political party:

- **Members** comprise users directly affiliated with political parties, including elected representatives and affiliated organizations, being users which are highly active in politics.
- **Supporters** are users closely aligned with political parties but not as actively engaged as members.
- **Sympathizers** are users with a loose connection to politics, making it more challenging to associate them directly with political parties due to their lower level of engagement which translates in very low activity related to political activities.

By incorporating different levels of political involvement we facilitate a more granular analysis of political landscapes, enhancing our comprehension of the multifaceted nature of political leaning. Additionally, we simulate realistic political scenarios to assess the methods in more feasible real-life situations. Member, supporter and sympathizer sets are designed to represent different levels of difficulty, with more involved user predictions expected to be easier than less involved user predictions. This step was taken to provide a more realistic evaluation of the proposed methods, facing situations that could happen in real life.

3.3. Data Collection

Once the regions and the political parties are selected, we begin with the data collection process. The first step is to manually generate a seed dataset by labeling Twitter users from each of the identified parties. In a second step, we automatically extract users which are related to the already labeled users, categorizing them as interacting users. Finally, we extracted the timelines of all collected users to obtain considerable amounts of both textual and interaction-based data.

(1) **Labeling of Member users:** In line with previous studies [41, 6, 9, 23], we employed a similar technique based on set of seed users to collect data for our study. We began by selecting a group of users who are affiliated with the political parties of interest, including elected members and associated organizations. These users were manually identified and labeled with the corresponding political party. For each region, we curated a list of member users independently, which served as the basis for data collection. The number of user's labeled by class can be seen in Table 1 (Member column).

(2) **Supporter and Sympathizer evaluation datasets:** We further gather two datasets for each area: one consisting of supporter users and a second one consisting of sympathizer users. To construct these evaluation sets, we first extracted users based on their political party affiliations and their followers as it has previously been done in similar approaches [7, 19]. Thus, supporter users were defined as those who followed five or more member users from a specific political party, while sympathizer users followed two or fewer users from each party [24]. We collected 100 users for each party from each region for both supporter and sympathizer sets, and these users were automatically labeled based on their party affiliation. Then we select users with available data to create the final datasets as seen in Table 1 (Supporter and Sympathizer columns).

Region	Party	Member	Supporter	Sympathizer
SCT	SNP	181	91	74
	SCU	59	86	81
	SL	52	88	72
	SGP	42	82	77
	SLD	24	90	84
	total	358	437	388
WAL	WL	55	92	77
	WC	42	91	75
	PC	42	91	72
	WLD	27	98	81
	total	166	372	305
NIR	SF	79	92	37
	DUP	61	44	54
	APNI	52	62	66
	UUP	57	52	48
	SDLP	58	54	65
	total	307	304	270

Table 1: Manually labeled (Member) users and automatically labeled users (Supporter and Sympathizer) for realistic evaluation, by region and class.

(3) **Timeline extraction:** In this phase we will get data to characterize the labeled twitter users based on text and interactions. Regarding textual data, we collected 120 tweets per Member user and 60 tweets for each Supporter and Sympathizer users. To achieve a balance between the number of users and available data, we excluded labeled users with insufficient data and discarded tweets with fewer than 10 tokens. In the case of interaction-based data, we first identified all the users who had interacted via retweets with each of the labeled users, referred to as *interacting* users. After that all the available retweets were extracted from the timelines of both the labeled and interacting users.

Political party selection and manual labeling of Member datasets were carried out in September 2022. Supporter and Sympathizer users evaluation sets were built on October 2022. Twitter data extraction was undertaken during October 2022, collecting the timelines of all the identified users. In Table 2 it can be seen the shape of the final corpus for each region: (i) Manually labeled users (Members) and gathered tweets (120 per user); (ii) automatically labeled users for the realistic evaluation (Supporters and Sympathizers) and the collected tweets for each of them (60 per user); (iii) interaction based retweet data from labeled and interacting users.

4. Methods

We conduct experiments using various methods to extract user representations from text and interaction data. On the one hand, text-based features are employed to represent users using textual data. On the other hand, interaction-based features derived from retweets are utilized to compare them with the textual approaches. Finally we propose a combination of textual and interaction features in our HTIM approach. User representations are then used to perform political leaning inference via alignment to the political parties included in our dataset.

4.1. Text-based Features

We explore a range of text-based feature extraction techniques in order to represent users. The extraction process involves generating user vectors based on the textual content associated

		SCT	WAL	NIR
Members	users	358	166	307
	tweets	42,960	19,920	36,840
	tokens	1,400k	789k	653k
Supporters	users	437	372	304
	tweets	26,220	22,320	18,240
	tokens	654k	676k	497k
Sympathizers	users	388	305	270
	tweets	23,280	18,300	16,200
	tokens	1,194k	523k	436k
Interactions	interacting users	87k	62k	21k
	retweets	19M	21M	4M

Table 2: Final dataset composition for each region.

with each user, namely, their tweets. By utilizing tweets as input data, we aim to capture users' preferences and behaviors. With that purpose we will employ text-based user representations to predict users' political leaning, following similar methodologies to those employed in previous studies [18, 29, 42].

Term Frequency Inverse Document Frequency. The tfidf statistical measure assesses the relevance of a word to a document within a set of documents. By lowering the impact of words that occur too frequently in the selected text collection the most salient features are selected. Our use of this approach is motivated by the fact that a limited set of words significantly impact the final predictions [43] and the remarkable results obtained in other similar text classification approaches [29, 17]. In this case, all the tweet collections from each user are considered as individual documents. The obtained tfidf vectors for each author or user are used to learn a classifier, proposing a user level classification.

Word Embeddings. We use Word2vec [25] (w2v) to encode each of the tweets into text-based vector representations. To accomplish this we train our own models from scratch in order to fit all the words from our datasets into the vocabulary. A separate language model is trained for each region, aiming to capture the prevailing expressions within specific communities. These models are trained using default hyperparameters (C-BOW, negative sampling, 5 epochs, 5 words window size) but different dimensions are considered to select the optimal configuration. For our purposes, we extract tweet vectors one by one, representing each tweet as the average of its word vectors [44, 18].

Transformers. We use pre-trained Transformer models [26] as they have garnered considerable attention in recent years for text classification tasks, including user profiling through textual data [17]. These models enable context and meaning representation by analyzing the relationships among tokens in a text sequence. Transformer-based contextualized embeddings values are modified depending on the surrounding words and their order, while static embedding methods such as word2vec represent words with fixed vector values. Four multilingual models have been selected for our experiments:

- **mBERT** [45] model is the multilingual version of BERT [45] pre-trained with the largest 104 languages in Wikipedia. Rather than simply predicting the next word in the sequence, the BERT model takes into consideration all of the words in the sequence, thereby developing a more profound comprehension of the context. BERT pre-trains bidirectional

representations from unlabeled text by considering both left and right context in all layers based on two pre-training objectives, namely, mask-language modeling and next sentence prediction.

