Sentiment and Emotion Classification in Low-resource Settings

Jeremy Barnes
HiTZ Basque Center for Language Technologies
Ixa NLP Group, University of the Basque Country UPV/EHU
jeremy.barnes@ehu.eus

Abstract

The popularity of sentiment and emotion analysis has lead to an explosion of datasets, approaches, and papers. However, these are often tested in optimal settings, where plentiful training and development data are available, and compared mainly with recent state-of-the-art models that have been similarly evaluated.

In this paper, we instead present a systematic comparison of sentiment and emotion classification methods, ranging from rule- and dictionary-based methods to recently proposed few-shot and prompting methods with large language models. We test these methods in-domain, out-of-domain, and in cross-lingual settings and find that in low-resource settings, rule- and dictionary-based methods perform as well or better than few-shot and prompting methods, especially for emotion classification. Zero-shot cross-lingual approaches, however, still outperform in-language dictionary induction.

1 Introduction

Affective computing, including sentiment and emotion classification, has been research focuses inside of the Natural Language Processing (NLP) community for many years (Mohammad, 2016; Poria et al., 2023). This has lead to an incredible number of research directions and papers published on these topics, ranging from rule-based and dictionary-based approaches Turney (2002); Lee et al. (2010); Taboada et al. (2011); Staiano and Guerini (2014), to supervised training of deep learning models (Xu et al., 2019; Barbieri et al., 2022; Samuel et al., 2022) and finally to few-shot and prompting of large language models (Brown et al., 2020; Min et al., 2022; Plaza-del Arco et al., 2022). This also means that a systematic comparison of the benefits and weaknesses of models has not been performed, as each often individual papers compare only against more recent state-of-the-art models, and do not take into account previous approaches.

Like many other research areas in NLP, sentiment and emotion classification are dependent on domain and language-specific training data for optimal performance and this high-quality task-specific data is always in short supply as we apply our models to a constantly evolving set of scenarios.

The objective of this paper is therefore to identify trends in sentiment and emotion classification, especially regarding low-resource settings. As such, we attempt to address the following research questions:

- **RQ1**: Given a limited number of examples per class (<100), what method currently performs best?
- **RQ2**: Do these methods suffer domain transfer equally?
- **RQ3**: How well do these results hold for languages other than English?

To address these questions, we perform experiments[1] on 10 sentiment classification datasets and two emotion classification datasets in 8 languages with a number of low-resource approaches. Specifically, we compare dictionary-based methods, rule-based methods, few-shot methods and prompting methods on the English datasets. We simultaneously test the out-of-domain performance for each of the methods that demand training data. Finally, we also perform cross-lingual experiments.

We find that rule- and dictionary-based methods often perform on par with few-shot approaches in low-resource settings, especially on emotion classification and are more robust to domain changes, while prompting similarly provides promising results. Zero-shot cross-lingual approaches, however, still outperform in-language dictionary induction for languages other than English, suggesting that more work could be done in this area.

[1]Code to reproduce the experiments available at https://github.com/jerbarnes/low_resource_sa_emo.

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 290–304

2 Related Work

Current state-of-the-art models for sentiment and emotion classification are dominated by language models that have been pretrained on large corpora and then fine-tuned for each specific task (Sharma et al., 2020; Barnes et al., 2022). Although ELMO (Peters et al., 2018), BERT (Devlin et al., 2019), and its variants were the first to provide evidence for the usefulness of language modeling as a transfer learning objective, there has since been an explosion and it is somewhat difficult to navigate which current models give the best performance on many datasets.

Besides the fully supervised setup, many larger language models also show signs of being able to learn a task with less data, allowing for non-trivial zero- or few-shot performance. The most common way to achieve this zero or few-shot ability is by prompting a model using a Natural Language Inference model, trained to determine whether a premise is true/false, given a hypothesis. This model can then be applied to new tasks by reformulating the input and labels (Schick and Schütze, 2021; Min et al., 2022).

For few-shot prompting, we can make use of the generative abilities of language models by providing demonstrations input/label pairs and asking for a final label (Brown et al., 2020; Lin et al., 2022). More recently, the results of models trained using instruction tuning suggests that these models generalize well to unseen tasks (Chung et al., 2022).

The same kinds of large language models trained on multilingual corpora also allow for *zero-shot cross-lingual transfer*, where a model is fine-tuned on a task in a high-resource source language and then tested on an under-resourced language (Pires et al., 2019; Conneau et al., 2020). However, these approaches have rarely been compared to previous dictionary-based methods.

2.1 Rule and Dictionary-based methods

Rule and dictionary-based methods are common for sentiment and emotion analysis, in part due to their simplicity and interpretability. Early work focused on automatically inferring polarity dictionaries for categorizing words (Hatzivassiloglou and McKeown, 1997) or texts (Turney and Littman, 2003; Kamps et al., 2004). Taboada et al. (2011) propose *SoCal*, one of the most popular rule-based methods for sentiment analysis, which uses a set of dictionaries with sentiment scores for certain parts of speech (adjectives, adverbs, nouns, intensifiers, and verbs) plus rules for interacting with negation, irrealis, and other sentiment shifting phenomena.

