Universal Joy
A Data Set and Results for Classifying Emotions Across Languages

Sotiris Lamprinidis
Copenhagen Business School
sla.msc@cbs.dk

Federico Bianchi
Bocconi University
f.bianchi@unibocconi.it

Daniel Hardt
Copenhagen Business School
dha.msc@cbs.dk

Dirk Hovy
Bocconi University
dirk.hovy@unibocconi.it

Abstract

While emotions are universal aspects of human psychology, they are expressed differently across different languages and cultures. We introduce a new data set of over 530k anonymized public Facebook posts across 18 languages, labeled with five different emotions. Using multilingual BERT embeddings, we show that emotions can be reliably inferred both within and across languages. Zero-shot learning produces promising results for low-resource languages. Following established theories of basic emotions, we provide a detailed analysis of the possibilities and limits of cross-lingual emotion classification. We find that structural and typological similarity between languages facilitates cross-lingual learning, as well as linguistic diversity of training data. Our results suggest that there are commonalities underlying the expression of emotion in different languages. We publicly release the anonymized data for future research.

1 Introduction

Emotions are fundamental to human experience across languages and cultures. The nature of emotions and their linguistic expression is a topic of enduring interest across disciplines such as psychology, linguistics, philosophy, and neuroscience. Emotion researchers have investigated the existence of basic emotions such as anger, fear, disgust, sadness, and happiness (Ekman, 2016), all of which were already described in the 19th century by Darwin ([1872] 1998) and Wundt (1896). Furthermore, Ekman (2016) reports a growing consensus concerning the universality of emotions across languages and cultures. Computational linguistics can help shed light on the way in which emotions are expressed in the languages of the world.

However, most existing research has focused either on English or used very small multilingual data sets. We present a new dataset, **Universal Joy** (UJ), of over 530,000 anonymized public Facebook posts distributed across 18 languages, labeled with five different emotions: anger, anticipation, fear, joy, and sadness. This dataset represents a substantial advance over prior datasets, both in terms of its size and its linguistic diversity. It provides a strong empirical foundation for exploring basic questions about the nature and expression of emotions across the languages of the world. Figure 3 shows the heatmap of relative emotion distribution for each language. In this paper, we use this dataset to explore multilingual emotion classification.

We first perform emotion classification in a **monolingual setting**, i.e., training and testing on a single language. We then expand to a **cross-lingual setting**, i.e., where the training data contains other languages in addition to the test data language. Finally, we test how well we can do in a **zero-shot learning** setting; here the training data does *not* include any data in the language of the test data – a setting particularly relevant for low-resource languages.

Overall, we find consistent effects of cross-lingual learning, which raises several interesting issues: first, it suggests that accurate emotion detection might be possible even for low-resource languages. Accurate models for such languages might be achievable with substantial amounts of training data from high-resource languages like English. More generally, however, we explore *why* cross-lingual learning works, and what linguistic circumstances support or hamper such learning. We explore three main factors: **code-switching**, **typological closeness** of training and test languages, and **linguistic diversity in the training data**.

We hope the richness of this data set opens up exciting future research avenues, and release the models and the complete anonymized dataset at https://github.com/sotlampr/universal-joy.

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 62–75
April 19, 2021. ©2021 Association for Computational Linguistics

Contributions 1) We publish a new dataset of over 530k anonymized public Facebook posts in 18 languages, labeled with five basic emotions. 2) We show results for various classification setups, including transfer learning setups like cross-lingual and zero-shot learning. 3) We analyze the sources of cross-lingual learning in depth, including the effect of code-switching, typological closeness, and linguistic diversity.

2 Data

The dataset described here is a substantially reorganized and cleaned version of one previously described, but not released Zimmerman et al. (2015). It was collected in October 2014 by searching for public Facebook posts with a Facebook "feelings tag". We did verify publication with the data protection officer of the main institution. For a Data Statement (Bender and Friedman, 2018), see Appendix A.

We remove any duplicates, and classify each instance's language using three methods: `langid`,[1] `cld3`,[2], and `FastText`.[3] We keep only instances where at least two of these methods agree. We manually evaluate 200 randomly selected instances labeled `deu`, `fra`, `eng`, `ita`, and `spa`,[4] and find the average precision of our method is 0.97(±0.04).

To anonymize the data, we remove identifying information by replacing names with the special token `[PERSON]`. Where possible (Dutch, English, French, German, Portuguese, Spanish, Italian), we use the `spacy`[5] NER to replace any PERSON entities. For languages without `spacy` support, we either use the Stanford CoreNLP NER tagger (Manning et al., 2014) and replace `PERSON`-tagged words (Chinese), or replace all given names and surnames found in Wiktionary for the respective languages (Bengali, Burmese, Hindi, Indonesian, Khmer, Malay, Romanian, Tagalog, Thai, Vietnamese).

Finally, we perform some additional preprocessing steps to replace any number with `0`, and the Facebook-specific tags *"with [PERSON]"* with the special token `[WITH]`, *"at [LOCATION]"* with the special token `[LOCATION]`, and photos, emails, and URLs with the special tokens `[PHOTO]`, `[EMAIL]`, and `[URL]`, respectively.

