XLM-EMO: Multilingual Emotion Prediction in Social Media Text

Federico Bianchi
Bocconi University
Via Sarfatti 25
Milan, Italy
f.bianchi@unibocconi.it

Debora Nozza
Bocconi University
Via Sarfatti 25
Milan, Italy
debora.nozza@unibocconi.it

Dirk Hovy
Bocconi University
Via Sarfatti 25
Milan, Italy
dirk.hovy@unibocconi.it

Abstract

Detecting emotion in text allows social and computational scientists to study how people behave and react to online events. However, developing these tools for different languages requires data that is not always available. This paper collects the available emotion detection datasets across 19 languages. We train a multilingual emotion prediction model for social media data, XLM-EMO. The model shows competitive performance in a zero-shot setting, suggesting it is helpful in the context of low-resource languages. We release our model to the community so that interested researchers can directly use it.

1 Introduction

Emotion Detection is an important task for Natural Language Processing and for Affective Computing. Indeed, several resources and models have been proposed (Alm et al., 2005; Abdul-Mageed and Ungar, 2017; Nozza et al., 2017; Xia and Ding, 2019; Demszky et al., 2020, inter alia) for this task. These models can be used by social and computational scientists (Verma et al., 2020; Kleinberg et al., 2020; Huguet Cabot et al., 2020) to better understand how people react to events through the use of social media. However, these methods often require large training sets that are not always available for low-resource languages. Nonetheless, multilingual methods (Wu and Dredze, 2019) have risen across the entire field showing powerful few-shot and zero-shot capabilities (Bianchi et al., 2021b; Nozza, 2021).

In this short paper, we introduce a new resource: XLM-EMO. XLM-EMO is a model for multilingual emotion prediction on social media data. We collected datasets for emotion detection in 19 different languages and mapped the labels of each dataset to a common set *{joy, anger, fear, sadness}* that is then used to train the model. We show that XLM-EMO is capable of maintaining stable performances across languages and it is competitive against language-specific baselines in zero-shot settings.

We believe that XLM-EMO can be of help to the community as emotion prediction is becoming an interesting and relevant task in NLP; the addition of a multilingual model that can perform zero-shot emotion prediction can be of help for many low-resource languages that still do not have a dataset for emotion detection.

Contributions We release XLM-EMO which is a multilingual emotion detection model for social media text. XLM-EMO shows competitive zero-shot capabilities on unseen languages. We release the model in two versions a base and a large to adapt to different possible use-cases. We make the models[1] and the code to train it freely available under a Python package that can be directly embedded in novel data analytics pipelines.[2]

2 Data and Related Work

We surveyed the literature to understand which datasets are available in the literature and with which kinds of emotions. Details on how we operate on this data can be found in the Appendix, here we give an overview of the transformation pipeline we have adopted and which datasets have been included.

[1] Models can be found at https://huggingface.co/MilaNLProc/

[2] See https://github.com/MilaNLProc/xlm-emo, where we also release other details for replication.

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 195 - 203

The datasets we have collected and used in this paper are presented in Table 1 with the method of annotation and the linguistic family of the language. Figure 1 shows instead the class distribution.

We describe here the general guidelines we have used to create this dataset, readers can find details for each dataset in the Appendix. For all the datasets we removed the emotions that are not in the set *joy, anger, fear, sadness* (e.g., Cortiz et al. (2021), Vasantharajan et al. (2022), Shome (2021) used the 27 emotions from GoEmotion (Demszky et al., 2020) and we just collected the subset of our emotions). We have some exceptions to Twitter data, as the Tamil dataset Vasantharajan et al. (2022) contains YouTube comments.

Some data was impossible to reconstruct because the tweets do not exist anymore and thus only a subset is still available (e.g., Korean (Do and Choi, 2015)). For some languages, we decided to apply undersampling in order to limit the skewness of the final distribution (e.g., both Shome (2021) and Cortiz et al. (2021) provide dozens of thousands of tweets). To simplify reproducibility, we will release the exact data extraction scripts that we have used to collect our data.

There are papers that we have not included in our research: Vijay et al. (2018) introduce a Hindi dataset that contains Hindi-English code switched text. However, Hindi is Romanized and only a few of this data has been used to pre-train XLM. Sabri et al. (2021) released a collection of Persian tweets annotated with emotions, however, their data has not been evaluated in a training task and thus we decided not to include it in our training. We also found a dataset for Japanese Danielewicz-Betz et al. (2015), however, the dataset is not publicly available.