- **DistilmBERT** [46] as the multilingual version of DistilBERT is a smaller and faster Transformer model distilled from BERT, with 40% fewer parameters and 60% faster performance while maintaining over 95% of BERT's performance on the GLUE benchmark.
- **XLM-RoBERTa** [47] is a multilingual version of RoBERTa-base [48] pre-trained on a large multilingual corpus containing 100 languages. As a robustly optimized BERT approach, it is trained on a 10 times larger dataset than the used in BERT and using a dynamic masking technique, byte-pair encoding tokenization and without the next-sentence prediction objective.
- **XLM-T** [49] is an extension of the XLM-RoBERTa base model further trained on 198 million multilingual tweets. Given its focus on Twitter-based data, it is particularly relevant to assess its performance in tasks that are specific to this social media platform.

To ensure consistency across all categories and prevent overfitting, we employ the Transformers models without fine-tuning, meaning that we use the default frozen weights as done in other approaches [50]. Text features are extracted separately for each tweet, treating each individual tweet as a sequence. We explored three distinct approaches for representing the text of tweets using transformers: (a) *start-of-sequence* initial token embedding is used as the entire tweet representation [45]; (b) *average* value of the output embeddings of all words in the tweet to represent each tweet as the average of its word vectors [44, 18]; (c) *max-pool* value of all the words in a tweet to extract the most salient features from every word-embedding, by taking the maximum values among all the word vectors [51].

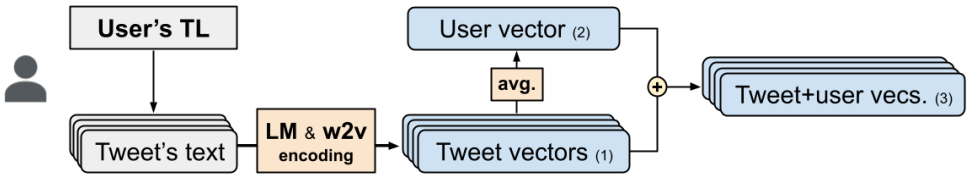


Figure 1: Pre-trained Transformer-based Language Models (LM) and Word2vec embeddings (w2v) usage to extract text-based user representations: (1) at Tweet level, (2) at user level and (3) the combination between tweet and user features.

While the tfidf method is capable of directly representing all of a user's content at once through user-level features, Word2vec and Transformers are able to extract information only at the tweet level. As the tweet level information is not sufficient to represent users due to the small amount of text, we generate and add user-level textual features to each of the tweets representation as shown in Figure 1. So, we obtain the textual representations of all the tweets for each user (1) which are then concatenated and averaged to obtain a user-based vector representation (2), emulating similar approaches [42, 44, 52, 19]. The final representation (3) consists of the concatenation of each tweet vector with the user-based vector representation. The combined user and tweet representations for each tweet are used to train a classifier. After predicting the labels at tweet level, a majority voting strategy is employed to infer the user label by considering the various tweet labels associated with the same author [17, 50].

4.2. Interaction-based Features

The intuition is that of building interaction based-user representations able to capture sociopolitical information, without any textual data. We focus on retweets as they have been

previously proven effective to perform user classification [20, 53, 8, 54, 29]. User interactions are brought into a low dimensional vector space, modeling retweets into user level representations. Three distinct unsupervised techniques are used to represent users, namely DeepWalk [27], Node2vec [28] and Relational Embeddings [29]. Other approaches such as GCN [55], GAT [56] or TIMME [57] are discarded as they are not suitable for unsupervised learning and have high memory requirements.

DeepWalk (DW) algorithm simulates random walks among connected nodes in a network to learn feature representations. It predicts the context or neighbors of an instance using the Skip-gram method [25]. The context is generated by random walks among surrounding connected data points, with the length and number of walks determining the context.

Node2vec (N2V) method, similar to DeepWalk, introduces two parameters, p and q, to influence the network structure during random walks. The return parameter (p) determines the likelihood of revisiting nodes and the in-out parameter (q) controls the probability of exploring unexplored graph areas.

Relational Embedings (RE) method aims to predict which user retweeted another user among all the collected interaction pairs. Unlike generating random walks among nearest neighbors, this approach focuses on the real relationships between two users.

We use all identified users, including those labeled and interacting users, together with their retweets, to provide input for the techniques mentioned above for training the models. Following previous work [8, 54, 29], by applying any of the three methods above we obtain low dimensional dense interaction-based representations. The user-based features are generated without applying any data filtering to any of the models. In order to create meaningful features while keeping the dimensionality low, we set the feature dimensions to 20 for DeepWalk, node2vec, and Relational Embeddings, drawing from previous studies [8, 54, 29]. For node2vec and DeepWalk, we use the default parameter values typically employed by these algorithms: walks_per_node = 10, walk_length = 80, window or context_size = 10, and one epoch [27, 28]. Specifically for node2vec, we set p=1 and q=0.5 to emphasize network community-related information [28].

	SCT				WAL				NIR				average			
Dims.	50	100	200	300	50	100	200	300	50	100	200	300	50	100	200	300
tfidf	22.4	39.3	48.4	<u>57.2</u>	45.3	54.8	64.2	<u>64.8</u>	30.3	40.4	55.5	<u>60.2</u>	32.7	44.8	56.0	<u>60.7</u>
w2v	57.9	62.3	63.5	<u>64.0</u>	61.8	61.3	61.3	<u>64.0</u>	54.2	56.3	57.5	<u>57.9</u>	58.0	60.0	60.8	<u>62.0</u>

Table 3: F1 macro score results 10 fold CV on SCT, WAL and NIR Members datasets. Algorithms used to generate the features: tfidf and w2v. Underlined values represent best result for each algorithm in each dataset.

		SCT				WAL				NIR			
	Transformers	B	dB	R	Rt	B	dB	R	Rt	B	dB	R	Rt
Features	start-of-sequence	35.9	33.7	07.2	37.2	55.7	55.3	41.0	56.4	45.6	42.0	10.4	39.0
	average	54.2	48.6	12.7	39.7	68.4	64.0	52.0	56.1	56.7	49.3	20.1	40.1
	max-pool	<u>67.6</u>	<u>73.4</u>	<u>47.6</u>	<u>60.7</u>	<u>75.6</u>	<u>75.1</u>	<u>58.0</u>	<u>64.4</u>	<u>68.4</u>	<u>72.1</u>	<u>50.7</u>	<u>57.4</u>

Table 4: F1 macro score results for 10 fold CV on SCT, WAL and NIR Members datasets. Algorithms used to generate the features: mBERT (B), DistilmBERT (dB), XLM-RoBERTa (R), XLM-T (Rt). Underlined values represent best results for each algorithm in each dataset.