Similar to the approach of Zimmerman et al. (2015), we map the 27 initial emotion tags into five labels of basic emotions: anger, anticipation, fear, joy, and sadness (see mapping in Appendix Table 10). We choose this label set for several reasons: first, it is similar to the lists of basic emotions proposed in the psychological literature (Ekman, 2016; Plutchik, 1994). Second, each of the five labels is well-represented in the Facebook data, whereas the original tags are highly imbalanced and often rare. Finally, it is similar to the lists of basic emotions used in recent NLP studies, and facilitates comparison with recent work by e.g., Abdul-Mageed and Ungar (2017).

As training languages, we choose all languages with more than 35 samples for the least frequent emotion (i.e., fear). The size and distribution of emotions for each dataset is presented in table 11.

Distribution of Languages There is a wide variety in the amount of data per language, ranging from 284,265 posts for English, the most frequent language, to 869 posts for Bengali (see table 1).

ISO code	Samples	Language	Family
ben	869	Bengali	Indo-European
cmn	4909	Chinese	Sino-Tibetan
deu	5902	German	Indo-European
eng	284265	English	Indo-European
fra	6557	French	Indo-European
hin	1823	Hindi	Indo-European
ind	6201	Indonesian	Austronesian
ita	6709	Italian	Indo-European
khm	977	Khmer	Austroasiatic
mya	953	Burmese	Sino-Tibetan
nld	2201	Dutch	Indo-European
por	31326	Portuguese	Indo-European
rom	1940	Romanian	Indo-European
spa	31326	Spanish	Indo-European
tgl	4909	Tagalog	Austronesian
tha	3803	Thai	Tai-Kadai
vie	3956	Vietnamese	Austroasiatic
zsm	4908	Malay	Austronesian

Table 1: Languages in Universal Joy data set

Distribution of Emotions per Language There are significant differences in the prevalence of each

[1] https://github.com/saffsd/langid.py

[2] https://github.com/google/cld3

[3] https://fasttext.cc/docs/en/language-identification.html (Joulin et al., 2016b,a)

[4] Here and in what follows, we use standard ISO language codes, which are given in table 1.

[5] https://spacy.io

emotion across languages, as shown in table 2. The relative distributions are shown in Figure 3. Fear is a very rare emotion in all languages, while joy is the most common. Anticipation, the second most frequent class, is especially prevalent in English . There are also differences in joy (more prevalent in Spanish) and sadness (more prevalent in Portuguese).

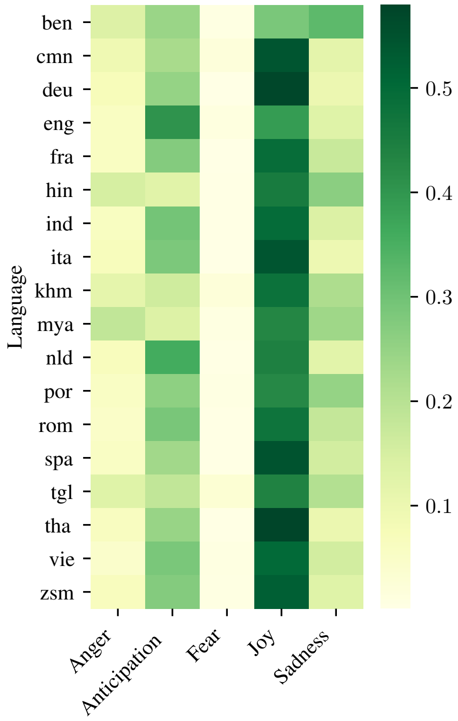

Figure 1: Heatmap of relative emotion distribution per language.

2.1 Training Datasets

We create three versions of the dataset for training purposes:

- **Small:** this version includes the five languages with sufficient training data for the least frequent emotion, fear: namely `eng`, `spa`, `por`, `cmn`, `tgl`. This dataset is balanced by language, so that there are 2,947 posts for each language.
- **Large:** this version includes 29,364 posts for each of the three most frequent languages: `eng`, `spa`, `por`. Note that each of these is a superset of the corresponding language in the Small training set.
- **Huge:** this version contains 283,853 posts from the single most frequent language, `eng`.

2.2 Test Datasets

For each training language, we create fixed-size development and test sets, stratified by emotion and following a 70:15:15 ratio with respect to the Small version of the dataset. Thus for `eng`, `spa`, `por`, `cmn`, `tgl`, the test and dev sets each consist of 631 posts.

The rest of the languages are combined in a separate test set, labeled *low-resource*. Note that for the purposes of this paper, what we call *low-resource languages* are the thirteen languages with insufficient training data in our corpus. This includes languages such as German and French, which are not, in general, low-resource languages. The low-resource test set for each of these languages simply consists of all the posts in that language. The low-resource sets will allow us to broadly measure zero-shot performance (see Section 4.3).

3 Methods

We model the task as a series of binary classification problems, one per emotion, similarly to Mohammad et al. (2018) (this allows for the theoretical case where there are multiple emotions in one instance). We use Logistic Regression and Multilingual BERT as classifiers, to probe for various lexical and syntactic properties of the task.

