English Prompts are Better for NLI-based Zero-Shot Emotion Classification than Target-Language Prompts

Patrick Bareiß
University of Stuttgart
Germany
patrick.bareiss@ims.uni-stuttgart.de

Roman Klinger
University of Bamberg
Germany
roman.klinger@uni-bamberg.de

Jeremy Barnes
University of the Basque Country
Spain
jeremy.barnes@ehu.eus

ABSTRACT

Emotion classification in text is a challenging task due to the processes involved when interpreting a textual description of a potential emotion stimulus. In addition, the set of emotion categories is highly domain-specific. For instance, literature analysis might require the use of aesthetic emotions (e.g., finding something beautiful), and social media analysis could benefit from fine-grained sets (e.g., separating anger from annoyance) than only those that represent basic categories as they have been proposed by Paul Ekman (anger, disgust, fear, joy, surprise, sadness). This renders the task an interesting field for zero-shot classifications, in which the label set is not known at model development time. Unfortunately, most resources for emotion analysis are English, and therefore, most studies on emotion analysis have been performed in English, including those that involve prompting language models for text labels. This leaves us with a research gap that we address in this paper: In which language should we prompt for emotion labels on non-English texts? This is particularly of interest when we have access to a multilingual large language model, because we could request labels with English prompts even for non-English data. Our experiments with natural language inference-based language models show that it is consistently better to use English prompts even if the data is in a different language.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

emotion, prompts, cross-linguality, natural language inference

ACM Reference Format:
Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English Prompts are Better for NLI-based Zero-Shot Emotion Classification than Target-Language Prompts. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3589335.3651902

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651902

1 INTRODUCTION

Pretraining large language models (LLMs) on large amounts of text and subsequently fine-tuning them for a specific task constitutes a *de facto* state of the art to address several natural language processing (NLP) tasks, *e.g.*, sentiment analysis [43, 48], question answering [36], or natural language inference [36, 55]. This includes emotion classification, a popular and important task with many datasets from various domains [4, 23, 29, 37, i.a.].

Most work on emotion analysis has been performed in English [see 4], although there has been some work in other languages [7, 40, 44, i.a.]. However, the difficulty and high cost of annotating a large emotion classification dataset means that most languages do not have any resources available. In such a situation, zero-shot cross-lingual methods are of interest.

Driven by the increasing abilities of LLMs to generalize across tasks, recent research has shifted away from fine-tuning models for each new task, instead focusing on zero and few-shot learning [42, 54], and oftentimes reformulating the original tasks as natural language inference (NLI) [5]. This approach enables the use of a language model that has been fine-tuned on an NLI dataset to perform a new task without further tuning the model [38, 39]. This reformulation can be done programatically, creating and filling prompt templates that correspond to NLI premises and hypotheses. Such zero-shot classification can achieve good results [50], including emotion classification [34].

Such NLI-based approach to emotion classification checks if a specific sentence entails information of the classification instance using the prompt. For instance, given a sentence "I won in the lottery", an NLI model shall return a high entailment probability for the prompt "This sentence expresses joy" but a low probability for "This sentence expresses anger". We assume the standard setup for zero-shot classification using NLI, in which the model is not further fine-tuned for emotion classification.

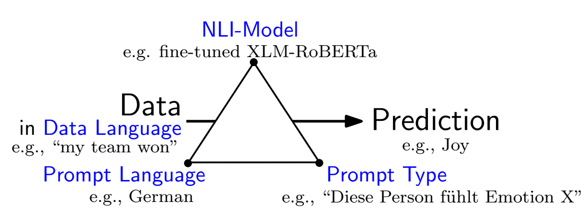

Figure 1: We study the interaction of *data* and *prompt language*, while considering the underlying *NLI-model* and the role of the *prompt type*.

In the established supervised learning regime, obtaining models for a low-resource target-language that is different from the language of the available training data, *i.e.* cross-lingual model induction, has been approached commonly by either 1) transforming the data in some way to create target-language data – oftentimes using translation or label projection, or 2) using model transformations to create a language-agnostic model.

However as many NLI models are inherently multilingual, they can perform a task in a low-resource target-language without additional training when used in a zero-shot manner and thus inducing training data in the target-language or making the model language-agnostic is unnecessary. Instead, the object of focus for cross-lingual transfer shifts to identifying the prompt most optimally suited for classifying data in the target-language. As the prompt does not need to be in the same language as the data are in, one approach is to use an existing known well-performing prompt in a high-resource and well-studied language such as English directly. On the one hand this makes sense as English is commonly the most prevalent language in the training data of multilingual models and is thus likely a prompt written in it will perform well. On the other hand it also appears sensible to match the prompt language with the data language as common multilingual datasets used for training NLI models (such as XNLI [9]) only contain matched examples, *e.g.*, German prompts with German data and thus a mismatch would be out-of-distribution for the training data and potentially results in worse performance. To address this, a well performing prompt in English could be translated to the data language. But then it still remains unclear if the kind of phrasing used to specify the prompt in the original language will be equally as useful in the target language. Especially for emotion classification different words can carry slightly different connotations in different languages. Right now answering these questions of optimal cross-lingual prompt transfer is relatively unexplored for most tasks [53], with no related research available concerning cross-lingual emotion classification.