4.3. HTIM: Hybrid Text-Interaction Modeling

We propose a hybrid approach which integrates both the textual content expressed by a user and their corresponding social media interactions. When using text representation methods at tweet level (Figure 2), the textual representation of the tweets for each of the users (1) are

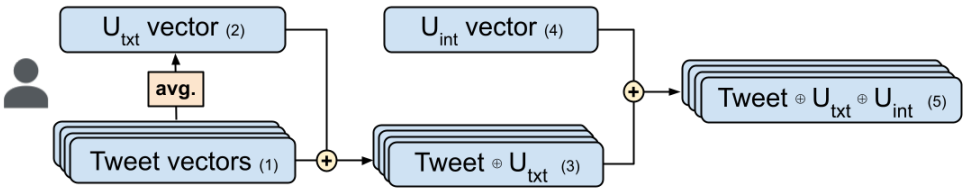


Figure 2: Hybrid Text-Interaction Modeling (HTIM) for each user tweet by tweet: (1) Tweet level text representation, (2) user level text representation, (3) the combination between tweet and user text features, (4) user level interaction representation and (5) final hybrid representation concatenating all vectors.

averaged to obtain a vector characterizing each user (2) and concatenated with each of the tweets written by the author (3). Afterwards, (4) we concatenate them with the interaction-based representations to obtain a final (5) hybrid feature for each of the tweets. Those hybrid representations are used to train a classifier and to predict the labels at tweet level, and we subsequently use a majority voting strategy to infer the user label by considering the various tweet labels associated with the same author [17, 50].

Alternatively, when text representation methods (i.e. tfidf) facilitate the extraction of text features at the user-level, hybrid features are created at user level, concatenating user-level text and interaction features. User-level hybrid representations are used to train a classifier and to predict the labels at user level.

5. Experimental Setup

We leverage the obtained representations in order to conduct political leaning inference. We do so in two different sets of experiments. First, we focus on determining the optimal configuration to obtain good quality textual representations. Second, we will apply the best textual representations, together with interaction-based representations, in our HTIM hybrid approach, evaluating them on the Member, Supporter and Sympathizer datasets across the 3 regions.

Textual Methods Selection. For each region, we experiment with the users in the group of *Members* using a 10 fold cross-validation (CV) setting, i.e., a 10% of the users are left for evaluation while all the others are used for training. The primary objective of this experiment is to compare the performance of user representation methods. The representations obtained using the different methods are used to train a RBF-kernel Support Vector Machine (SVM) classification algorithm. We use the scikit-learn implementation [58] with default configuration.

With respect to tfidf and w2v, different dimension values were tested, as shown in Table 3. Ultimately, the best dimension value for both methods is set to 300. Particularly, for tfidf, as the dimension value increases, the results tend to improve. Regarding Transformer-based methods, the results in Table 4 demonstrate that *max-pooling* proves to be the most effective strategy. The best textual user representation methods are used to be compared to and combined with interaction-based user representation methods.

Experimental Settings. We will extract representations from tweets and interactions to train independent user classification models for each of the political regions: SCT, WAL and NIR. Each of the regions will have its own user representations to observe the performance of the methods on different scenarios.

In order to compare the quality of our proposed user representation methods considering the level of political engagement of the users, we evaluate separately on the Member, Supporter and Sympathizer datasets. On the one hand, we train and evaluate a SVM (RBF kernel) classifier [58] using 10 fold CV with the Members dataset. On the other hand, we use the

Members dataset to train another SVM (RBF kernel) classifier and then evaluate the model on the Supporter and Sympathizer datasets. To ensure consistency across all categories and prevent overfitting, we have employed the default or automatic hyperparameters for all the aforementioned classifiers. In addition, majority and random label predictors are added as baselines for each of the datasets.