For emotion classification, there has been a good deal of work on creating dictionaries. Mohammad and Kiritchenko (2015) use word-association measures with emotional hashtags to create a large emotion dictionary from social media text, while Mohammad (2018) instead use best-worst scaling to crowdsource an emotion intensity dictionary. Buechel et al. (2016) adapt affective lexicons to historical German texts and use these to characterize emotional trends in various genres of writing across several centuries. Buechel et al. (2020) furthermore develop methods for inducing emotion dictionaries for 91 languages, but do not make use of these dictionaries for emotion classification.

For dictionary induction, Hamilton et al. (2016) propose a method to automatically induce domain-specific dictionaries and show their effectiveness across a number of historical and modern text classification tasks. An et al. (2018) similarly propose a method to create a semantic axis, *SemAxis*, in an embedding space and successfully create dictionaries for tasks beyond sentiment analysis, despite having small amounts of data available. In this approach, we create an average vector for positive V^+ and negative V^- sentiment by averaging the vectors for seed words from an embedding space, such as Word2Vec or FastText. We can then define the axis vector as the difference of the two:

$$V_{axis} = V^+ - V^-$$

To use the semantic axis that we have created, we can measure the cosine distance of another embedding and the semantic axis.

$$score(w)V_{axis} = \text{cosine dist}(w, V_{axis})$$

If the score is positive, we can assume the word is positive and vice versa, and expand the positive and negative seed dictionaries to cover all lemmas in the test set, effectively creating a high-coverage dictionary. We then use this dictionary to generate the semantic orientation score of a text.

However, most of these techniques have not been recently compared to what are considered state-of-the-art models under low-resource settings.

	Dataset	lang	Train	Dev	Test
Sentiment	MPQA	EN	987	337	299
	SemEval	EN	3,737	413	1,791
	OpeNER	EN	1,210	174	347
	OpeNER	ES	1,029	147	296
	GermEval	DE	6,444	772	1,490
	ASTD	AR	2,468	353	706
	NoReC	NO	2,675	516	417
	MultiBooked	EU	789	113	227
	NArabizi	DZ	564	75	92
	Maltese	MT	595	85	171
Emotion	SSEC	EN	2,329	583	1,956
	EnISEAR	EN	720	80	201

Table 1: Statistics regarding the sentiment and emotion datasets.

3 Data

In this section we describe the datasets that are used for experimentation. The statistics are shown in Table 1 (see Tables 6 and 7 in the Appendix for further details).

Sentiment datasets As we want to explore how well methods work across a number of domains and languages, we choose to explore binary sentiment classification. We use the binary version of the following datasets, where any strong positive/negative has been mapped to positive/negative and neutral has been removed. Using only binary sentiment classification allows for us to compare across a larger number of datasets and languages.

MPQA: Wiebe et al. (2005) annotate English news wire texts with a complex set of annotation types. We map the polarities to sentences and keep those sentences that contain a majority of one polarity, such that we have only positive and negative sentence-level annotations.

SemEval: The SemEval 2013 Shared Task 2 (Nakov et al., 2013) collected tweets and annotated them as positive, negative, or neutral. We keep only the positive and negative tweets.

OpeNER: Agerri et al. (2013) annotate English and Spanish (among others) hotel reviews for structured and aspect-based sentiment. We use the script from Barnes et al. (2018) to map these to sentence-level binary sentiment classification. **ES** is the Spanish data from this dataset.

AR: Nabil et al. (2015) annotate Arabic (both Modern Standard Arabic and various dialects) tweets. We remove the neutral and mixed classes.

DZ: Touileb and Barnes (2021) annotate Northern African Arabizi social media posts for sentiment. In this case, we use the transliterated Arabic script version of the dataset and remove the neutral class.

MT: The data for Maltese (Dingli and Sant, 2016; Cortis and Davis, 2019) comes from the combination (Martínez-García et al., 2021) of two smaller datasets.

DE: The GermEval 2017 Shared Task (Wojatzki et al., 2017) released annotated data for several subtasks on German social media texts. We use the document-level data (task B) and remove mixed and neutral.

EU: Barnes et al. (2018) annotate Basque hotel reviews for structured sentiment. We map these to sentence-level binary sentiment classification, using the script provided with the data.

NO: Velldal et al. (2018) provide a collection of professional reviews from news outlets. We keep the binary document-level data.

Emotion datasets For emotion classification we use the SSEC (Schuff et al., 2017) and EnISEAR (Troiano et al., 2019) datasets. The SSEC dataset reannotates a stance and sentiment dataset of political tweets with crowd-sourced labels for eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust). The EnISEAR dataset, on the other hand, crowd sources descriptions of events tied to emotions (anger, disgust, fear, guilt, joy, sadness, shame), as well as how readers perceive these events.

For the SSEC, we separate 583 examples from the training set for development. For EnISEAR, we split the fully labeled data into train (70%), dev (10%), and test (20%). For EnISEAR, we use the crowd sourced annotations for emotion labels, rather than the prior emotion to align with SSEC. For both datasets, we take the view that any number of annotations is valid (the 0.0 strategy in SSEC) and accept any label that has been assigned to an example by at least one annotator.

4 Experimental Setup

In this section, we describe the approaches for three experimental setups (monolingual English sentiment classification, monolingual English emotion classification, and cross-lingual sentiment classification) from most resource intensive to least.