3.1 Logistic Regression Models

We use the `scikit-learn`[6] implementation to extract TFIDF-weighted bag-of-words (BOW) features, and train Logistic Regression **LR** models with L2 regularization ($C = 1.0$) and balanced label loss-weighting. We include BOW features for comparison purposes – in particular, to assess the extent to which cross-lingual effects might arise from code switching or other forms of token overlap across languages. For the **LR** models we use the same tokenization as in 3.2, from the pre-trained multilingual BERT model.

3.2 Multilingual BERT

To optimize performance, we use the multilingual BERT (mBERT) model.[7] We follow Devlin et al. (2019) in optimizing the model using a machine with an Intel i9-9940X CPU, 32GB RAM and a NVIDIA Quadro RTX 6000 GPU.

The loss $\mathcal{L}$ is the mean over the individual losses $l \in L$ for each emotion:

[6] `https://scikit-learn.org`

[7] `https://github.com/google-research/bert`

Language	Anger	Anticipation	Fear	Joy	Sadness	Sum
cmn	**460**	**1094**	**104**	**2666**	585	4909
deu	425	**1475**	8	**3388**	**606**	5902
eng	16842	**115793**	3258	**111211**	**37161**	284265
fra	382	**1788**	22	**3222**	1143	6557
ind	382	**1841**	32	**3077**	869	6201
ita	472	**1910**	20	**3656**	**651**	6709
por	1776	**8103**	218	13363	**7866**	31326
spa	1795	**7285**	**150**	**17175**	4921	31326
tgl	**647**	**914**	**159**	2166	**1023**	4909
zsm	326	1344	34	**2566**	638	4908
Sum	23507	141547	4005	162490	55463	387012

Table 2: Number of samples per language and emotion, for top ten languages, significant outliers at $\alpha = 0.05$ in **bold**, using χ^2 test $\chi^2 = 1780054.57, dof = 36, N = 50, p \ll 0.001$.

$$\mathcal{L} = \sum_{i=1}^{E} L_i(y_i, x) \big/ E$$

$$L_i(y_i;\, x) = -w_i \left[y_i \log(p(Y_i|\, x)) + (1 - y_i) \log(1 - p(Y_i|\, x)) \right]$$

$$p(Y|\, x) = \text{Sigmoid}(\bar{h}(x) W + b)$$

where E is the number of emotions $e \in \mathcal{E}$, $\vec{y} \in \{0, 1\}^E$ is a one-hot vector of the target emotion, $x \in X$ are the input byte-pair pieces, $\bar{h}(x) : X \to \mathbb{R}^{768}$ is the mean-pooled output of the BERT `[CLS]` token for input x, $W_{768 \times E}$ and $\vec{b} \in \mathbb{R}^E$ are learnable parameters, and $P(Y|x)$ is the predicted probability distribution over the emotions.

We use instance weighting to address the high class imbalance. For each emotion, we weight positive class instances as $w_i = N_{e=\neg\mathcal{E}_i} / N_{e=\mathcal{E}_i}$, i.e., the inverse proportion of negative examples to positive examples, averaged for all languages.

We linearly increase the learning rate for half an epoch and then linearly decay it until the end of the training. For the monolingual and zero-shot learning task, we select the model with the highest macro-averaged F1-score across all emotions on the target language development set. For the cross-lingual task, we choose the model with the best average score on all languages.

Monolingual Classification The simplest setting is the classification of emotions within one language – that is, the test, development, and training data are all taken from the same language. This provides a strong baseline for cross-lingual work.

Cross-lingual Classification In this setting, we test whether knowledge about emotions expressed in one language can be transferred to another language. Here, the training data includes one or more languages in addition to the language of the development and test data. Note that emotion distributions *differ* between languages. This likely affects performance and could be addressed by stratified resampling. However, that presupposes that all languages exhibit the same emotions to the same degree, which is by no means certain. So while sampling would improve performance, it would distort the "natural" distribution, and preclude future analysis of language-specific studies.

Zero-shot learning Here the training data does *not* include the language of the development and test sets. We use two versions of zero-shot training: single-language, and multilingual, depending on the number of languages present in the training data. In either case, the size of the training data is the same.

4 Results

We treat each emotion as a separate binary task, and compute a macro-average of the F1-scores for each of the six tasks. A random baseline model (table 15 in the Appendix) always predicts the positive class for all emotions, giving an average macro F1-score of around 0.3.

4.1 Monolingual English Tests

We evaluate our methods on available data sets for emotion classification in English (Abdul-Mageed and Ungar, 2017; Troiano et al., 2019). We compare performance on the EmoNet test data as well as the English test data from ISEAR (Troiano et al., 2019). In addition, we report performance on the English test data from our Universal Joy dataset. These tests are designed to assess two points: whether our data set is comparable to previously published data, and whether the models we use are performant enough to enable meaningful investigations.

We obtained the EmoNet dataset (Abdul-Mageed and Ungar, 2017) from the authors of the paper. The benchmark shared by the authors contains 80k tweet IDs. However, some of the tweets do not exist anymore, and some contain emotions we are not considering in this work (i.e., *disgust*). Thus, after removal we were left with a test set of around 40K tweets. We were, therefore, unable to reproduce their full setup.