		SCT			WAL			NIR			average			
		Mem.	Sup.	Sym.	Mem.	Sup.	Sym.	Mem.	Sup.	Sym.	Mem.	Sup.	Sym.	ALL
Baselines	majority	13.4	06.9	06.4	12.4	09.9	10.1	08.2	09.3	04.8	11.3	08.7	07.1	09.0
	random	17.8	21.6	18.7	21.5	27.9	28.1	21.2	21.9	16.1	20.2	23.8	21.0	21.6
Interactions	RE	99.4	91.5	62.7	97.9	95.8	59.6	97.6	72.9	33.0	98.3	86.7	51.8	78.9
	N2V	80.8	62.0	17.4	67.5	48.5	11.1	60.8	21.6	07.0	69.7	44.0	11.8	41.9
	DW	80.9	61.6	22.9	77.5	57.0	14.7	72.2	23.6	06.6	76.9	47.4	14.7	46.3
Text	tfidf	57.2	52.4	29.5	64.8	59.5	26.1	60.2	45.8	25.9	60.7	52.6	27.2	46.8
	w2v	64.0	40.5	27.5	64.0	51.2	31.5	57.9	37.9	26.7	62.0	43.2	28.6	44.6
	B	67.6	47.8	27.7	75.6	55.2	33.9	68.4	48.5	34.4	70.5	50.5	32.0	51.0
	dB	73.4	59.0	36.9	75.1	64.1	33.6	72.1	49.8	28.5	73.5	57.6	33.0	54.7
	R	47.6	43.7	29.8	58.0	42.0	28.0	50.7	42.0	24.3	52.1	42.6	27.4	40.7
	Rt	60.7	48.3	34.5	64.4	52.2	32.7	57.4	44.0	27.8	60.8	48.2	31.7	46.9
HTIM	RE + tfidf	**99.7***	**97.7***	**74.2***	**99.2***	**98.4***	**67.0***	97.3	**82.8***	**48.1***	98.7*	**93.0***	**63.1***	**84.9***
	RE + w2v	99.4	95.6*	66.0*	98.5*	96.0*	64.4*	98.2*	72.7	37.2*	98.7*	88.1*	55.9*	80.9*
	RE + B	98.5	94.0*	63.1*	**99.2***	94.9	60.8*	97.7*	78.5*	45.8*	98.5*	89.1*	56.6*	81.4*
	RE + dB	99.4	95.4*	64.3*	**99.2***	94.7	61.6*	**98.4***	80.2*	42.9*	**99.0***	90.1*	56.3*	81.8*
	RE + R	99.4	94.6*	66.0*	98.6*	96.5*	64.7*	97.4	78.5*	41.1*	98.5*	89.9*	57.3*	81.9*
	RE + Rt	99.4	94.4*	66.3*	**99.2***	96.0*	62.3*	98.1*	78.4*	44.7*	98.9*	89.6*	57.8*	82.1*
	N2V + tfidf	74.5	44.5	11.3	42.2	26.7	11.8	32.5	10.3	06.7	42.8	34.6	11.1	29.5
	N2V + w2v	89.8*	56.0	28.6*	75.1*	60.2*	32.7*	73.1*	45.2*	25.4	79.3*	53.8*	28.9*	54.0*
	N2V + B	77.6	52.1	27.7	74.5	55.9	35.6*	68.0	48.3	34.6*	73.4*	52.1*	32.6*	52.7*
	N2V + dB	83.2*	60.5	35.0	74.8	65.1*	34.4*	71.6	50.1*	29.2*	76.5*	58.6*	32.9	56.0*
	N2V + R	75.4	47.1	28.2	58.4	45.6	26.7	52.5	37.3	22.6	62.1	43.3*	25.8	43.8*
	N2V + Rt	79.4	51.0	32.8	65.5	57.6*	30.6	57.9	42.8	28.1	67.6	50.5*	30.5	49.5*
	DW + tfidf	71.9	38.1	13.1	57.1	28.8	10.5	45.7	13.2	06.7	46.5	37.5	12.3	32.1
	DW + w2v	87.5*	58.0	30.1*	79.4*	64.2*	31.0	78.5*	46.2*	29.2*	81.8	56.1*	30.1*	56.0*
	DW + B	76.5	53.3	27.5	75.5	56.4	35.6*	69.2	49.2*	34.3	73.7	53.0*	32.5*	53.1*
	DW + dB	83.8*	62.0*	35.6	75.9	65.0*	33.9*	72.1	50.2*	29.4*	77.3*	59.1*	33.0*	56.4*
	DW + R	74.6	43.4	25.1	63.9	50.5	26.8	62.0	37.9	22.0	66.8	43.9*	24.6	45.1*
	DW + Rt	78.6	48.9	33.5	67.8	57.3	29.8	61.8	42.4	27.7	69.4	49.5*	30.3	49.8*

Table 5: Macro-F1 scores on SCT, WAL and NIR from Member (10 fold CV), Supporter and Sympathizer datasets. Algorithms used to generate the representations: Relational Embeddings (RE), Node2vec (N2V), DeepWalk (DW), tfidf, word2vec (w2v), mBERT (B), DistilmBERT (dB), XLM-RoBERTa (R), XLM-T (Rt). Values in **bold** represent best overall results for each dataset, while underlined values represent best results on text-only and interactions-only framework. Values with * represent when the combination of text and interactions gets better results than each on its own.

6. Analysis of Results

In this section we analyze the results obtained for different regions (SCT, WAL and NIR) on Member, Supporter and Sympathizer datasets, testing the methods through regions and different levels of political attachment. On the one hand, we are interested in evaluating standalone text-based and interaction-based methods. On the other hand, we compare standalone methods with the use of HTIM for combining interactions and text.

Text vs interaction representations. In Table 5 we can compare the results for text and interaction-based representations on their own, showing which data type is better to represent political leaning. Regarding text-based representations, Bert-based (B and dB) approaches generally yield superior results compared to Roberta-based (R and Rt), w2v and tfidf methods. Interestingly, despite being a smaller model, dB achieves superior results compared to B, being the best text based approach. Among Roberta-based approaches, Rt is significantly superior to the R method, given that the former is an extended version of R that has been retrained using Tweets. Bert-based (B and dB) approaches also outperform interaction-based N2V and DW on politically less engaged users. However, interaction-based approaches tend to be better with politically engaged Member users.

In conclusion, max-pooled DistilmBERT Transformer model is the best approach to tackle political leaning inference with textual data, outperforming other Transformer configurations

as well as w2v and tfidf baselines. Furthermore, we have to remark that none of the textual approaches come close to surpassing the performance of interaction-based RE representations, which consistently yield superior results across all regions and political attachments. However, RE is not able to perform well with politically less engaged users, as the performance drops as the political involvement decreases.

Representations using HTIM. Taking a wider look into the whole Table 5, the results indicate that the use of our proposed HTIM for the integration of text and interaction representations leads to a notable improvement in results as compared to their independent usage (ALL column on Table 5). The RE method consistently yields better results when compared to other approaches based on interactions or text on their own. However, incorporating any of the extracted text representations alongside the RE representations often leads to improved results. This is particularly beneficial for politically less engaged users since their interactions alone may not provide enough information to determine their orientation accurately.

Thus, RE combined with Transformer-based representations (B, dB, R, Rt) generally perform better than when the RE are on their own. Especially, RE combined with Rt (RE+Rt) outperformed RE method for the whole 3 regions and all the political engagements. Hence, the combination of RE with Transformer-based representations (B, dB, R, Rt) typically yields superior performance compared to standalone RE. Furthermore, tfidf representations, which do not perform well when combined with *DW* and *N2V* representations, considerably enhance the results achieved by *RE* when combined with them. Thus, results on the Sympathizer datasets improve more than 10 points in average across the three regions, while for Supporter they increased more than 5 points. Considering that the added *tfidf* representations are based on the most significant terms per user, we can hypothesize that certain referential terms may serve as anchor terms for specific political parties.

Results for different levels of engagement. We next look at the performance scores with a focus on the three levels of engagement, i.e. members, supporters and sympathizers. Results for these three groups as shown in Figure 3 indicate that, as hypothesized, politically less engaged users are more difficult to predict for all the selected approaches. This observation is supported by the downward trend of all the lines, visually showing a steep performance decrease as engagement fades. This trend in turn reinforces the need for a data collection strategy like the one we defined here to collect not only actively engaged users, but also those with more modest levels of engagement which do need to be considered in these analyses. Furthermore, we also observe that political leaning inference of highly engaged users is best achieved by exploiting their content sharing actions or interactions (N2V and DW); this however changes when we shift our focus towards less engaged users, where the use of textual content becomes more crucial as the interactions alone do not suffice (i.e. tfidf, dB and Rt). Interestingly, the combination of both data types through HTIM further improves the results, especially when the hard-to-beat RE method underperforms on politically less engaged users.