4.1 Sentiment classification

Supervised: To provide an upper-bound of fully supervised in-domain models, we use DistilBERT (Sanh et al., 2019), and RoBERTa base and large (Zhuang et al., 2021). To simulate low-resource scenarios, we train the same models with varying amounts of training data (200, 100, and 20 examples). We finetune these models for 5 epochs, with a learning rate of 2e-5, a weight decay of 0.01, and a batch size of 16 on a single Tesla T4 GPU. We take the best model on the development set for testing.

Few-shot: In this scenario, we assume we have a development set and a limited number of training examples (200, 100, 20). We train the same models in the same way as fully supervised training, but with the reduced training set size. We again take the best model on the development set for testing.

Prompting: In this scenario, we assume we have only a few training examples. We explore few-shot prompting (concretely 2-shot) using two OPT models (Zhang et al., 2022): namely, the 125 million and 1.3 billion parameter versions. We prompt these models by giving them 2 positive and negative examples with the following template (an example from the hotel domain):

(1) `I didn't like the hotel. Label: negative. We loved the hotel. Label: positive. {text}. Label:`

We take the first predicted token as the predicted label.

Rule-based: In this scenario, we assume no training data whatsoever. We compare these models with the rule-based SoCal system (see details in Section 2.1). This approach requires a large initial effort to create the rules and dictionaries, but afterward can be applied to new data without retraining.

Dictionary-based: Finally, we also compare simpler dictionary-based approaches which do not include rules, and instead rely on a simpler scoring procedure for each text:

$$score(\text{text}, D) = \frac{1}{|D|}\frac{1}{|\text{text}|}\sum_{w \in \text{text}} s_e(w, D)$$

where D is a sentiment dictionary, either containing a list of words with positive orientation D_{pos} or negative D_{neg}, and s_e is a function that returns 1 if a word w is in D, otherwise 0. The $score$ function therefore returns the average score of a text, normalized by the length of the text and by the length of the dictionary D. To predict the aggregate semantic orientation (positive or negative), we divide the positive score by the negative score

$$\text{semantic orientation} = \frac{score(\text{text}, D_{pos})}{score(\text{text}, D_{neg})}$$

If this orientation is greater than a certain λ, we will assume that the orientation is positive and return 1, otherwise we will assume it is negative, and return 0.

We can then use available sentiment dictionaries to estimate the semantic orientation of a text. For all dictionary-based methods, we further preprocess the texts by tokenizing and lemmatizing the text using spaCy.[2] For sentiment dictionaries, we use the available HuLiu dictionary (Hu and Liu, 2004), the NRC Hashtag sentiment dictionary (Mohammad et al., 2013), and the MPQA subjectivity and sentiment dictionary (Wiebe et al., 2005).

Dictionary induction: Finally, it is also possible to automatically create a sentiment or emotion dictionary from a small seed dictionary. In this case, we use the SemAxis method (An et al., 2018) with a small seed dictionary of 10 words per class. We limit the expansion of the dictionaries to tokens found in the test set and allow only words which have a cosine ≥ 0.15 to reduce likely noisy.

We compare the use of three embedding spaces to induce the new dictionaries: 200 dimensional GloVe embeddings trained on Twitter data (Pennington et al., 2014), 300 dimensional FastText embeddings trained on Wikipedia data (Bojanowski et al., 2017), and 300 dimensional FastText embeddings trained on Wikipedia and the GigaWord corpus[3] (Fares et al., 2017).

[2] Found at `https://spacy.io/`.

[3] These can be found at `http://vectors.nlpl.eu/repository/20/22.zip`

4.2 Emotion classification

Supervised and Few shot: Given that both the SSEC and EnISEAR datasets are multi-label, we train the models using a one-vs-all approach, effectively creating a binary version of the dataset for each emotion and training a binary classifier. Like the sentiment experiments, we use DistilBERT, RoBERTa-base, and RoBERTa-large. The training procedure is the same as with sentiment. We perform experiments with 200, 100, and 20 training examples for the few shot experiments.

Prompting: For prompting, we use the Flan T5 models (Chung et al., 2022) (base and large), which are instruction tuned models. We performed initial experiments with the same OPT models used for sentiment analysis, but found that the multi-label nature of emotion classification was better covered using the Flan T5 models. For prompting the SSEC dataset, we use the following template:

(2) `What emotions are found in this text (Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, None)?: {text}`

where `text` is the text to be classified. For EnISEAR, we replace the emotions with anger, disgust, fear, guilt, joy, sadness, and shame. We assume that any mention of these words in the generated text is a predicted label.

Dictionary-based: As emotion classification in the datasets we use is a multi-label task, we cannot use the semantic orientation score as is. Instead, we set a threshold value $\lambda = 1$ and predict any label where $score(\text{text}, D_{emotion}) > \lambda$. This allows for our dictionary-based approach to predict multiple labels.

We use the NRC emotion dictionary as an emotion dictionary (Mohammad and Kiritchenko, 2015), which contains 16,862 entries with annotations for 8 emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust), which were compiled semi-automatically using word–emotion association scores with hashtags.