Therefore this paper aims at answering the following question: *How do we best transfer prompts for zero-shot emotion classification from a high-resource language to a low-resource language?* We study the relation between the *data language* and the *prompt language*, while also analyzing the impact of changes to the *prompt type* (the phrasing of the prompt) and the underlying multilingual NLI *model*. Figure 1 shows a visual representation of this setup. Concretely, we focus on the following research questions:

- **RQ1.** Should we translate the *prompt language* to match the *data language* or leave it in English? *(English is better)*
- **RQ2.** Is the performance of different *prompt types* stable across different *data languages*? *(yes)*
- **RQ3.** How consistent are the results across different NLI *models*? *(they are consistent)*

Our evaluation is based on 3 corpora spanning 18 languages with 7 prompt types [34] and 6 multilingual NLI [8, 22, 41] models.

2 RELATED WORK

2.1 Multilingual Emotion Classification

While much early work on emotion classification in NLP focused on English [1, 26, 28], approaches and datasets to classify emotions in multiple languages, including low-resource ones, have expanded more recently.

Bianchi et al. [3] collect social media emotion data across 19 languages and use it to train an inherently multilingual model. Becker et al. [2] investigates this supervised setting with multiple experiments. Multiple labelled multilingual emotion classification corpora exist for use in this setting, such as Universal Joy [tagged Facebook comments, 21], de/enISEAR [crowd-sourced self-reported event descriptions, 44] or EmoEvent [tweets, 35]. Gupta [14] explores the use of multilingual models in conjunction with unsupervised, adversarial training, *i.e.*, unlabelled data instead.

De Bruyne [10] has pointed out problems with such approaches, e.g., that the concept of an emotion is to some extend dependent on the language and associated culture itself making multilingual approaches inherently more difficult to apply. De Bruyne et al. [11] find evidence for this, suggesting that typologically dissimilar languages in particular utilize language-specific representations for classification in a single multilingual model. Havaldar et al. [15] also investigate this and suggest to work towards better monolingual models as well as culturally balanced corpora for training.

2.2 Prompt-based Learning for Emotion Classification

Prompt-based learning for emotion classification is an attractive alternative to more data-intensive approaches [3]. Plaza-del Arco et al. [34] explore and evaluate a set of prompts extensively across multiple corpora for this reason. Prompt-based approaches can also be used in more complicated settings: Yi et al. [49] propose a prompt-based approach for emotion classification in conversation, a task often difficult for more traditional approaches. They achieve this by first using a language model to extract commonsense features and use those to create a soft prompt then used for actual classification. Another area where prompt-based learning has seen success is in multimodel emotion classification, i.e, where the input consists not only of text but also audio or video. Zhao et al. [52] use a pretrained language model in conjunction with a prompt and combine the resulting embeddings with data from other modalities. Jeong et al. [19] employ something similar but focus only on the combination of text and audio. However, previous work does not evaluate these techniques in a multilingual setting.

2.3 Multilingual and Cross-lingual Prompting

There is only limited work on multilingual prompting, which has, however, shown already some promising results. As an example, Zhao and Schütze [53] explore few-shot cross-lingual NLI and finetune multilingual LLMs with both English and translated prompts, finding that prompting outperformed standard supervised training in few-shot and multilingual scenarios. Fu et al. [13] experiment with multi-task multilingual prompting on a number of tasks (summarization, NER, QA, topic classification, sentiment analysis and NLI). They find that training on larger amounts of available

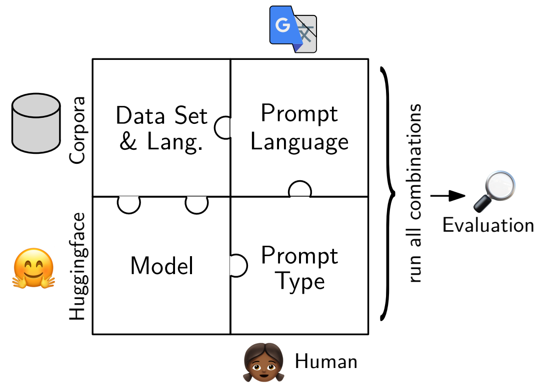

Figure 2: Overview of our experimental setting. We compare models from Huggingface and multiple prompt types for NLI-based emotion classification from Plaza-del Arco et al. [34]. Across them, we study the relation between the data language and the prompt language for 18 languages. To obtain the prompt in various languages, we apply Google Translate. An example setup would be the German subset of the Universal Joy corpus with an XLM-RoBERTa NLI model and the prompt as "This person feels X" translated to German (or left in English).

English datasets leads to benefits for both in-language training, as well as for a cross-lingual zero-shot scenario. They also report that training the models uniformly on English prompts performs better. Huang et al. [18] find that initializing soft prompts with embeddings taken from multilingual LLMs performs better than translation or soft prompting with random initialization. Kim and Komachi [20] concentrate on discovering target-language examples that zero-shot prompting cannot predict. Nie et al. [31] instead propose to retrieve similar source-language examples and use source-language prompting to improve performance on a target language. Finally, Tu et al. [45] show that prompt-tuning multilingual LLMs can outperform fine-tuning in a cross-lingual setting. However, this previous work does not evaluate any approach on emotion analysis.