Results across regions. We are also interested in looking at how results differ across the three regions under study, i.e. Scotland (SCT), Wales (WAL) and Northern Ireland (NIR). We show results for each region in Figure 4, which shows that performance scores across regions also vary depending on the level of political engagement. Performance is consistenly high and comparable across all three regions when we look at highly engaged users (i.e. members). These trends vary more across regions when we look at less engaged users. As performance decreases for less engaged users across all three regions, we see that this drop is more prominent to some extent for WAL but particularly for NIR. Performance is more stable across regions when text-based methods (tfidf, dB and Rt) are used than when interaction-based methods (N2V and DW) are used, given that the availability of interaction data varies across regions. The weakest

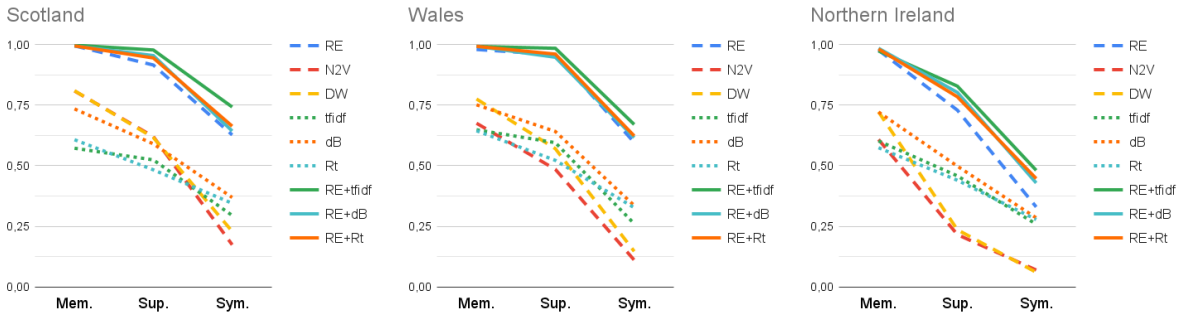


Figure 3: Performance variations for interaction-based approaches (RE, N2V and DW), best text-based approaches (tfidf, dB and Rt) and corresponding HTIM approaches (RE+tfidf, RE+dB and RE+Rt) among different levels of political engagement on SCT (left), WAL (center) and NIR (right) datasets.

overall results occur within the NIR region, not least when interaction-based approaches are used on less engaged users. These weak results are however mitigated through the use of HTIM, especially when used in combination with RE+tfidf, which again proves to be a more robust strategy to be used, both to ensure consistency across regions as well as to better generalize on less engaged users.

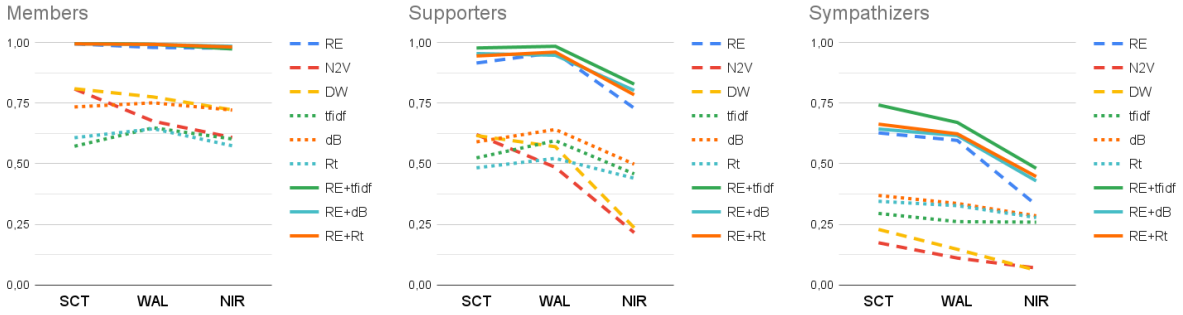


Figure 4: Performance variations for interaction based approaches (RE, N2V and DW), best text based approaches (tfidf, dB and Rt) and corresponding HTIM approaches (RE+tfidf, RE+dB and RE+Rt) among regions for members (left), supporters (center) and sympathizers (right) datasets.

7. Discussion

In this section we will perform a more comprehensive examination of the reported findings. We will do this by conducting an error analysis and creating visual representations of user data for the most effective method, namely, HTIM RE+tfidf.

Error analysis. Interaction-based RE representations have a hard-to-beat performance, but when combined with text representations, they perform even better. Thus, when combining any textual representations with RE representations, they demonstrate comparable or superior performance. Interestingly, transformer-based approaches are not the optimal representations to combine with RE. The most effective representations for combination are tfidf representations. HTIM representation of RE and tfidf (RE+tfidf) yielded significantly improved results, particularly for SCT and NIR Sympathizer users, surpassing the F1-score of standalone RE by more than 10 points. To visualize the improvement, confusion matrices are plotted for these datasets.

RE+tfidf representations achieve a macro-F1 score that is more than 10 points higher than RE representations on their own for SCT Sympathizer users. Further analysis of the confusion

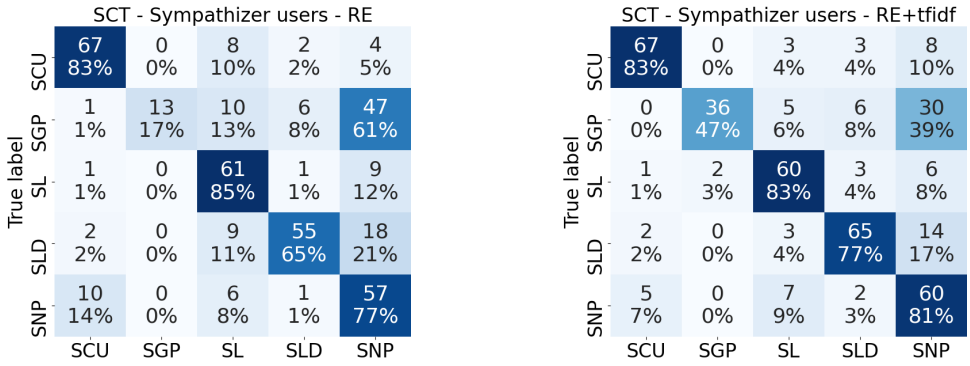


Figure 5: Confusion matrices of SCT Sympathizer users trained with RE (left) or RE+tfidf (right) representations.

matrices for RE and RE+tfidf representations (Figure 5) reveals that users from other parties are often classified as SNP in RE representations, which is understandable considering the SNP's prominence in the region. Specifically, SGP users are consistently misclassified as SNP users. However, when the RE+tfidf model is utilized, the classification becomes more accurate. The RE+tfidf model enhances the accuracy of the classification process, despite certain SGP Sympathizer users remain misclassified as SNP. That misclassification happens between users of two close parties, as both parties have ideological confluences and they govern together in the Scottish Parliament.