Dictionary induction: Similarly, we can use an adapted version of the SemAxis method to induce emotion dictionaries. In this setting, we create a semantic axis vector for each emotion we wish to propogate. For example, to create a semantic axis for 'anger' V_{anger}, we create the positive pole vector $V^{+}{}_{anger}$ by averaging the vector representation of seed words for 'anger' and the negative pole vector $V^{-}{}_{anger}$ by averaging the vectors of all other seed words.

Once we have the semantic axis vectors for each emotion, we can expand the original seed dictionaries by taking any word whose vector representation has a positive cosine distance with the semantic axis. As with sentiment, we take a conservative estimate and allow only words which have a cosine > 0.15 to reduce noise.

We then use the same prediction procedure as with the dictionary-based approach.

4.3 Cross-lingual generalization

We also compare zero-shot cross-lingual performance of multilingual large language models (MLLMs), in this case XLM-RoBERTa base and large, to dictionary induction. For the MLLM experiments, we train on one of the three English corpora (MPQA, OpeNER, and SemEval) and test the best model on the English development data on all non-English corpora.

For the dictionary induction experiments, we use the SemAxis method with FastText embeddings (Bojanowski et al., 2017), as these have embeddings available for most languages. For NArabizi (DZ), we use the embeddings trained on modern standard Arabic as a proxy.

4.4 Evaluation

For both sentiment and emotion classification datasets, we evaluate using Macro F_1, as the distribution of labels is unbalanced and we are interested in knowing how well the models perform on the less frequent labels as well.

5 Results

In this section we detail the results for sentiment classification, out of domain performance, emotion classification, and cross-lingual transfer.

5.1 Sentiment classification

Table 2 shows the Macro F_1 of the sentiment classification approaches on the English datasets (MPQA, OpeNER, and SemEval), as well as the average of all results per each approach.

The fully supervised upper-bound achieves an average F_1 of 91.2, showing strong performance for this binary classification task.

		MPQA	OpeNER	SemEval	Avg.
Supervised	DB	86.3	92.7	90.1	91.2
	RBB	87.2	94.4	91.0	
	RBL	92.0	95.3	91.5	
FewShot-200	DB	84.7	77.4	70.9	80.9
	RBB	80.8	93.4	86.8	
	RBL	67.9	80.5	86.2	
FewShot-100	DB	59.0	65.3	66.5	56.5
	RBB	62.5	81.9	45.3	
	RBL	38.5	31.1	58.1	
FewShot-20	DB	49.0	23.7	47.4	40.5
	RBB	36.9	42.1	42.2	
	RBL	39.3	42.1	42.2	
Prompted	OPT-125m	34.0	52.4	51.8	56.9
	OPT-1.3B	59.7	84.1	59.5	
Rules	SoCal	74.9	83.9	74.0	77.6
Dictionary	HuLiu	61.4	71.4	59.3	61.6
	NRC Hash	52.7	67.4	68.6	
	MPQA	60.7	60.2	52.5	
Induced	Twitter	61.9	65.1	67.7	61.8
	NLPL22	58.2	61.8	59.6	
	FastText	53.6	66.8	61.4	

Table 2: Results on sentiment analysis (MacroF1). DB: DistilBERT, RBB: RoBERTa-base, RBL: RoBERTA-large.

In the low-resource scenario, FewShot-200 is the best performing approach (80.9), followed closely by the rule-based SoCal (77.6). The dictionary-induction methods (61.8) and dictionary-based methods (61.6) achieve quite similar results, followed by prompting (56.9) and the few-shot methods using 100 examples (56.5) and 20 (40.5).

In general the RoBERTa-large model suffers more in the few-shot scenarios, losing 3.4-20 percentage points (pp) compared to RoBERTa-base. For prompting, however, the opposite is true, as the 1.3 billion parameter model performs 21.7 pp better than the 125 million parameter model. This ties in well with research indicating that the size of the language model leads to better few-shot performance (Brown et al., 2020).

Surprisingly, dictionary-based methods perform better than FewShot-100 or prompting large language models. Even more surprising is that inducing a sentiment lexicon from as few as 10 labeled words can outperform careful hand-designing of these dictionaries.

Approach	Avg. In	Avg. Out	$\overline{TL}_{A\rightarrow B}$
Supervised	91.2	84.6	13.3
FewShot-200	80.9	70.6	20.7
FewShot-100	56.5	47.4	18.1
FewShot-20	40.5	31.5	18.2

Table 3: We show the average in-domain results (Avg. In), average out-of-domain results (Avg. Out) and average domain transfer loss ($\overline{TL}_{A\rightarrow B}$) for the supervised models on English sentiment analysis.

Therefore, revisiting **RQ1**, we can say for binary sentiment classification, *fine tuning a model on as few as 100 examples per class gives competitive in-domain performance*. For anything less, *rule-based methods perform better*.

5.2 Out of domain performance of sentiment classification

Unlike prompting and dictionary-based approaches, supervised and few-shot methods are tied heavily to the domain they are trained with. In order to quantify the loss in performance of supervised models, we measure *domain transfer loss*, which is defined in Equation 1:

$$TL_{x\rightarrow y} = S_{x\rightarrow x} - S_{x\rightarrow y} \quad (1)$$

where $TL_{x\rightarrow y}$ is the difference of the Macro F_1 score $S_{x\rightarrow x}$ of a model fine-tuned on domain x and tested in the same domain, and the score $S_{x\rightarrow y}$ of the model fine-tuned on x and tested on domain y.