3 EXPERIMENTAL SETTING

For our experiments we use 6 multilingual NLI models, 3 emotion corpora in 18 languages, and 7 prompt types. All experimentation is performed in a zero-shot setting – using no training or development data. We explain the details in the following section. Figure 2 depicts this setup.

3.1 Data

We use three different emotion corpora which combine multiple languages. The *de/enISEAR* corpora [44] are manually created emotion-triggering event descriptions collected by crowdsourcing. The authors asked workers to describe an event that caused in them a predefined emotion. It consists of 1001 instances for both English and German, respectively, across 6 emotions.

The *Universal Joy (UJ)* corpus [21] stems from Facebook posts in 18 languages (see Table 1 for a list). The motivation for creating this resource was to explore how emotions manifest across languages. We use the predefined test split (containing data for 5 comparatively higher resource languages), downsampled to 981 instances for each of the languages. For the remaining 13 languages (comparatively lower resource languages) there is only one version of the dataset containing all their respective instances. We subsample all of them to a maximum of 981 instances. The data set contains 7 emotion categories.

Table 1: List of languages used by Universal Joy (UJ) and more generally throughout the paper, sorted alphabetically by shorthand.

Shorthand	Name	Shorthand	Name
bn	Bengali	ms	Malay
de	German	my	Burmese
en	English	nl	Dutch
es	Spanish	pt	Portuguese
fr	French	ro	Romanian
hi	Hindi	th	Thai
id	Indonesian	tl	Tagalog
it	Italian	vi	Vietnamese
km	Khmer	zh	Chinese

The *EmoEvent* corpus [35] consists of manually annotated Tweets in Spanish and English. We remove all instances with the emotion labelled as 'other' as well as 12 empty instances. This leads to 792 instances for English and 830 for Spanish across 7 emotions.

3.2 Models

We now describe the details of the 6 NLI models used for our experiments, including which base language model was used and what NLI dataset it was fine-tuned on.

Natural Language Inference Datasets: The NLI datasets we use for fine-tuning are the Multi-Genre Natural Language Inference corpus [MNLI, 47], the Cross-lingual Natural Language Inference corpus [XNLI, 9], the Adversarial Natural Language Inference Dataset [ANLI, 32] and finally the Tasksource dataset [41]. MNLI is a collection of 433k English sentence pairs with entailment information, while XNLI contains 7500 new English test examples following the annotation procedure of Williams et al. [47], and then uses manual translation to 15 languages in order to create a final dataset of 112.5K combined development and testing examples. ANLI is a collection of NLI instances specifically designed to be difficult for state-of-the-art models to solve, while Tasksource is a collection of 500 smaller datasets, including many for NLI.

Model Architectures: We use pretrained multilingual language models that have been fine-tuned on the NLI corpora described above. This allows us to study the effects of *model* and *prompt language* separately. If we instead used monolingual models, these two variable always have to coincide, making it harder to trace where an effect comes from. In order to maximize the generality of our claims, we sample a variety of model architectures for our experiments.

Concretely, we experiment with:

- a XLM-RoBERTa-large model fine-tuned on MNLI & ANLI,

Table 2: A list of the NLI models we use for our experiments. The names are links to the respective HuggingFace models. All of them have either a differing architecture or differing fine-tuning datasets to ensure a diverse sample of different models.

Name	Fine-tuned On	Base Model
XLM-RoBERTa	XNLI/ANLI	XLM-RoBERTa-large
MiniLMv2	XNLI/MNLI	Distilled XLM-RoBERTa-large
Ernie	XNLI/MNLI	RoBERTa
XLM-V	XNLI/MNLI	XLM-V-base
mDeBERTa	XNLI/MNLI	mDeBERTa
mDeBERTa-TS	Tasksource	mDeBERTa (v3)

- a distilled version of XLM-RoBERTa-large (MiniLMv2, Wang et al. 46 fine-tuned on MNLI and XNLI,
- Ernie [51] fine-tuned on MNLI and XNLI,
- XLM-V [24] fine-tuned on MNLI and XNLI,
- mDeBERTa [16, 17] fine-tuned on MNLI and XNLI,
- and mDeBERTa-TS, which has been fine-tuned on the Tasksource dataset [41].

We take the models from the Huggingface Hub[1] with all but XLM-RoBERTa and mDeBERTa-TS being introduced by Laurer et al. [22]. The information on each *model* can be found in Table 2.

3.3 Prompt Types

To use NLI models in a zero-shot manner, we encode the data point we want to classify as the premise and each of the possible labels (in our case emotions) as the hypothesis and then choose the label with the highest probability of being entailed by the premise.

To represent the labels, we use seven (of eight total[2]) *prompt types* proposed by Plaza-del Arco et al. [34]. We define a prompt as a mapping from the input text x and emotion label e to a template T, where T can be:

$T_{\text{Emo-Name}}$ x: e
$T_{\text{Expr-Emo}}$ x: This text expresses e
$T_{\text{Feels-Emo}}$ x: This person feels e
$T_{\text{WN-Def}}$ x: This person expresses wn(e)
$T_{\text{Emo-S}}$ x: $syn(e)$
$T_{\text{Expr-S}}$ x: This text expresses syn(e)
$T_{\text{Feels-S}}$ x: This person feels syn(e)

where wn(e) is a function that maps an emotion to its WordNet definition [27] and syn(e) is a function that maps an emotion to 6 predefined synonyms. For the all prompt templates that use syn(e), we run the model on all 6 prompts and take average entailment probability as the final prediction.