French and German are collected through the translation of Spanish (Mohammad et al., 2018) tweets using DeepL.[3] For Chinese, we use the messages found in the NLPCC dataset (Wang et al., 2018). Note that this dataset has some internal code-switching.

The most similar work to ours is the work by Lamprinidis et al. (2021). Lamprinidis et al. (2021) introduces a dataset collected through distant supervision on Facebook and covers 6 main languages for training and a set of 12 other languages that can be used for testing. We will run a comparison with this model in Section 3.3.

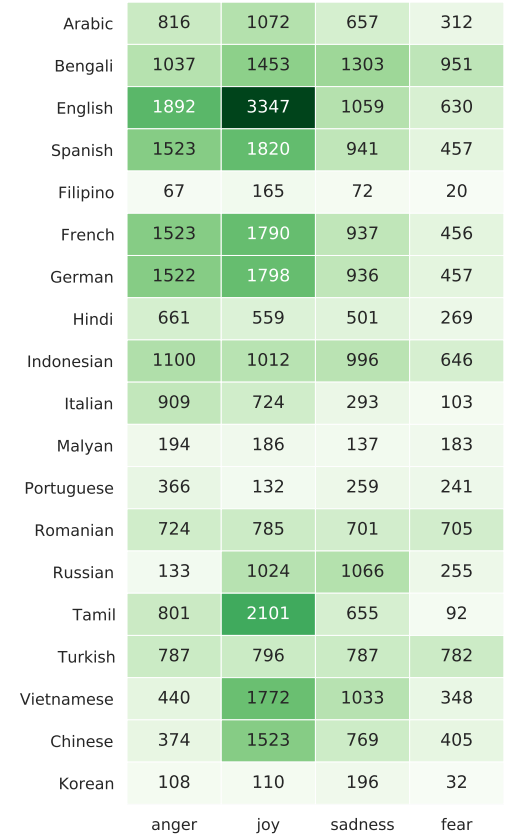

Figure 1: Label distribution. German, French have different numbers because some API translations failed.

3 Experiments

We perform three different experiments. The first one is meant to show the performance of XLM-EMO across the different languages. The second one evaluates how well XLM-EMO works on a zero-shot task in which data from one language is held out; we focus on testing three languages: English, Arabic, and Vietnamese. The third evaluation shows the performance of XLM-EMO on additional datasets different from those used for training on which we compare our model with other state-of-the-art models.

3.1 Performance on Test Set

We fine-tune 3 different models: XLM-RoBERTa-base (Conneau et al., 2020), XLM-RoBERTa-large (Conneau et al., 2020) and Twitter-XLM-RoBERTa (Barbieri et al., 2021). The first two are trained on data from 100 languages while the latter is a fine-tuned version of XLM-RoBERTa-base on Twitter data.

We use 10% for validation (we evaluate the

[3]We are aware that this process might introduce bias in the model as described by Hovy et al. (2020)

Language	Reference	Method	Family
English	Mohammad et al. (2018)	Manual Annotation	Indo-European
Spanish	Mohammad et al. (2018)	Manual Annotation	Indo-European
Arabic	Mohammad et al. (2018)	Manual Annotation	Afroasiatic
French	-	Translation	Indo-European
German	-	Translation	Indo-European
Chinese	Wang et al. (2018)	Manual Annotation	Sino-Tibetan
Korean	Do and Choi (2015)	Manual Annotation	Koreanic
Romanian	Ciobotaru and Dinu (2021)	Manual Annotation	Indo-European
Russian	Sboev et al. (2020)	Manual Annotation	Indo-European
Indonesian	Saputri et al. (2018)	Manual Annotation	Austronesian
Bengali	Iqbal et al. (2022)	Manual Annotation	Indo-European
Italian	Bianchi et al. (2021a)	Manual Annotation	Indo-European
Portuguese	Cortiz et al. (2021)	Distant Supervision	Indo-European
Turkish	Güven et al. (2020)	Distant Supervision	Turkic
Filipino	Lapitan et al. (2016)	Manual Annotation	Austronesian
Malay	Husein (2018)	Distant Supervision	Austronesian
Hindi	Shome (2021)	Translation	Indo-European
Vietnamese	Ho et al. (2019)	Manual Annotation	Austroasiatic
Tamil	Vasantharajan et al. (2022)	Manual Annotation	Dravidian

Table 1: Languages used in this work

Language	**Lang-Specific** (large)	**XLM-EMO ZeroShot** (large)	**XLM-EMO Trained** (large)
Arabic	**0.91**	0.81	0.88
English	0.83	0.82	**0.85**
Vietnamese	**0.84**	0.77	0.82

Table 2: Comparison between the language-specific models, the zero-shot XLM-EMO and an XLM-EMO that has been trained also on the additional data used for language-specific models plus all the other languages. Results are computed over the average of 5 different seeds.