As we have two test domains $B = \{b_{domain1}, b_{domain2}\}$ for each training domain x, we average over these using Equation 2:

$$\overline{TL}_{x\rightarrow B} = \frac{1}{N_B} \sum_{\substack{i\in B \\ i\neq x}} S_{x\rightarrow x} - S_{x\rightarrow i} \quad (2)$$

		SSEC	EnISEAR	Ave.
Supervised	DB	74.6	72.1	67.6
	RBB	71.3	56.0	
	RBL	68.1	63.2	
FewShot-200	DB	55.5	62.8	54.1
	RBB	53.0	55.6	
	RBL	50.2	47.7	
FewShot-100	DB	45.6	47.3	45.8
	RBB	42.1	57.2	
	RBL	39.0	43.3	
FewShot-20	DB	42.8	43.3	41.6
	RBB	39.4	43.3	
	RBL	37.5	43.3	
Prompted	FlanT5-base	51.5	58.9	57.7
	FlanT5-large	47.6	72.6	
Seed Dict.		37.4	47.9	42.7
Dictionary	NRC	52.2	46.4	49.3
Induced	Twitter	62.0	53.2	54.4
	NLPL22	53.0	45.7	
	FastText	53.9	58.8	

Table 4: Macro averaged F_1 for emotion classification results on the SSEC and EnISEAR datasets. DB: DistilBERT, RBB: RoBERTa-base, RBL: RoBERTA-large.

Finally, we compute the average domain transfer loss for all models of a certain approach $A = \{\text{supervised, few shot, } \ldots\}$ by computing the average of the domain transfer losses $\overline{TL}_{x \to B}$ for all models in the approach:

$$\overline{TL}_{A \to B} = \frac{1}{N_A} \sum_{i \in A} \overline{TL}_{i \to B} \quad (3)$$

Table 3 shows the average in-domain results (Avg. In), average out-of-domain results (Avg. Out) and average domain transfer loss (TL) for the supervised models (the full results table can be found in Table 8 in Appendix A). Models finetuned in a supervised fashion achieve the best in-domain (91.2) and out-of-domain (84.6), with the smallest transfer loss (13.3).

Although FewShot-200 achieves relatively good in-domain performance (80.9), it has the largest transfer loss (20.7), with the out-of-domain performance dropping to 70.6, 7 pp. below SoCal. This suggests that it is highly dependent on the few training examples seen being in-domain and that it cannot be safely applied out-of-domain.

Finally, both FewShot-100 and FewShot-20 have similar transfer losses (18.1/18.2), although the already low in-domain performance (56.5/40.5) means that using these models either in-domain or out-of-domain is impractical.

In contrast, the prompting, rule-based, and dictionary-based approaches do not suffer from this and perform more consistently across domains.

Therefore, the answer to **RQ2** is that *rule-based methods perform better across domains that few-shot supervision methods.*

5.3 Emotion classification

Table 4 shows the Average Macro F_1 scores for all approaches on the two emotion classification datasets, as well as the averaged score per approach (results for each emotion can be found in Tables 9 and 10 in the Appendix).

Again, the fully supervised upper bound achieves the best F_1 (67.6), where DistilBERT achieves much better performance than either RoBERTa model. RoBERTA-base achieves poor performance on EnISEAR, RoBERTa-large consistently performs quite poorly, suggesting that it requires either more data or more careful fine-tuning than was used here.

The best performing method in the low-resource setting is prompting (57.7), followed by dictionary induction (54.4) and Few-shot 200 (54.1). The dictionary-based method, as well as the FewShot-100 and -20 approaches, perform quite poorly (49.3, 45.8, and 41.6 respectively), with the latter achieving worse performance than the 10 word per emotion seed dictionary (42.7).

In contrast to prompting OPT models for sentiment analysis, the FlanT5-large model does not consistently improve over the base model, achieving a quite low score on the SSEC dataset (47.6).

Similar to the sentiment experiment, the induced emotion dictionaries perform as well or better than previously compiled emotion dictionaries (NRC).

Returning to **RQ1**, *for emotion analysis prompting or dictionary induction perform better than few shot approaches.*

5.4 Cross-lingual sentiment classification

The results of the cross-lingual experiments can be seen in Table 5. In general, the XLM-RoBERTa models perform much better than the dictionary induction approaches (10-20 pp). However, this depends heavily on the source language corpus used to train, as several XLM-RoBERTa results are lower than their respective dictionary induction

Model	train	Test Lang							Avg. on Test
		self	DE	ES	AR	NO	EU	DZ	MT	
maj. baseline			46.2	45.2	33.5	41.8	45.8	39.1	39.4	41.6
XLM-RoBERTa-base	MPQA	87.1	65.9	89.9	68.8	74.3	80.2	52.8	54.9	69.5
	OpeNER	93.0	73.3	90.8	72.4	75.5	79.0	57.5	58.3	72.4
	SemEval	88.9	71.0	89.0	73.1	75.1	82.4	71.3	58.8	74.4
XLM-RoBERTa-large	MPQA	89.1	62.7	84.0	62.3	74.2	80.3	50.9	30.9	63.6
	OpeNER	95.6	72.8	93.8	77.1	82.9	87.2	72.2	40.2	75.2
	SemEval	90.9	67.6	88.4	75.0	77.0	83.6	76.9	51.2	74.2
FlanT5-base			69.9	77.9	36.3	43.9	14.5	26.4	44.5	44.8
FlanT5-large			73.1	93.4	89.7	86.7	90.9	97.6	82.6	87.7
dictionary induction			50.1	59.9	62.9	41.8	45.8	58.7	50.0	52.7

Table 5: Results on cross-lingual sentiment analysis (MacroF1).

approach (large trained on MPQA and tested on AR, DZ, or MT for example).