We extend the prompts provided by Plaza-del Arco et al. [34] to cover the labels *anticipation* in UJ and *surprise* in EmoEvent, and add 6 manually created synonyms for each.

We use Google Translate to obtain prompts in the 18 languages of our data. Table 5 shows some examples. We performed a manual analysis of the prompts in a subset of the languages (German, Spanish) and confirm that the quality of translation is generally high.

[1] https://huggingface.co/models
[2] The original paper additionally uses a prompt that uses all synonyms for a particular emotion from the Emolex dictionary [30]. We omit this prompt due to computational constraints.

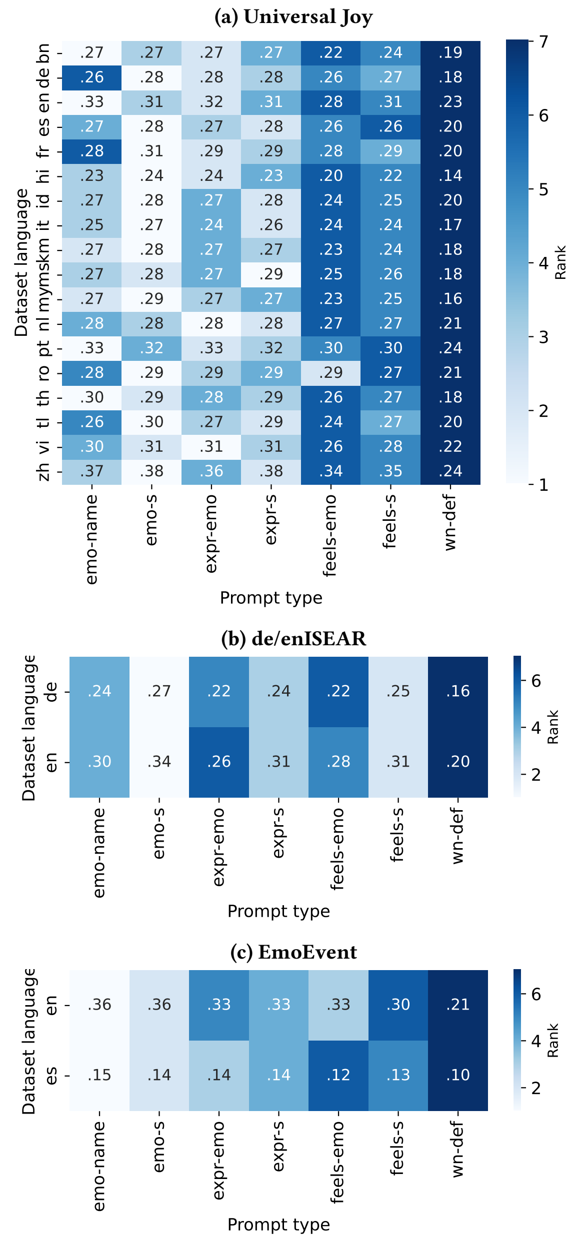

Figure 3: Interaction of *prompt types* and *data languages*. Each cell contains the average F_1 across NLI *models*. The prompt is always in English. The color corresponds to the rank and therefore indicates consistency of the results.

3.4 Controlling for Variables of Interest

Although it is in principle interesting to evaluate all possible combinations of the four variables *model*, *data language*, *prompt language* and *prompt type*, due to practical limitations, we restrict the *prompt*

Table 3: Comparison (macro-F_1 across emotion categories) of the performance of using the English prompt for emotion classification or a translation to the data language (RQ1). The various scores are averages across *prompt types* and NLI *models*. EmoE: EmoEvent; de/enIS: en/deISEAR.

	Dataset language																					
	Universal Joy																		EmoE		en/deIS	
Prompt lang.	bn	de	en	es	fr	hi	id	it	km	ms	my	nl	pt	ro	th	tl	vi	zh	en	es	de	en
English	**25**	**26**	**30**	**26**	**28**	**21**	**26**	**24**	**25**	**26**	**25**	**27**	**31**	**27**	**27**	**26**	**28**	**34**	**32**	**13**	**23**	**29**
Translated	22	24	—	24	26	19	24	23	23	25	19	24	28	26	25	20	27	31	—	13	22	—

language to English and the translated target data language. This restriction is motivated by the fact that English is well-represented in all training sets of the *models* we test. By matching the *prompt language* and *data language* via machine translation, on the other hand, we capture a common use case in NLP. In total, we evaluate 1470 combinations for UJ and 126 for both de/enISEAR and EmoEvent this way.

4 RESULTS

Overall, models perform within the expected range for zero-shot classification with a larger number of labels. Macro F_1 scores run from 0.03–0.5, depending on the combination of model, prompt type, and language. We therefore set out to answer the research question posed in the introduction.

4.1 RQ1: Should we translate the *prompt language* to match the *data language* or leave it in English?

Overview. Multilingual NLI models can process prompts in either English or the target language. It is reasonable to assume that the performance would be higher if the data and prompt languages are the same. Here we test this hypothesis.

Results. Table 3 shows the results of all models on the three emotion corpora. The rows correspond to the prompt language (English or translated to the data language) and the columns show the data language. We report the macro F_1 scores for each emotion classification setting, averaging over *models*, *prompt types*, and *emotion label* for each target language in the three data sets.