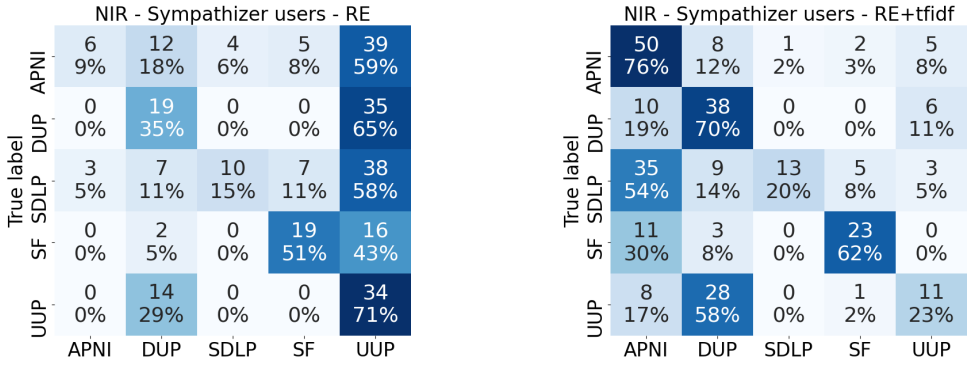


Figure 6: Confusion matrices of NIR Sympathizer users trained with RE (left) or RE+tfidf (right) representations.

In terms of the performance of RE and RE+tfidf representations on NIR Sympathizer users, the inclusion of tfidf text representations results in a significant improvement of over 15 points in the F1 macro score. When examining the confusion matrices for both type of representations (Figure 6), several differences can be observed. The RE representations predominantly classify users as UUP users (Figure 6 left), whereas the RE+tfidf representations primarily classify users as APNI users (Figure 6 right). Moreover, the RE+tfidf representations struggle to identify UUP users as DUP users, which is an issue also present in the RE representations. One similarity between the two representations is that both perform better at classifying SF Sympathizer users compared to users from other parties, while facing difficulties in classifying SDLP Sympathizer users. Specifically, the RE+tfidf representations tend to misclassify SDLP users as APNI users and UUP users as DUP users. These failures can be attributed to the

ideological proximity of these parties, making the RE+tfidf approach more accurate than the RE representations.

Data visualization. Combining RE interaction representations with tfidf textual representations achieves considerably better or similar results across all datasets especially enhancing the results for users that are less attached. In order to better understand and explain the effectiveness of the RE+tfidf HTIM representations, we visualize Member, Supporter and Sympathizer users datasets for the three regions, SCT, WAL and NIR by performing t-SNE dimensionality reduction into 2 dimensions.

SCT *(Figure 7):* In the Members dataset (Figure 7 left), the different users are clearly grouped and defined by party. The SNP (●) and SGP (●) are isolated and separated from the other parties, while the unionist parties SL (●), SLD (●), and SCU (●) are clustered together. The positioning of the pro-independence parties (SNP and SGP) and the unionist parties (SL, SLD, and SCU) in different locations on the chart indicates a high polarization on the issue of national identity. Within the unionist community, the SLD serves as a link between SCU and SL, occupying a central political position. The Supporter dataset chart (Figure 7 center) displays less attached users, but still tight-knit communities can be observed. The representation places the SGP between the SNP and SLD, indicating a more central position of SGP party among supporter users. The representation of Sympathizer users (Figure 7 right) becomes sparser, with less homophily around political parties. The SCU occupies one extreme of the plot, while the SLD and SL take central positions. On the other extreme, the SNP and SGP are mixed, with the SGP closer to central positions. Despite the government treaty and the close alignment between SNP and SGP, the mixture between both parties is evident among Sympathizer users, demonstrating a significant level of closeness as observed in the error analysis.

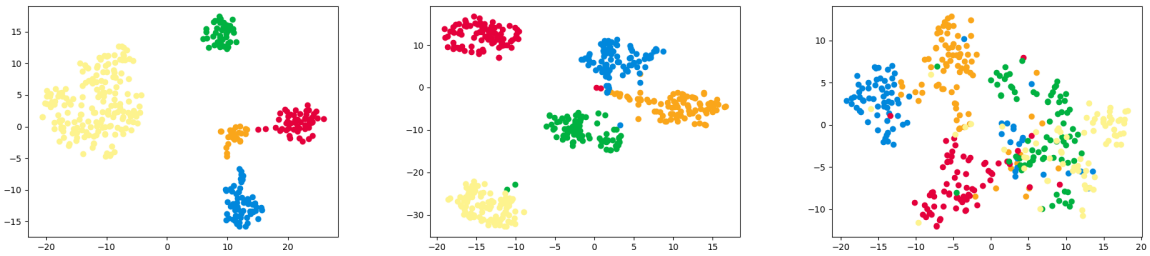


Figure 7: Visualization of t-SNE 2 dimension reduction of RE+tfidf HTIM representations for SCT Member (left), Supporter (center) and Sympathizer (right) user datasets.

WAL *(Figure 8):* Regarding Member users plot (Figure 8, left), the WLD (●) party occupies a central position, as it is considered ideologically situated at the political center. The remaining parties are positioned in the periphery of the diagram, with WL (●) located between PC (●) and WC (●). The visualization of Supporter users (Figure 8 center) shows that pro-independence PC users are isolated while the unionist WLD, WL and WC parties form their own cluster with labour (WL) and liberal (WLD) parties taking central positions. This configuration is similar to the SCT representation, with two prominent stances as pro-independence versus unionist. The representation of Sympathizer users (Figure 8, right) becomes sparser, similar to SCT, losing the clear distinction between the various communities. PC and a combined group of WLD and WC occupy opposite extremes of the plot, while WL appears to remain in the center, highlighting the political similarities from the perspective of sympathizers.