Table 3 shows results on English test data, using the two LR models as well as mBERT on the small, large, and huge training data. For a prediction, we take the output probabilities from EmoNet and use the most probable emotion that is in our set of five emotions.

		Test			
Model	Training	UJ	EmoNet	ISEAR	Avg.
EmoNet	EmoNet	0.23	0.47	0.41	0.37
LR	Small	0.45	0.31	0.24	0.33
	Large	0.52	0.41	0.31	0.41
	Huge	0.52	0.47	0.37	0.45
mBERT	Small	0.46	0.40	0.48	0.45
	Large	0.58	0.48	0.46	0.51
	Huge	**0.63**	**0.55**	**0.49**	**0.56**

Table 3: Macro-F1 score for different models on various English datasets.

The results show that using more training data in a LR model or any of the mBERT model improves performance across the board and yields competitive results. We are therefore confident that our data collection and model choices produce meaningful results. But does this performance extend to other test languages than English?

4.2 Cross-lingual Tests

We now turn to cross-lingual investigation: table 4 shows results using a variety of training sets with the five test sets from eng, por, spa, cmn, and tgl. We show results for both LR models and mBERT on monolingual, cross-lingual, and zero-shot Universal Joy data.

In addition to the Small, Large, and Huge training sets described above, we test mBERT on some additional training data combinations. First, we divide Small into Indo-European (Small-IE) and non-Indo-European (Small∼IE). We also combine the Large and Small training datasets from all languages to test whether more diversity balances out more data. This dataset comprises five languages, but only about half as many instances as Huge English.

In general, mBERT models outperform the LR models. Furthermore, the mBERT models frequently show positive cross-lingual effects; that is, training data improves performance even when it is from a language other than the test language. For example, small-mono for spa is 0.43, while small-all (including all five languages) is 0.45. Small-all on average is 0.53, while small-mono is 0.51. On the other hand, large-mono (0.57) is better than large-all (0.56). Perhaps cross-lingual improvements are easier to obtain when the monolingual model is weaker.

Using training data based on language families (IE and ∼IE, respectively), indicates typological effects (which we explore further in Sections 5.2 and 5.3). Specifically, training on non-Indo-European languages results in higher performance for cmn and tgl (though not the highest overall).
Table 5 shows mBERT results for monolingual and zero-shot training. Unsurprisingly, the best results always involve training on the same language as the test language (see diagonal), and more data helps.

Table 6 compares mBERT with the LR models on zero-shot. In particular, we compare single-language zero-shot with multilingual. We see that the multilingual scores are consistently higher than the single-language scores, across all models. This provides evidence for the benefit of diversity in the training data. One reason could be a wider range of ways to express emotions. We will investigate this in more detail in Section 5.

We consistently see cross-lingual learning capabilities with the mBERT models. The zero-shot scores are significantly above the random baseline (paired one-sided t-test): zero-shot vs. random: $t = 4.08, dof = 24, p < 0.001$ monolingual vs. zero-shot: $t = 3.17, dof = 24, p = 0.002$ cross-lingual vs. monolingual: $t = 1.28, dof = 24, p < 0.105$. Zero-shot vs. random and mono-

Model	Dataset	eng	por	spa	cmn	tgl	Avg
LR	Small-mono	0.45	0.48	0.41	0.46	**0.59**	0.48
	Large-mono	0.52	0.50	0.46			0.49
mBERT	Small-mono	0.46	0.48	0.43	**0.67**	0.53	0.51
	Large-mono	0.58	0.59	0.55			0.57
	Huge English	**0.63**					
	Small-all	0.45	0.50	0.45	**0.67**	0.56	0.53
	Small-IE.	0.49	0.52	0.44	0.41	0.32	0.44
	Small-∼IE.	0.46	0.38	0.36	0.66	0.56	0.48
	Large-all	0.54	**0.60**	0.53			0.56
	Large-all & Small-all	0.57	0.59	**0.56**	**0.67**	0.57	**0.59**

Table 4: Macro-F1 results for mono & cross-lingual learning on the Universal Joy data. mBERT and LR models.

lingual vs. zero-shot are significant at Bonferroni-corrected significance level $a = 0.05/3 = 0.0167$. Cross-lingual frequently performs better than monolingual, but with considerable variation.

4.3 Zero-shot Learning for Low-Resource Languages

Table 7 shows zero-shot results when testing on the low-resource languages with the Small, Large, Large&Small, and Huge training sets. Small is also split typologically: Small-IE consists only of the Indo-European languages `eng`, `spa`, and `por`, while Small ∼IE consists of the remaining languages in Small, `cmn` and `tgl`. We compare against a monolingual result for each language, using a LR system described in Section 3.1, taking the 10-fold cross-validation average. [8] For Indo-European languages, zero-shot models consistently outperform the monolingual model; furthermore, larger zero-shot models tend to do better, although the linguistically diverse L&S model often does better than the much-larger Huge model (which is only English). The zero-shot models rarely do well with the non-Indo-European languages. This is not surprising, since most of the data in the zero-shot models comes from Indo-European languages.