For some data sets and languages, the performance is lower than for others, which we interpret as a varying difficulty of the respective data sets. More interestingly for our RQ is that the English prompt performance outperforms the target language prompts in all cases of the Universal Joy Data Set (average F_1 difference of 0.025). For EmoEvent, the performance is roughly the same, while for de/enISEAR, there is only a minor difference for the English–German pair (of 0.013).

We therefore conclude that it is generally better or equally beneficial to use an English prompt for performing emotion classification in a target language. This observation is in line with previous work [12, 18, 53], which finds that translating a prompt to a target language for other tasks has no benefit and often directly harms model performance.

4.2 RQ2: Is the performance of different *prompt types* stable across different *data languages*?

Overview. Small variations to a prompt can lead to a drastic change in classification performance [25, 34]. Therefore, we ask if if there is any concrete prompt type that performs particularly well or poorly across all languages. Or instead, is the choice of prompt type to use for emotion classification tied tightly to the target language?

Results. We show the results in Figure 3 for the three datasets. Each cell in Figure 3 shows an average across *models* for a combination of a *prompt type* (x-axis) and a *data language* (y-axis). The color in the heatmaps represents the rank of each prompt type, *i.e.*, the rank of the average performance for each prompt type compared to the other 6 (for a given row, *i.e.*, *data language*). Given the results of RQ1 above, we fix the *prompt language* for this heatmap to be English.

Figure 3 indicates that the best performing *prompt types* are consistent across target languages. The best-performing prompt for English data on UJ (emo-name) is also in the top-3 *prompt types* for 11 other languages. The best overall *prompt type* for other target languages, however, is emo-s, which achieves the top ranking results in 11 languages. Wn-def is consistently the worst performing *prompt type*, followed by feels-emo and feels-s. The results on EmoEvent and de/enISEAR are comparable to UJ.

To quantify the consistency across prompts in Figure 3, we calculate the average Kendall's τ between all pairs of rows. The correlation of different *prompt types* between the languages is .64 for UJ, .9 for de/enISEAR, and .62 for EmoEvent.

We conclude that there is a strong relation between the performance of a prompt in English and the target languages. Therefore, we expect a good prompt for English data to be good for data in other languages. Similarly, we observe that prompt templates that ask the model to estimate what a concrete actor is feeling (feels-emo, feels-s) generally perform worse than others.

4.3 RQ3: How consistent are the results across different NLI *models*?

Overview. The NLI models we use vary in number of parameters, size, variety of pretraining data, and NLI-datasets used for fine-tuning. Therefore, we explore whether the effects found for RQ1 and RQ2 generally hold across models. More specifically, we

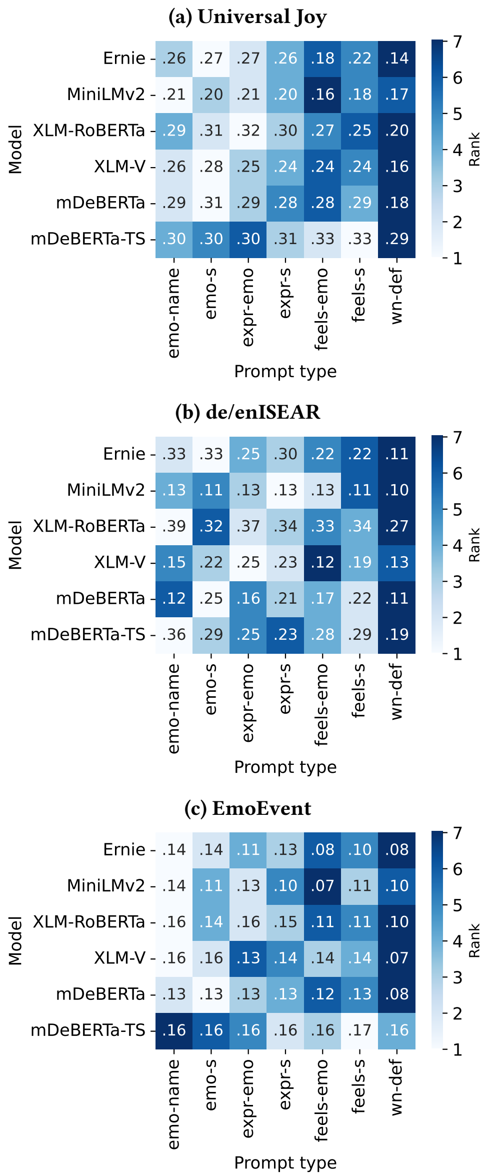

Figure 4: Interaction of *NLI-models* and *prompt types.* Cells are macro-average F_1 scores across *prompt and data languages.* English data is omitted, as we are interested in the results on the target languages.

study whether the results for prompt language or prompt type vary particularly for specific models.

Results – Prompt type performance across models. Figure 4 shows the relation between *model* and the *prompt type.* Each cell in the heatmaps shows an average across *models* for a combination of a *prompt type* (x-axis) and a *model* (y-axis). We are interested in the performance consistency on low-resource languages and therefore exclude English. Similarly to the results above, the rank shows the consistency of the performance of a prompt type across models.