Curiously, the large version performs worse than the base version when trained on MPQA or SemEval. Like with the previous experiments, this may suggest that the larger models need more data or require more careful tuning than we performed in our experiments. In either case, it is important to note that simply increasing the size of the cross-lingual model will not necessarily result in better results.

Finally, the results of all models are generally worse for Narabizi (DZ) and for Maltese (MT), which is unsurprising, as they have little or no pretraining data in XLM-RoBERTA. The one exception is the Flant5-large, which achieves very good results on both. It is unclear what exactly causes this difference in multilingual ability, especially for low-resource languages like Narabizi and Maltese, although larger models are known to memorize training data (de Wynter et al., 2023) and both of these datasets are available in text format. Therefore, we cannot rule out data contamination as the source of such a jump in performance.

Finally, the cross-lingual models achieve an average of 71.5, compared to 66.3 for prompting or 52.7 for dictionary induction. Thus, we can cautiously venture that for **RQ3**, cross-lingual methods allow for the best results, although prompting larger multi-lingual LLMs may also provide good results in the future.

6 Conclusion and future work

In this paper we have performed experiments on 10 sentiment datasets and two emotion classification datasets in 8 languages with a number of low-resource approaches (dictionary-based methods, rule-based methods, few-shot methods and prompting methods). The main experiments were performed on the English language datasets (3 sentiment and 2 emotion), while further experiments were performed in 7 additional languages.

These results confirm that under ideal circumstances, fully supervised models perform much better than low-resource approaches. However, in low-resource settings (lack of training data, domain shift), these same models quickly lose performance and rule-based and dictionary-based approaches perform on par or even better if there is a domain shift involved.

While prompting achieved impressive performance in our experiments, given that the models were not explicitly trained, this came at a price. Namely, such approaches for languages other than English are currently not available or not on par with English versions. This area will surely be explored in the near future, but this current gap is nonetheless a product of the over-reliance on English in NLP.

The strong cross-lingual performance of the XLM-RoBERTa models suggests that cross-lingual approaches, especially those designed for adapting to new languages, scripts (Pfeiffer et al., 2021), or generally enabling ever more multilingual pretraining (Lauscher et al., 2020; Pfeiffer et al., 2022).

We find conflicting evidence on the importance

of model size for low-resource performance. On the one hand, prompting the larger OPT model for sentiment classification gave consistently better results. On the other hand, RoBERTa-large suffered much more in out-of-domain classification and generally performed worse than RoBERTA-base on emotion classification in all data regimens. For prompting in emotion classification, FlanT5-large did not lead to consistent gains over the base version and finally, XLM-RoBERTa-large similarly performed worse than the base version on cross-lingual sentiment classification. This finding seems to indicate that some of the promised few-shot performance found in large language models is either lacking or requires careful tuning.

In the future, it would be interesting to expand this comparison to other dictionary induction methods, such as cross-lingual propagation (Buechel et al., 2020), or high-coverage expansion (Köper and Schulte im Walde, 2016). Given the promising results from the simple prompting approaches we used in our experiments, further research on how to expand these models to new languages and tasks would be of great use.

Finally, multi-lingual few-shot approaches (Lauscher et al., 2020) could also be compared, as it is often possible to use a few examples in the target language.

7 Limitations

In this paper, we only explore binary sentiment classification, as it is enables cross-lingual experiments to be somewhat comparable. However, this is a simplified task, which should be taken into account when interpreting the results. Our multilingual datasets also come from various domains and, although we try to control for this in English, this does lead to some effect in the results. Finally, for emotion detection, we only experiment in English.

We also chose only a few representative methods for each approach (few-shot, prompting, rule-based, etc). This was a necessary simplification given the large number of available models, and care was given to choose truly representative methods for each approach. However, some relevant methods may not be represented here.

Finally, we only report the results for a single run for the supervised models, rather than the average of 5-10 runs as is common. We compensate by averaging over results on several datasets and across several methods.

Acknowledgements

This work has been partially supported by the HiTZ center and the Basque Government (Research group funding IT-1805-22).

We also acknowledge the funding from the following projects: DeepKnowledge (PID2021-127777OB-C21) project funded by MCIN/AEI/10.13039/501100011033 and by FEDER Una manera de hacer Europa.

References

Rodrigo Agerri, Montse Cuadros Cuadros, Seán Gaines, and German Rigau. 2013. Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, 51(0):215–218.