Below we investigate these results in more detail to assess the factors that facilitate cross-lingual learning.

5 Analysis of Cross-lingual Effects

Our results show a wide range of cross-lingual effects. In many cases, they are quite substantial, while in other cases we observe no effect. We believe these differences are due to the linguistic properties of the languages that the models pick up on. Here we examine some of the factors involved in this: code switching, typological closeness, and linguistic diversity. The results can shed light on the similarities and differences in how emotions are expressed in different languages.

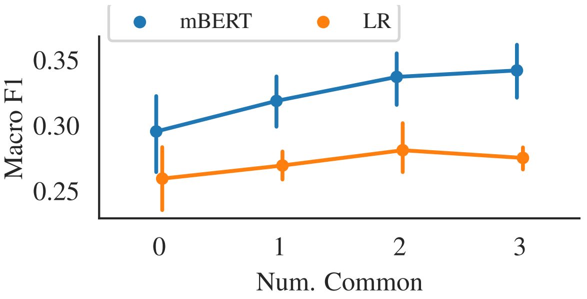

Figure 2: Number of shared WALS features as a function of zero-shot performance for various models.

5.1 Code Switching

A post involves code-switching if it combines multiple languages, as in the following:

"We love you guys ... proud of you.. [PHOTO] jongens heel veel succes vanavond ..!!!!" (the English translation of the Dutch part is *"guys lots of luck tonight"*)

This post is classified as Dutch, but includes text in English. BoW (bag of word) models are therefore well-suited to take advantage of code switching. This post provides information about both Dutch and English, since there are several tokens in each language associated with the labeled emotion, anticipation. A model could learn here that "love" is associated with anticipation and use this on English test data, even though the training example

[8] For languages that have $k < 10$ instances per any emotion, we do a k-fold cross-validation.

Dataset	Language	eng	por	spa	cmn	tgl
Small	eng	**0.46**	0.38	0.34	0.39	0.34
	por	0.44	**0.48**	0.38	0.38	0.31
	spa	0.34	0.39	**0.43**	0.34	0.31
	cmn	0.42	0.34	0.32	**0.67**	0.25
	tgl	0.38	0.34	0.32	0.36	**0.53**
Large	eng	**0.58**	0.38	0.36	0.39	**0.45**
	por	0.42	**0.59**	0.38	**0.40**	0.44
	spa	0.42	0.44	**0.55**	0.38	0.31

Table 5: mBERT macro-F1 results in mono-lingual setting on two sets of universal joy data. Best result for each training dataset per language in bold.

Dataset	Method	Model	eng	por	spa	cmn	tgl	Avg
Small	Zero-shot single-lang avg	LR	0.26	0.24	0.24	0.17	0.22	0.23
		mBERT	0.40	0.36	0.34	0.37	0.30	0.35
	Zero-shot multilingual	LR	0.28	0.28	0.26	0.16	0.26	0.25
		mBERT	0.47	0.40	0.34	0.39	0.41	0.40
Large	Zero-shot single lang avg	LR	0.30	0.29	0.29			0.29
		mBERT	0.42	0.41	0.37			0.40
	Zero-shot multilingual	LR	0.33	0.35	0.31			0.33
		mBERT	0.45	0.45	0.39			0.43

Table 6: Macro-F1 results for zero-shot learning with Small and Large training sets. Single lang avg. is based on average results of models for each language other than test language. Multilingual involves a single model with the same amount of training data as single-language, but evenly mixed among the different languages. Results below the random baseline in gray.

is classified as Dutch.

However, observe in table 6 that the BoW models (LR) perform poorly in zero-shot learning. They are near or below baseline, except in the Large multilingual case. mBERT, by contrast, is consistently above the baseline and consistently better than the BoW models. This suggests that code-switching is *not relevant* to the cross-lingual effects we have observed.

5.2 Typological Closeness

A natural hypothesis, explored by Singh et al. (2019), is that cross-lingual effects are stronger for typologically close languages; that is, scores are higher when training and test language are closely related. Following Pires et al. (2019), we compute typological closeness as overlap on selected features from the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). These features are particularly relevant to word order of categories. In Figure 2, we plot for each pair of languages <l1,l2> the number of WALS features[9] shared between l1 and l2 against the zero-shot score obtained when training on l1 and testing on l2. Indeed, more shared features correlate with better performance, especially in mBERT models.

Table 8 shows the correlation between performance and several other measures of similarity between languages, including the number of shared bigrams and emoticons, the shared WALS features, and the proximity in a genealogy tree. Compare the striking positive correlation between shared WALS features and performance (0.27) for the mBERT model, while there is no such correlation for the two LR models, corroborating Figure 2. This suggests that these particular WALS features related to word order, are relevant to abstract features of the mBERT models, while they are irrelevant for the LR models. In contrast, see the flipped correlation

[9] We keep features that all of our languages have annotations for: *81A* (Order of Subject, Object and Verb), *82A* (Order of Subject and Verb), and *83A* (Order of Object and Verb).