Table 4: Performance in macro-F_1 across emotion categories for the *models* and *prompt languages* in the Universal Joy data set. Each cell represents an average across *prompt types* and *data languages.* We average over the *data languages.* English is omitted as we are mainly interested in consistency on low-resource languages.

	Prompt language					
	UJ		de/enISEAR		EmoEvent	
Model	en	transl.	en	transl.	en	transl.
Ernie	**.25**	.21	**.26**	.24	.11	**.11**
mDeBERTa	**.29**	.26	.17	**.18**	**.12**	.12
mDeBERTa-TS	**.31**	.30	.26	**.28**	**.16**	.16
MiniLMv2	**.21**	.17	**.12**	.11	**.11**	.10
XLM-RoBERTa	**.28**	.27	**.36**	.32	**.14**	.12
XLM-V	**.25**	.23	**.20**	.16	**.13**	.13

We see a high consistency across models, with the exception of mDeBERTa-TS. For most models, either emo-name or emo-s are the best performing prompt types, while WN-def has the lowest or second lowest performance across all models. The average correlations of the performances for the prompt types across models is lower than across for languages with 0.4 on UJ, 0.18 for de/enISEAR, and 0.23 for EmoEvent.

This is mostly due to the outlier mDeBERTa-TS. Omitting this last row in the heatmaps from the correlation calculations leads to 0.7 on UJ, 0.52 for EmoEvent and 0.22 for de/enISEAR. We presume that this is attributable to the use of the Tasksource dataset [41], which is specific to this model.

Therefore, we conclude that the finding of RQ2 holds consistently on the majority of models.

Results – Prompt language performance across models. Finally, we show the results for both English and the translated prompts across languages for all data sets in Table 4. For all models, leaving the prompt untranslated performs better on UJ and for the majority of models on en/deISEAR and EmoEvent (4 out of 6 for both cases), strengthening our results from RQ1.

Overall these results indicate that our findings on the superior performance of English prompts from RQ1 are consistent across models.

5 ANALYSIS

To provide an intuition of the results, we show prompts with predictions in Table 5. We acknowledge that these results are too few to gain any particular generalizable observations, but hope that they still provide a better idea about how our methods work and the results were obtained.

The top part of the table shows instances in which the English and the translated prompt leads to the same predictions. Most instances contain event descriptions that are clearly connotated with an emotion. Becoming father (Example 1) is predominantly related to joy and both the English and the German model infer this emotion to be most appropriate. Similarly clear is the assignment of shame for the event of sweating (Example 2). In Examples 2, 3, and

Table 5: Examples of predictions with English and German prompts. The model is XLM-RoBERTa, the data is the German portion of de/enISEAR. The prompt is expr-emo. Correct predictions are printed in bold. The top part of the table shows examples where both the English and the German prompt lead to the same result, while the predictions differ in the bottom part.

	Sentence	True Label	English Prompt Pred.	German Prompt Pred.
1.	Ich fühlte ..., als ich Vater wurde. (I felt ... when I became a father.)	Joy	**Joy** (Prompt: This text expresses joy)	**Joy** (Prompt: Dieser Text drückt Freude aus)
2.	Ich fühlte ..., weil ich zu dick bin (I felt ... because I am too fat.)	Shame	Guilt (Prompt: This text expresses guilt)	Guilt (Prompt: Dieser Text drückt Schuld aus)
3.	Ich fühlte ..., als ein Onkel starb. (I felt ... when an uncle died.)	Fear	Sadness (Prompt: This text expresses sadness)	Sadness (Prompt: Dieser Text drückt Trauer aus)
4.	Ich fühlte ..., als ich absagen musste (I felt ... when I had to cancel.)	Sadness	Shame (Prompt: This text expresses shame)	Shame (Prompt: Dieser Text drückt Scham aus)
5.	Ich fühlte ..., als ich geschwitzt habe (I felt ... when I sweated.)	Shame	**Shame** (Prompt: This text expresses shame)	**Shame** (Prompt: Dieser Text drückt Scham aus)
6.	Ich fühlte ..., als mein Hund krank war. (I felt ... when my dog was sick.)	Fear	**Fear** (Prompt: This text expresses fear)	Sadness (Prompt: Dieser Text drückt Trauer aus)
7.	Ich fühlte ..., als ich befördert wurde. (I felt ... when I got promoted.)	Joy	Shame (Prompt: This text expresses shame)	**Joy** (Prompt: Dieser Text drückt Freude aus)
8.	Ich fühlte ..., als der Urlaub vorbei war. (I felt ... when the vacation was over.)	Sadness	**Sadness** (Prompt: This text expresses sadness)	Joy (Prompt: Dieser Text drückt Freude aus)
9.	Ich fühlte ..., als ich vor ihrem Grab stand (I felt ... standing in front of her grave.)	Sadness	Shame (Prompt: This text expresses shame)	**Sadness** (Prompt: Dieser Text drückt Trauer aus)
10.	Ich fühlte ..., als ich schwer erkältet war. (I felt ... when I was severely cold.)	Sadness	Fear (Prompt: This text expresses fear)	**Sadness** (Prompt: Dieser Text drückt Trauer aus)

4, one might argue that both labels are correct and the predicted labels are acceptable labels for the text.