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461, Melbourne, Australia. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu,

Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2016. Feelings from the Past—Adapting affective lexicons for historical emotion analysis. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 54–61, Osaka, Japan. The COLING 2016 Organizing Committee.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Keith Cortis and Brian Davis. 2019. A social opinion gold standard for the Malta government budget 2018. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 364–369, Hong Kong, China. Association for Computational Linguistics.

Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. 2023. An evaluation on large language model outputs: Discourse and memorization.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexiei Dingli and Nicole Sant. 2016. Sentiment analysis on maltese using machine learning. In *Proceedings of The Tenth International Conference on Advances in Semantic Processing (SEMAPRO 2016)*, pages 21–25.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Jaap Kamps, Maarten Marx, Robert J. Mooken, and Maarten de Rijke. 2004. Using wordnet to measure semantic orientations of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC '04, pages 1115–1118.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2595–2598, Portorož, Slovenia. European Language Resources Association (ELRA).

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antonio Martínez-García, Toni Badia, and Jeremy Barnes. 2021. Evaluating morphological typology in zero-shot cross-lingual transfer. 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: Long Papers)*, pages 3136–3153, Online. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.

Saif M. Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Herb Meiselman, editor, *Emotion Measurement*. Elsevier.

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. ASTD: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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*, pages 6805–6817, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2023. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*, 14(1):108–132.

David Samuel, Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2022. Direct parsing to sentiment graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

470–478, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Jacopo Staiano and Marco Guerini. 2014. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–433, Baltimore, Maryland. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Samia Touileb and Jeremy Barnes. 2021. The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3700–3712, Online. Association for Computational Linguistics.

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*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian review corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 1–12, Berlin, Germany.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

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*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Appendix

Dataset	lang	domain	Train	Dev	Test	Pos %
MPQA	EN	news	987	337	299	48.5
SemEval	EN	social media	3,737	413	1,791	72.2
OpeNER	EN	hotel reviews	1,210	174	347	72.7
OpeNER	ES	hotel reviews	1,029	147	296	82.6
GermEval	DE	social media	6,444	772	1,490	18.2
ASTD	AR	social media	2,468	353	706	50.2
NoReC	NO	reviews	2,675	516	417	67.1
MultiBooked	EU	hotel reviews	789	113	227	84.7
NArabizi	DZ	social media	564	75	92	52.0
Maltese	MT	social media	595	85	171	31.8

Table 6: Sentiment dataset statistics, including the percentage of positive examples for the sentiment datasets.

	lang	Train	Dev	Test	Anger	Anticipation	Disgust	Fear	Guilt	Joy	Sadness	Shame	Surprise	Trust
SSEC	EN	2,329	583	1,956	16.9	15.7	12.7	10.7	–	12.0	15.4	–	6.5	10.0
EnISEAR	EN	720	80	201	17.5	–	11.5	11.8	17.0	10.5	17.3	14.5	–	–

Table 7: Emotion dataset statistics, including the relative distribution of labels for the emotion classification datasets are also shown.

	Train	MPQA			OpeNER			SemEval		
	Test	MPQA	OpeNER	SemEval	MPQA	OpeNER	SemEval	MPQA	OpeNER	SemEval
Fully Supervised	DistilBert	86.3	84.4	82.3	77.5	92.7	85.6	66.8	91.5	90.1
	RoBERTa-base	87.2	90.3	87.1	79.7	94.4	88.4	82.7	94.0	91.0
	RoBERTa-large	92.0	90.4	86.3	75.2	95.3	87.2	78.6	94.8	91.5
FewShot-200	DistilBert	84.7	86.0	83.0	64.8	77.4	57.3	38.7	37.8	70.9
	RoBERTa-base	80.8	84.8	84.9	71.9	93.4	80.6	77.3	92.4	86.8
	RoBERTa-large	67.9	46.0	48.9	70.8	80.5	78.2	72.7	94.4	86.2
FewShot-100	DistilBert	59.0	54.6	52.7	59.3	65.3	60.8	47.3	38.4	66.5
	RoBERTa-base	62.5	44.3	44.8	57.9	81.9	66.4	29.3	42.1	45.3
	RoBERTa-large	38.5	42.0	42.4	45.4	31.1	46.8	43.2	35.4	58.1
FewShot-20	DistilBert	49.0	21.5	23.2	38.4	23.7	24.4	36.5	22.4	47.4
	RoBERTa-base	36.9	21.5	21.3	29.3	42.1	42.2	29.3	42.1	42.2
	RoBERTa-large	39.3	24.4	46.8	29.3	42.1	42.2	29.3	42.1	42.2

Table 8: Cross-domain results on sentiment analysis (Macro F_1).

		Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Ave.
Supervised	DistilBERT	76.6	62.6	80.5	73.5	78.7	75.4	69.9	79.8	74.6
	RoBERTa-base	78.2	70.6	73.3	68.3	72.8	68.9	63.5	75.0	71.3
	RoBERTa-large	80.7	68.4	76.9	66.2	74.3	69.2	69.6	39.5	68.1
FewShot-200	DistilBERT	76.7	46.8	72.1	58.2	38.0	57.9	42.2	51.7	55.5
	RoBERTa-base	50.5	57.1	60.8	48.6	62.0	63.1	42.2	39.5	53.0
	RoBERTa-large	78.4	38.5	73.2	53.8	38.0	37.7	42.2	39.5	50.2
FewShot-100	DistilBERT	57.7	30.8	61.5	46.0	38.1	45.4	42.2	42.9	45.6
	RoBERTa-base	57.4	38.1	42.7	43.6	38.0	35.2	42.2	39.5	42.1
	RoBERTa-large	38.9	46.7	34.8	37.1	38.0	35.2	42.2	39.5	39.0
FewShot-20	DistilBERT	39.4	44.0	46.8	37.1	38.1	55.0	42.2	39.4	42.8
	RoBERTa-base	38.9	38.1	31.8	37.1	38.0	49.5	42.2	39.5	39.4
	RoBERTa-large	26.7	38.1	34.8	37.1	38.0	43.4	42.2	39.5	37.5
Prompted	FlanT5-base	64.1	38.4	43.9	58.7	49.9	52.1	47.9	57.0	51.5
	FlanT5-large	33.7	29.1	68.7	48.4	72.4	35.0	47.3	46.4	47.6
dictionaries	seed dictionary	29.4	35.0	35.9	40.3	40.0	32.0	43.6	43.3	37.4
	NRC	58.6	51.5	56.1	48.9	56.6	53.9	46.4	45.4	52.2
Induced	SemAxis	77.8	76.2	63.6	58.1	55.8	70.3	42.5	51.6	62.0
	NLPL22	48.6	55.1	51.6	55.8	61.0	49.7	47.4	54.7	53.0
	FastText	60.4	38.1	58.7	55.8	61.1	50.9	53.5	53.1	53.9

Table 9: Per class and Macro averaged F_1 for emotion classification results on the SSEC dataset.

		Anger	Disgust	Fear	Guilt	Joy	Sadness	Shame	Ave.
Supervised	DistilBERT	77.8	80.1	74.6	77.9	82.1	66.6	45.4	72.1
	RoBERTa-base	84.6	81.6	74.1	78.7	89.4	74.7	65.6	78.4
	RoBERTa-large	82.8	83.7	65.7	41.7	55.0	70.2	43.5	63.2
FewShot-200	DistilBERT	70.4	57.5	44.3	71.8	76.3	64.1	55.4	62.8
	Roberta-base	57.9	45.1	44.3	72.7	85.0	40.7	43.5	55.6
	Roberta-large	72.9	45.1	44.3	41.7	45.8	40.7	43.5	47.7
FewShot-100	DistilBERT	70.2	45.1	44.3	41.7	45.8	40.7	43.5	47.3
	Roberta-base	73.6	45.1	44.3	70.4	82.9	40.7	43.5	57.2
	Roberta-large	41.7	45.1	44.3	41.7	45.8	40.7	43.5	43.3
FewShot-20	DistilBERT	41.7	45.1	44.3	41.7	45.8	40.7	43.5	43.3
	Roberta-base	41.7	45.1	44.3	41.7	45.8	40.7	43.5	43.3
	Roberta-large	41.7	45.1	44.3	41.7	45.8	40.7	43.5	43.3
Prompted	FlanT5-base	60.3	54.8	62.9	43.7	80.9	64.8	44.8	58.9
	FlanT5-large	53.0	66.3	82.7	77.4	91.9	81.1	55.4	72.6
dictionaries	seed dictionary	41.4	45.2	59.2	48.1	50.7	44.6	45.8	47.9
	NRC	50.6	48.7	40.6	48.1	39.7	51.2	45.8	46.4
Induced	NLPL22	54.5	50.9	50.2	50.2	53.7	58.3	54.9	53.2
	FastText	22.4	70.6	18.6	55.9	40.7	61.9	50.0	45.7
	Twitter	49.7	64.8	71.2	53.3	57.5	55.5	59.8	58.8

Table 10: Per class and Macro averaged F_1 for emotion classification results on the enISEAR dataset.

Class	Seed Words									
Positive	good	nice	happy	beautiful	wonderful	enjoy	love	best	terrific	great
Negative	bad	mean	terrible	sad	ugly	hate	dislike	disgusting	worst	stressful
Anger	angry	mad	annoyed	hate	annoying	furious	upset	irritated	irritating	displeased
Anticipation	want	wanting	desire	anticipate	anticipating	wait	waiting	expect	expecting	hope
Disgust	yuck	disgusting	nasty	revolting	repulsive	despicable	nauseated	repugnant	shocking	vile
Fear	scared	afraid	fear	worried	worry	scary	dangerous	dark	panic	terror
Joy	happy	content	joyful	fun	cheerful	cheerfulness	cheer	delighted	ecstatic	elated
Sadness	sad	unhappy	melancholy	sorrowful	sorrow	gloomy	gloom	pessimistic	heartbroken	depressed
Surprise	wow	surprise	surprised	amazed	gobsmacked	stunned	shocked	dazed	astonished	startled
Trust	trust	trustworthy	confidence	confident	sure	faith	conviction	convinced	belief	truthful
Guilt	guilt	guilty	culpability	disgrace	regret	remorse	penitence	remorseful	sorry	wrong
Shame	ashamed	embarrassed	embarrassing	humiliating	humiliated	stigma	scandal	scandalous	shame	shameful

Table 11: Seed dictionaries for each class.