NIR *(Figure 9):* A closer examination of the political party affiliations of Member users

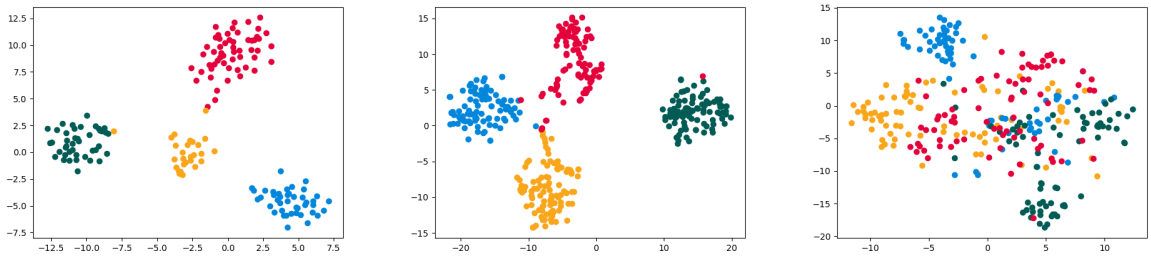


Figure 8: Visualization of t-SNE 2 dimension reduction of RE+tfidf HTIM representations for WAL Member (left), Supporter (center) and Sympathizer (right) user datasets.

(Figure 9 left) reveals that DUP (●), UUP (●) and APNI (●) are positioned each in their own cluster characterized by liberal-conservative, right-wing, and unionist ideologies. On the other side of the graph, SF (●) and SDLP (●) occupy a distinct cluster with left-wing and pro-Irish orientations. When considering the representation of Supporter users (Figure 9 center), significant polarization is observed between SF and the combined group of DUP and UUP, while APNI and SDLP assume more central positions. The visual representation effectively captures the ideological disparities, including the divisions between left and right orientations as well as pro-Irish and unionist perspectives. It is evident that the political parties aligned with unionist and right-wing ideologies are far apart from those with pro-Irish and left-wing leanings. In the representation of Sympathizer users (Figure 9 right), the data becomes too sparse to extract any meaningful information. APNI sympathizers appear scattered throughout the plot, while UUP and DUP sympathizers occupy a similar space, confirming the challenges highlighted in the error analysis.

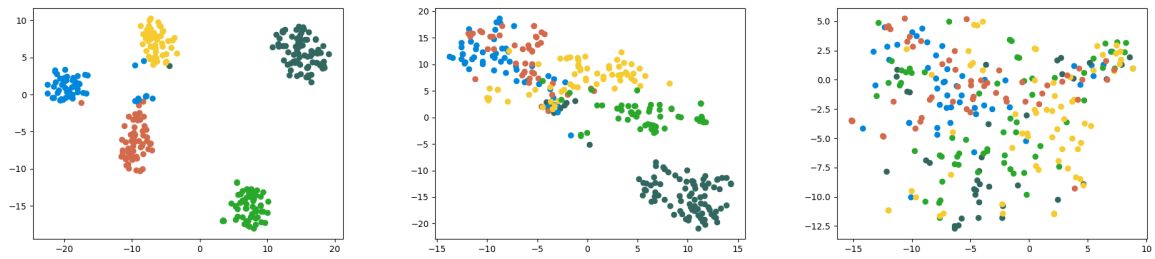


Figure 9: Visualization of t-SNE 2 dimension reduction of RE+tfidf HTIM representations for NIR Member (left), Supporter (center) and Sympathizer (right) user datasets.

While user representations for the Members dataset are clearly grouped, political party communities lose homophily when political attachment decreases and become sparser among Supporter and Sympathizer users datasets. It is noticeable that users not only are grouped depending on their assigned political party, but also that parties are positioned depending on their political similarities specially when communities are less defined. This phenomenon occurs for every annotated political party and for the three selected regions, meaning that the analyzed representations are able to consistently capture political information across different frameworks.

8. Conclusion and Future Work

In this work, we are the first to delve into the ability to predict the political leaning of social media users across different regions with multi-party systems using text and interactions

as a data source. To achieve this, we propose Hybrid Text-Interaction Modeling (HTIM), a framework that enables integrating both data sources into the same model. To perform the experiments, we develop a new dataset spanning three UK regions, where we label users with different levels of political engagement (Members, Supporters and Sympathizers) with respect to the political party they align with.

A look at the performance of each data source individually (i.e. text-based and interaction-based) shows that interactions are more effective in inferring political leaning than text-based approaches. However, these representations tend to underperform when dealing with users who have weaker political engagements. To overcome this limitation, the use of our proposed HTIM to combine RE interaction representations with all the proposed textual representations results in considerable improvements across all datasets, especially with politically less attached users. The results are consistent across the regions and the different levels of political engagement, demonstrating its robustness. All in all, we demonstrate that our proposed HTIM achieves consistently improved performances, with a slight improvement on highly engaged users, but a remarkable improvement with those less engaged.

Considering the improvement obtained from the combination of textual and interaction data, we need to conduct further research to extract hybrid representations. We can experiment with various models using different datasets with missing information to address more realistic scenarios. As the collected datasets include interaction and textual data, we are able to try different configurations. Furthermore, we want to implement the proposed data extraction and user representation techniques in various other tasks, including hate-speech, disinformation or propaganda detection.

References

[1] G. Sartori, Politics, ideology, and belief systems, American Political Science Review 63 (2) (1969) 398–411.

[2] S. Akhtar, V. Basile, V. Patti, A new measure of polarization in the annotation of hate speech, in: AI* IA 2019–Advances in Artificial Intelligence: XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings 18, Springer, 2019, pp. 588–603.

[3] K. Hristakieva, S. Cresci, G. Da San Martino, M. Conti, P. Nakov, The spread of propaganda by coordinated communities on social media, in: Proceedings of the 14th ACM Web Science Conference 2022, 2022, pp. 191–201.

[4] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, A. Flammini, Political Polarization on Twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, 2011.

[5] P. Barberá, G. Rivero, Understanding the political representativeness of twitter users, Social Science Computer Review 33 (2015) 712 – 729.

[6] P. Barberá, Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data, Political Analysis 23 (2015) 76 – 91.

[7] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, R. Bonneau, Tweeting from left to right: Is online political communication more than an echo chamber?, Psychological science 26 (10) (2015) 1531–1542.

[8] K. Darwish, P. Stefanov, M. Aupetit, P. Nakov, Unsupervised user stance detection on twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 14, 2020, pp. 141–152.

[9] V. R. K. Garimella, I. Weber, A long-term analysis of polarization on twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 11, 2017.

[10] F. M. F. Wong, C. W. Tan, S. Sen, M. Chiang, Quantifying political leaning from tweets and retweets, Proceedings of the International AAAI Conference on Web and Social Media (2013).

[11] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media, Proceedings of the National Academy of Sciences 118 (9) (2021) e2023301118.

[12] K. Garimella, G. De Francisci Morales, A. Gionis, M. Mathioudakis, Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship, in: Proceedings of the 2018 World Wide Web Conference, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, p. 913–922.