The lower part of the table shows instances in which the labels inferred by the English and the translated prompt differ. As often the case for prompt-based predictions, it is difficult to infer any patterns from these instances. In Example 6 (description of a sick dog), both fear (English prompt) and shame (German prompt) are reasonable assignments. In Example 8 (vacations being over), the German prompt is more prone to spurious correlations to the associations of vacations with joy than the English prompt. English 7 (being promoted) and Example 9 (standing in front of a person's grave) are challenging to interpret – the labels predicted by the English prompt make no sense compared to the German, data language, prompts.

We observe that there are indeed cases in which the data language prompts outperform the English prompts, but there are also cases in which the English prompts are less sensitive to potential biases in underlying data. While these observations are hard to generalize, given the few instances, they motivate future research which we will mention in the next section.

6 CONCLUSION AND FUTURE WORK

With this paper, we studied if English prompts for emotion classification work well across various data languages and if the results are robust to changes of the underlying language model and reformulations of the prompt. We found that generally English prompts outperform the prompts in the respective data languages, and except for one underlying model, they hold robustly across them.

Our main results support previous work that multilingual language models often perform better on a task when the prompt is kept in English, even for target languages that are typologically far from English [12, 18, 53]. This suggests that multilingual models have an inherent bias towards English, no matter what the target language is.

There are two exceptions to this general observations. First of all, we only had one underlying language model that has been fine-tuned on different NLI data. This model showed differing results for some prompt types and therefore this variation on the setup requires more future attention. It is important to better understand how the training data of the language model and the prompt interact, and particularly how this affects the transferability of prompts across languages.

Secondly, we saw in the analysis that some instances do show more reasonable results for target language prompts. While, overall, this does not justify the use of target language prompts, understanding better what such instances have in common might help to improve the development of languages in other languages than English. This is important for the majority of people who want to use multilingual language models interactively but do not have a sufficient command of English.

Further, we did focus on the setup in which the language model is fixed and only the prompts receive variations. It may be assumed that slightly adapting the language model to perform similarly on a target language as it does on English could change the overall results and enable other language prompts to perform comparably well. This required approaches of cross-lingual model alignment under consideration of specific prompts – a research task that we are not aware received any attention yet.

Additionally, in this paper, we concentrated on prompting for emotion classification, where we predict a single label for each text. However, emotion labels are not mutually exclusive. Therefore, future work needs to also consider prompting for multilabel emotion classification [33]. While a simple conversion of single labels to binary predictions would likely lead to comparable results, models that can exploit label relations might behave differently.

Finally and more broadly, future work could benefit from the exploration of prompt-based cross-lingual transfer for less restrictive styles of prompting as compared to ones based on NLI. For instance, prompting-based on next-token prediction of autoregressive languagel models like GPT-3 [6] allows the specification of (1) task instructions as well as (2) few-shot examples, which is not easily possible for NLI-based prompting. The impact of these features when choosing a prompt for cross-lingual transfer is not well understood and will certainly benefit from additional work.

REFERENCES

[1] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from Text: Machine Learning for Text-based Emotion Prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Vancouver, British Columbia, Canada, 579–586. https://aclanthology.org/H05-1073

[2] Karin Becker, Viviane P Moreira, and Aline GL dos Santos. 2017. Multilingual emotion classification using supervised learning: Comparative experiments. *Information Processing & Management* 53, 3 (2017), 684–704.

[3] Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. XLM-EMO: Multilingual Emotion Prediction in Social Media Text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Association for Computational Linguistics, Dublin, Ireland, 195–203. https://doi.org/10.18653/v1/2022.wassa-1.18

[4] Laura-Ana-Maria Bostan and Roman Klinger. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2104–2119. https://aclanthology.org/C18-1179

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 632–642. https://doi.org/10.18653/v1/D15-1075

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Xiyao Cheng, Ying Chen, Bixiao Cheng, Shoushan Li, and Guodong Zhou. 2017. An emotion cause corpus for chinese microblogs with multiple-user structures. *ACM Transactions on Asian and Low-Resource Language Information Processing* 17, 1 (2017), 6:1–6:19. https://dl.acm.org/citation.cfm?id=3132684

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[9] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2475–2485. https://doi.org/10.18653/v1/D18-1269

[10] Luna De Bruyne. 2023. The Paradox of Multilingual Emotion Detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Association for Computational Linguistics, Toronto, Canada, 458–466. https://doi.org/10.18653/v1/2023.wassa-1.40

[11] Luna De Bruyne, Pranaydeep Singh, Orphee De Clercq, Els Lefever, and Veronique Hoste. 2022. How Language-Dependent is Emotion Detection? Evidence from Multilingual BERT. In *Proceedings of the The 2nd Workshop on Multilingual Representation Learning (MRL)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 76–85. https://doi.org/10.18653/v1/2022.mrl-1.7

[12] Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do Multilingual Language Models Think Better in English? arXiv:2308.01223 [cs.CL]

[13] Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot Prompt: Multilingual Multitask Prompt Training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9919–9935. https://doi.org/10.18653/v1/2022.emnlp-main.674

[14] Vikram Gupta. 2021. Multilingual and Multilabel Emotion Recognition using Virtual Adversarial Training. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 74–85. https://doi.org/10.18653/v1/2021.mrl-1.7

[15] Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual Language Models are not Multicultural: A Case Study in Emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Association for Computational Linguistics, Toronto, Canada, 202–214. https://doi.org/10.18653/v1/2023.wassa-1.19

[16] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv:2111.09543 [cs.CL]

[17] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*. https://openreview.net/forum?id=XPZIaotutsD

[18] Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot Cross-lingual Transfer of Prompt-based Tuning with a Unified Multilingual Prompt. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 11488–11497. https://doi.org/10.18653/v1/2022.emnlp-main.790

[19] Eunseo Jeong, Gyunyeop Kim, and Sangwoo Kang. 2023. Multimodal Prompt Learning in Emotion Recognition Using Context and Audio Information. *Mathematics* 11, 13 (2023), 2908.