Data	ben	deu	fra	hin	ita	nld	rom	Avg	ind	khm	mya	tha	vie	zsm	Avg
mono	0.26	0.37	0.42	**0.36**	0.32	0.36	0.35	0.35	**0.40**	0.27	**0.36**	**0.33**	**0.39**	**0.47**	**0.37**
S	0.35	0.38	0.42	0.30	0.37	0.39	0.35	0.37	0.34	0.31	0.21	0.32	0.34	0.34	0.31
L	0.34	0.37	0.42	0.29	0.37	0.40	**0.38**	0.37	0.35	0.32	0.29	**0.33**	0.37	0.37	0.34
L&S	**0.38**	0.38	**0.44**	0.30	**0.38**	0.40	0.37	**0.38**	0.35	0.32	0.28	0.32	0.38	0.37	0.34
Huge en	0.34	**0.39**	0.43	0.30	0.37	**0.42**	0.35	0.37	0.36	**0.34**	0.30	0.32	0.36	0.36	0.34

🟩 Indo-European 🟦 non Indo-European

Table 7: mBERT macro-F1 results for zero-shot learning on low-resource languages in Universal Joy data. Improvements hold mainly for IE languages.

Model	Bigrams	Emoticons	WALS	Genealogy
LR	**0.54**	**0.31**	0.01	**0.50**
mBERT	0.18	**0.41**	**0.27**	**0.57**

Table 8: Spearman's rank correlation between performance and various language similarity measures. Significant correlations at $\alpha = 0.05$, Bonferroni-corrected for each model, in bold.

of "lexical" features like bigrams and emoticons in the two model types. Genealogical proximity has a high correlation with performance in all models, but again is highest for mBERT.

All this supports the idea, also discussed in (Pires et al., 2019), that mBERT models are sensitive to abstract syntactic features that are shared across languages.

5.3 Linguistic Diversity

We find clear evidence that diversity of training languages improves performance. Table 6 shows a clear multilingual advantage in zero-shot learning; for all three models, the multilingual scores are higher than the monolingual scores. In table 7, the last two lines compare a diverse training set of 102k instances (Large&Small) to the more than twice as large Huge English training data (283k). The average results over the 13 low-resource languages of both are identical.

6 Related Work

Abdul-Mageed and Ungar (2017) collect tweets based on user-inserted hashtags; the resulting dataset, EmoNet, is similar to ours in that it uses a distant supervision approach. Above we presented results based on the EmoNet dataset and model. The SemEval 2018 Task 1 (Baziotis et al., 2018), involves classification and regression tasks for four emotions: joy, sadness, anger, and fear. The Affect in Tweets Dataset (Mohammad et al., 2018) is a small dataset that includes emotions annotated in multiple languages; it does not involve the cross-linguistic investigation of emotion that is central to the present work. Troiano et al. (2019) describe a small, bilingual emotion dataset, with English and German (ISEAR). Wang et al. (2018) describe a bilingual Chinese-English emotion dataset (NLPCC). We provide results on these datasets in table 9 in the Appendix; it's important to note that these datasets differ in important ways from the Facebook data in our dataset.

7 Conclusion

We introduce a new data set of over 530,000 anonymized Facebook posts from 18 languages, labeled with five basic emotions. We show that emotions can be reliably identified, both within and across languages, including zero-shot learning. This suggests substantial opportunities for transferring knowledge from high-resource to low-resource languages. In a detailed investigation of the factors supporting cross-lingual learning, we find evidence for the importance of linguistic diversity of training data as well as syntactic and typological similarities between languages.

These results provide intriguing evidence of deep commonalities in the linguistic expression of emotion across the languages of the world.

8 Ethical considerations

Collecting and publishing a data set from social media raises a number of ethical concerns. We have prepared and planned the release of this dataset in close consultation with the Data Protection Officer of the main institution for the publication. The data is completely anonymized, with all identifying information removed. It was collected from public postings on Facebook, accessed using the standard Facebook API for collection of such postings. Based on these considerations, the Data Protection Officer approved our plan to release the dataset, and certified that it complies with GDPR and other relevant requirements. The actual data collection was performed by graduate assistants as part of their studies, and the process involve no manual annotation. We also provide a data statement, to allow future researchers to assess any inherent bias. The primary aim of our study is academic, to understand the interplay between language and emotion.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at Semeval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 245—255.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Charles Darwin. [1872] 1998. *The expression of the emotions in man and animals*. Oxford University Press, USA. (Original work published 1872).

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.

Matthew S. Dryer and Martin Haspelmath. 2013. WALS online. https://wals.info/.

Paul Ekman. 2016. What scientists who study emotion agree about. *Perspectives on Psychological Science*, 11(1):31–34.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Robert Plutchik. 1994. *The psychology and biology of emotion.* HarperCollins College Publishers.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. 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.

Zhongqing Wang, Shoushan Li, Fan Wu, Qingying Sun, and Guodong Zhou. 2018. Overview of nlpcc 2018 shared task 1: Emotion detection in code-switching text. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 429–433. Springer.