[13] D. Preotiuc-Pietro, Y. Liu, D. J. Hopkins, L. H. Ungar, Beyond binary labels: Political ideology prediction of twitter users, in: ACL, 2017.

[14] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, K. Karahalios, Quantifying search bias: Investigating sources of bias for political searches in social media, Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (2017).

[15] F. Plà, L. F. Hurtado, Political tendency identification in twitter using sentiment analysis techniques, in: International Conference on Computational Linguistics, 2014.

[16] H. Yan, S. Das, A. Lavoie, S. Li, B. Sinclair, The congressional classification challenge: Domain specificity and partisan intensity, Proceedings of the 2019 ACM Conference on Economics and Computation (2019).

[17] J. A. G.-D. y Salud María Jiménez-Zafra y María-Teresa Martín Valdivia y Francisco García-Sánchez y L. Alfonso Ureña-López y Rafael Valencia-García, Overview of politices 2022: Spanish author profiling for political ideology, Procesamiento del Lenguaje Natural 69 (0) (2022) 265–272.

[18] T. Fagni, S. Cresci, Fine-grained prediction of political leaning on social media with unsupervised deep learning, ArXiv abs/2202.12382 (2022).

[19] A. Rashed, M. Kutlu, K. Darwish, T. Elsayed, C. Bayrak, Embeddings-based clustering for target specific stances: The case of a polarized turkey, in: ICWSM, 2021.

[20] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, F. Menczer, Predicting the political alignment of twitter users, in: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, IEEE, 2011, pp. 192–199.

[21] M. Pennacchiotti, A. M. Popescu, Democrats, republicans and starbucks afficionados: user classification in twitter, in: KDD, 2011.

[22] P. Lahoti, V. R. K. Garimella, A. Gionis, Joint non-negative matrix factorization for learning ideological leaning on twitter, Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (2017).

[23] Y. Hua, T. Ristenpart, M. Naaman, Towards measuring adversarial twitter interactions against candidates in the us midterm elections, in: ICWSM, 2020.

[24] J. Fernandez de Landa, A. Zubiaga, R. Agerri, Generalizing political leaning inference to multi-party systems: Insights from the uk political landscape, arXiv preprint arXiv:2312.01738 (2023).
URL https://arxiv.org/abs/2312.01738

[25] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[27] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, KDD '14, Association for Computing Machinery, 2014, p. 701–710.

[28] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).

[29] J. Fernandez de Landa, R. Agerri, Relational embeddings for language independent stance detection, arXiv e-prints (2022) arXiv–2210.

[30] E. Uberoi, N. Johnston, Political disengagement in the uk: who is disengaged?, House of Commons Library: Research Briefing Number 07501 (2022).

[31] A. Zubiaga, B. Wang, M. Liakata, R. Procter, Political homophily in independence movements: Analyzing and classifying social media users by national identity, IEEE Intelligent Systems 34 (2019) 34–42.

[32] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, SemEval-2016 task 6: Detecting stance in tweets, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 31–41.

[33] M. Hardalov, A. Arora, P. Nakov, I. Augenstein, Cross-domain label-adaptive stance detection, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 9011–9028.

[34] A. T. Cignarella, M. Lai, C. Bosco, V. Patti, P. Rosso, SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), CEUR-WS.org, 2020.

[35] R. Agerri, R. Centeno, M. Espinosa, J. Fernandez de Landa, R. Álvaro, VaxxStance@IberLEF 2021: Overview of the Task on Going Beyond Text in Cross-Lingual Stance Detection, Procesamiento del Lenguaje Natural 67 (2021) 173–181.

[36] F. Ferraccioli, A. Sciandra, M. D. Pont, P. Girardi, D. Solari, L. Finos, TextWiller@SardiStance, HaSpeede2: Text or Context? A smart use of social network data in predicting polarization., in: Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), CEUR Workshop Proceedings, 2020.

[37] R. Alkhalifa, A. Zubiaga, QMUL-SDS@SardiStance: Leveraging Network Inter-actions to Boost Performance on Stance Detection using Knowledge Graphs., in: Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020), CEUR Workshop Proceedings, 2020.

[38] M. Lai, A. T. Cignarella, L. Finos, A. Sciandra, Wordup! at vaxxstance 2021: Combining contextual information with textual and dependency-based syntactic features for stance detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings, 2021.

[39] G. A. Almond, S. Verba, The civic culture: Political attitudes and democracy in five nations, Princeton university press, 2015.

[40] A. F. Ponce, S. E. Scarrow, Party members vs. party sympathizers in a period of declining membership: Who does what (and with whom)?, Party Sympathizers in a Period of Declining Membership: Who Does What (and with Whom) (2013).

[41] A. Makazhanov, D. Rafiei, Predicting political preference of twitter users, Social Network Analysis and Mining 4 (2013) 1–15.

[42] I. R. Hallac, S. Makinist, B. Ay, G. Aydin, user2vec: Social media user representation based on distributed document embeddings, in: 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 2019, pp. 1–5.

[43] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, L. Carin, Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 440–450.

[44] T. Kenter, A. Borisov, M. de Rijke, Siamese CBOW: Optimizing word embeddings for sentence representations, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 941–951.

[45] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[46] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, in: Proceedings of NeurIPS EMC2 Workshop, 2019.

[47] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: ACL, 2020.

[48] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019).

[49] F. Barbieri, L. Espinosa-Anke, J. Camacho-Collados, A multilingual language model toolkit for twitter, arXiv preprint arXiv:2104.12250 (2021).

[50] J. Fernandez de Landa, R. Agerri, Hitz-ixa at politices-iberlef2023: Document and sentence level text representations for demographic characteristics and political ideology detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.

[51] V. Zhelezniak, A. Savkov, A. Shen, F. Moramarco, J. Flann, N. Y. Hammerla, Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors, in: International Conference on Learning Representations, 2019.

[52] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[53] W. Magdy, K. Darwish, N. Abokhodair, A. Rahimi, T. Baldwin, #isisisnotislam or #deportallmuslims? predicting unspoken views, in: Proceedings of the 8th ACM Conference on Web Science, WebSci '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 95–106.

[54] P. Stefanov, K. Darwish, A. Atanasov, P. Nakov, Predicting the topical stance and political leaning of media using tweets, in: ACL, 2020.

[55] T. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, ArXiv abs/1609.02907 (2017).

[56] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio', Y. Bengio, Graph attention networks, ArXiv abs/1710.10903 (2018).

[57] Z. Xiao, W. Song, H. Xu, Z. Ren, Y. Sun, Timme: Twitter ideology-detection via multitask multi-relational embedding, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 2258–2268.

[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.