[20] Hwichan Kim and Mamoru Komachi. 2023. Enhancing Few-shot Cross-lingual Transfer with Target Language Peculiar Examples. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 747–767. https://doi.org/10.18653/v1/2023.findings-acl.47

[21] Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal Joy A Data Set and Results for Classifying Emotions Across Languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Online, 62–75. https://aclanthology.org/2021.wassa-1.7

[22] Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis* (2022), 1–33. https://doi.org/10.1017/pan.2023.20

[23] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 986–995. https://aclanthology.org/I17-1099

[24] Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models. *arXiv e-prints*, Article arXiv:2301.10472 (Jan. 2023), arXiv:2301.10472 pages. https://doi.org/10.48550/arXiv.2301.10472 arXiv:2301.10472 [cs.CL]

[25] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9, Article 195 (jan 2023), 35 pages. https://doi.org/10.1145/3560815

[26] Rada Mihalcea and Carlo Strapparava. 2012. Lyrics, Music, and Emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, 590–599. https://www.aclweb.org/anthology/D12-1054

[27] George A. Miller. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. https://aclanthology.org/H94-1111

[28] Saif Mohammad. 2012. #Emotional Tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Association for Computational Linguistics, Montréal, Canada, 246–255. https://www.aclweb.org/anthology/S12-1033

[29] Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion Intensities in Tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*. Association for Computational Linguistics, Vancouver, Canada, 65–77. https://doi.org/10.18653/v1/S17-1007

[30] Saif Mohammad and Peter D. Turney. 2012. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29, 3 (2012). https://doi.org/10.1111/j.1467-8640.2012.00460.x

[31] Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-Lingual Retrieval Augmented Prompt for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 8320–8340. https://doi.org/10.18653/v1/2023.findings-acl.528

[32] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4885–4901. https://doi.org/10.18653/v1/2020.acl-main.441

[33] Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6542–6552. https://doi.org/10.18653/v1/2020.coling-main.575

[34] Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 6805–6817. https://aclanthology.org/2022.coling-1.592

[35] Flor Miriam Plaza del Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin. 2020. EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1492–1498. https://aclanthology.org/2020.lrec-1.186

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 1, Article 140 (jan 2020), 67 pages.

[37] K R Scherer and H G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology* 66, 2 (Feb. 1994), 310–328. https://doi.org/10.1037//0022-3514.66.2.310

[38] Timo Schick and Hinrich Schütze. 2021. Few-Shot Text Generation with Natural Language Instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 390–402. https://doi.org/10.18653/v1/2021.emnlp-main.32

[39] Timo Schick and Hinrich Schütze. 2021. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2339–2352. https://doi.org/10.18653/v1/2021.naacl-main.185

[40] Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021. Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics, Punta Cana, Dominican Republic (online), 67–79. https://doi.org/10.18653/v1/2021.latechclfl-1.8

[41] Damien Sileo. 2023. tasksource: A Dataset Harmonization Framework for Streamlined NLP Multi-Task Learning and Evaluation. arXiv:2301.05948 [cs.CL]

[42] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *Comput. Surveys* (2023). https://doi.org/10.1145/3582688

[43] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification?. In *Chinese Computational Linguistics*, Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu (Eds.). Springer International Publishing, Cham, 194–206. https://doi.org/10.1007/978-3-030-32381-3_16

[44] Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and Validating Event-focused Emotion Corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4005–4011. https://doi.org/10.18653/v1/P19-1391

[45] Lifu Tu, Caiming Xiong, and Yingbo Zhou. 2022. Prompt-Tuning Can Be Much Better Than Fine-Tuning on Cross-lingual Understanding With Multilingual Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5478–5485. https://doi.org/10.18653/v1/2022.findings-emnlp.401

[46] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 2140–2151. https://doi.org/10.18653/v1/2021.findings-acl.188

[47] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

[48] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[49] Jingjie Yi, Deqing Yang, Siyu Yuan, Kaiyan Cao, Zhiyao Zhang, and Yanghua Xiao. 2022. Contextual Information and Commonsense Based Prompt for Emotion Recognition in Conversation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 707–723. https://doi.org/10.1007/978-3-031-26390-3_41

[50] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3914–3923. https://doi.org/10.18653/v1/D19-1404

[51] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1441–1451. https://doi.org/10.18653/v1/P19-1139

[52] Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4703–4707. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9746910

[53] Mengjie Zhao and Hinrich Schütze. 2021. Discrete and Soft Prompting for Multilingual Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 8547–8555. https://doi.org/10.18653/v1/2021.emnlp-main.672

[54] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL]

[55] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, Huhhot, China, 1218–1227. https://aclanthology.org/2021.ccl-1.108