W. Wundt. 1896. *Grundriss der Psychologie (Outlines of Psychology)*. Leipzig: Engelmann.

Christopher Zimmerman, Mari-Klara Stein, Daniel Hardt, and Ravi Vatrapu. 2015. Emergence of things felt: harnessing the semantic space of facebook feeling tags. In *Proceedings of the Thirty Sixth International Conference on Information Systems*.

A Data Statement

We follow Bender and Friedman (2018) on providing a Data Statement for our corpus, in order to provide a fuller picture of the possibilities and limitations of the data, and to allow future researchers to spot any biases we might have missed.

CURATION RATIONALE We use Facebook postings originally collected in 2014; any identifying information of the authors has been removed by anonymization.

LANGUAGE VARIETY Eighteen different languages, as identified by language classification, therefore presumably mostly standard. Due to the setting (Facebook posts), some non-standard language is likely.

SPEAKER DEMOGRAPHICS Unknown, though gender could be inferred from first names before anonymization. Due to the setting, all authors need to have access to internet, which means a young demographic is likely.

SPEECH SITUATION Facebook posts self-labeled with emotions – i.e., short, written, spontaneous texts written synchronously with a broad audience in mind.

TEXT CHARACTERISTICS Wide range of topics, but confined broadly to emotional issues.

B Macro-F1 score for the results on other datasets

Results on additional multilingual datasets are given in table 9

C Datasets

The testing datasets are displayed in table 11.

D Logistic Regression Results

Table 12 gives monolingual and zero-shot results for the two logistic regression models, Uni-LR an BT-LR. Note that zero-shot results are generally at or below baseline.

E Code switching & Emoticons

For extracting the bigram types, we remove all punctuation and split the sentences into tokens using `nltk` (Loper and Bird, 2002) `word_tokenizer`. We keep bigrams with more than 5 occurences. For extracting the emoticons/punctuation, we keep only Unicode punctuation characters and use `nltk TweetTokenizer` to split the sentences into tokens. We discard any token consisting of a single character and replace all occurences of more than 3 consecutive punctuation characters with just 3 characters. We keep tokens with more than 5 occurences.

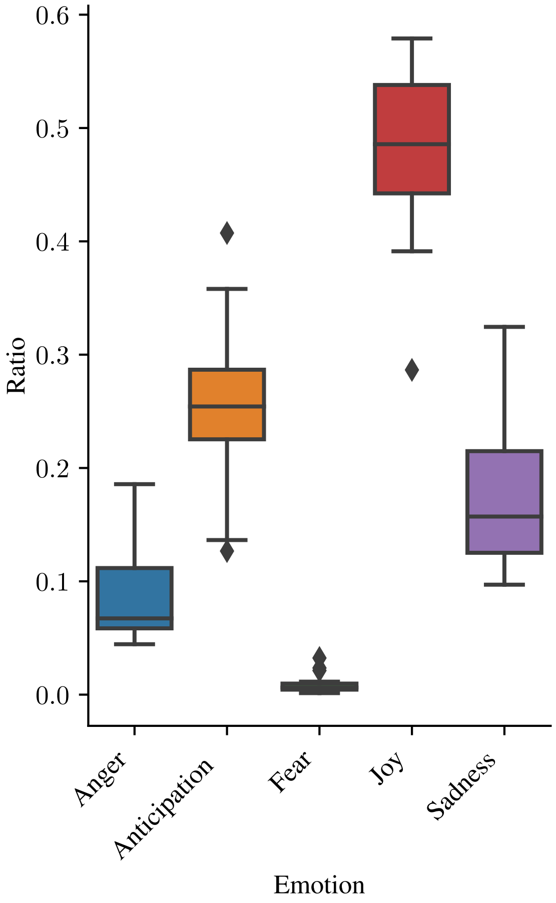

Figure 3: Box plot of emotion ratios over all languages

E.1 Expanded Results

Tables 16 through 18 give results for mBERT Small models, separated out for the different emotions.

Dataset	Language(s)	EmoNet	ISEAR eng	ISEAR deu	NLPCC	Avg
baseline		0.47	**0.52**	**0.50**	0.40	0.47
Small	eng	0.40	0.48	0.42	0.45	0.44
	cmn	0.31	0.40	0.30	0.43	0.36
	IE.	0.36	0.46	0.40	0.40	0.41
	∼IE.	0.37	0.46	0.38	0.42	0.41
	all	0.43	0.44	0.36	0.43	0.41
Large	eng	0.48	0.46	0.36	0.42	0.43
	all	0.46	0.43	0.31	0.41	0.40
Large & Small	all	0.50	0.48	0.38	0.44	0.45
Huge	eng	**0.55**	0.49	0.39	**0.48**	**0.48**

Table 9: Results on other datasets

Original Emotion	Mapped Emotion
accomplished	
amused	
angry	anger
annoyed	anger
awesome	
bad	
confident	
confused	
depressed	sadness
determined	
disappointed	sadness
disgusted	
down	sadness
excited	anticipation
fantastic	joy
great	joy
happy	joy
heartbroken	sadness
hopeful	anticipation
pissed	anger
proud	
pumped	anticipation
sad	sadness
scared	fear
super	joy
wonderful	joy
worried	fear

Table 10: Mapping from Facebook emotion to the 5 basic emotions.

Dataset	Lang.	Anger	Anticip.	Fear	Joy	Sadness	Total
UJ Testing	eng	58	400	11	384	128	981
	spa	56	228	5	538	154	981
	por	56	254	7	418	246	981
	cmn	92	218	21	533	117	981
	tgl	129	183	32	433	204	981
UJ Low Resource	ben	120	211	7	249	282	869
	deu	425	1475	8	3388	606	5902
	fra	382	1788	22	3222	1143	6557
	hin	274	231	8	830	480	1823
	ind	382	1841	32	3077	869	6201
	ita	472	1910	20	3656	651	6709
	khm	115	158	23	469	212	977
	zsm	326	1344	34	2566	638	4908
	mya	177	130	9	412	225	953
	nld	150	788	10	981	272	2201
	rom	97	560	8	923	352	1940
	tha	244	938	21	2202	398	3803
	vie	176	1137	39	1982	622	3956

Table 11: Testing Datasets

Model	Dataset	Language	eng	por	spa	cmn	tgl
Uni-LR	Small	eng	0.40	0.30	0.25	0.30	0.29
		por	0.28	0.48	0.31	0.25	0.27
		spa	0.27	0.38	0.38	0.26	0.25
		cmn	0.32	0.25	0.26	0.43	0.28
		tgl	0.31	0.29	0.26	0.26	0.56
	Large	eng	0.51	0.28	0.27	0.29	0.44
		por	0.28	0.52	0.35	0.27	0.29
		spa	0.32	0.37	0.47	0.31	0.29
BT-LR	Small	eng	0.45	0.21	0.22	0.24	0.26
		por	0.25	0.48	0.32	0.21	0.19
		spa	0.24	0.32	0.41	0.15	0.21
		cmn	0.29	0.19	0.21	0.46	0.21
		tgl	0.26	0.22	0.23	0.10	0.59
	Large	eng	0.52	0.23	0.26	0.21	0.41
		por	0.28	0.50	0.33	0.21	0.29
		spa	0.31	0.36	0.46	0.21	0.28

Table 12: Macro-F1 results in mono-lingual setting for the Logistic Regression models. Results below the baseline in gray.

Bigram	# Languages
i love	14
love you	12
good morning	9
of the	9
new year	9
the best	8
this is	8
in the	8
have a	8
thank you	7
happy birthday	7
so much	6
coming soon	6
we are	6
for the	6
i am	6
on the	5
see you	5
happy new	5
a nice	5
more info	4
to be	4
make up	4
you all	4
like share	4

Table 13: Prevalence of common bigrams between languages

Pattern	# Languages
!!	18
!!!	18
...	18
???	18
:)	18
..	18
:D	17
??	16
:(	16
***	16
!.	16
;)	16
!!!.	15
:-)	15
:’(	15
).	15
...!!!	15
,,,	15
...!	14
.(	14
”	14
,,	14
...#	13
!!.	13
!...	13

Table 14: Prevalence of common punctuation patterns between languages

language	anger	anticipation	fear	joy	sadness	avg
eng	0.11	0.58	0.02	0.56	0.23	0.30
por	0.11	0.41	0.01	0.60	0.40	0.31
spa	0.11	0.38	0.01	0.71	0.27	0.3
cmn	0.17	0.36	0.04	0.70	0.21	0.3
tgl	0.23	0.31	0.06	0.61	0.34	0.31
avg	0.15	0.41	0.03	0.64	0.29	0.30

Table 15: F1-scores for random baseline.

language	anger	anticipation	fear	joy	sadness	avg
eng	0.45	0.63	0.15	0.59	0.49	0.46
por	0.39	0.54	0.13	0.66	0.68	0.48
spa	0.25	0.44	0.29	0.71	0.46	0.43
cmn	0.67	0.47	0.86	0.75	0.61	0.67
tgl	0.47	0.45	0.52	0.65	0.55	0.53
avg	0.45	0.51	0.39	0.67	0.56	0.51

Table 16: F1-scores for mono-lingual classification.

language	anger	anticipation	fear	joy	sadness	avg
eng	0.42	0.65	0.24	0.58	0.49	0.47
por	0.28	0.54	0.00	0.61	0.55	0.4
spa	0.19	0.47	0.00	0.63	0.41	0.34
cmn	0.44	0.42	0.00	0.69	0.41	0.39
tgl	0.26	0.38	0.43	0.53	0.44	0.41
avg	0.32	0.49	0.13	0.61	0.46	0.40

Table 17: F1-scores for zero-shot multi-lingual classification

language	anger	anticipation	fear	joy	sadness	avg
eng	0.43	0.66	0.07	0.56	0.54	0.45
por	0.41	0.53	0.25	0.65	0.65	0.50
spa	0.33	0.49	0.18	0.72	0.53	0.45
cmn	0.63	0.47	0.89	0.77	0.62	0.67
tgl	0.51	0.45	0.65	0.64	0.57	0.56
avg	0.46	0.52	0.41	0.67	0.58	0.53

Table 18: F1-scores for cross-lingual classification