Model	ME	EE-EN	EE-ES
XLM-EMO	**0.62**	**0.66**	**0.73**
LS-EMO	0.58	0.44	-
UJ-Combi	0.35	0.52	0.51

Table 3: Results on the Out of Domain test. XLM-EMO performs better than the selected baseline.

model every 50 steps and get the best checkpoint) and 5% of data for the test. Figure 2 shows the comparison between the three different models averaged on 5 runs with different seeds. These results show that the model is able to maintain a stable performance even when trained on data from 19 languages. The overall average Macro-F1s for XLM-RoBERTa-large, XLM-RoBERTa-base and XLM-Twitter-base are 0.86, 0.81 and 0.84.

The results also indicate that XLM-RoBERTa-large is the best model; however, XLM-Twitter-base performs better than XLM-RoBERTa-base and this is probably because it is a Twitter-specific model. Unfortunately, at this date, a large version of XLM-Twitter does not exist.

For all languages but Korean and Filipino, the performance is reliable. This is probably because both do not occur frequently in the training data. It should be noted that also Chinese and Tamil have a performance that is slightly above 0.6 with the large model. Considering these results, we will refer to the fine-tuned XLM-RoBERTa-large as XLM-EMO and we will use it in the rest of the paper.

3.2 Zero-shot Tests

We run 3 zero-shot comparisons to show the model performance on unseen languages. We select Arabic, English, and Vietnamese. Target language data is split into training and test (80/20). A language-specific model is trained (we again select the best model based on checkpoints on validation that is 10% of the training data). We use language-specific BERT-large for all the three languages. [4][5][6].

[4] https://huggingface.co/bert-large-uncased

[5] https://huggingface.co/aubmindlab/bert-large-arabertv02-twitter

[6] https://huggingface.co/vinai/phobert-large

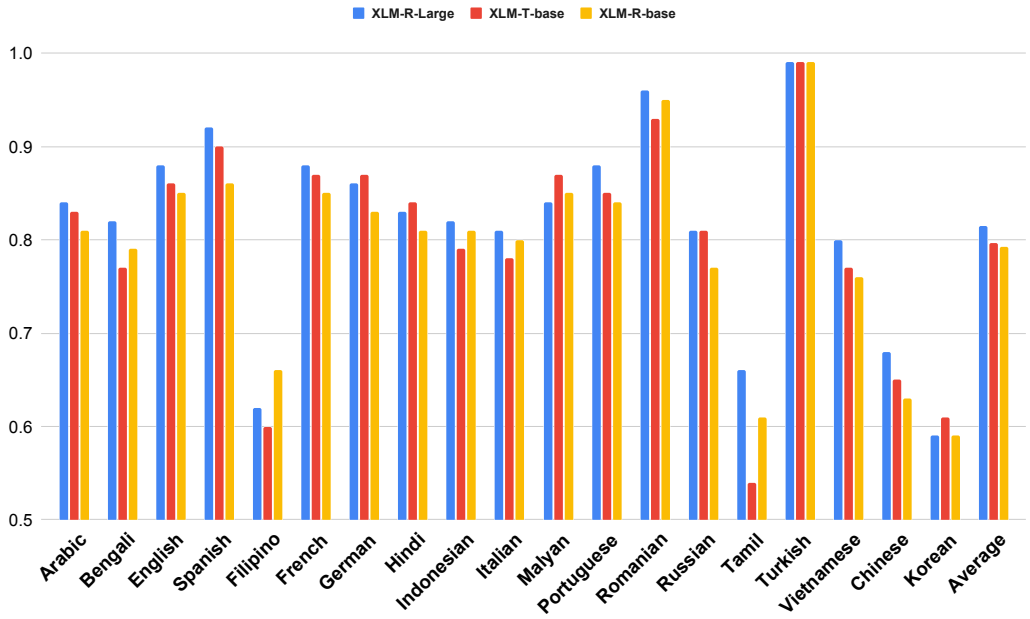

Figure 2: The performance (Macro-F1) of the three fine-tuned models across the various languages present in the test set. XLM-RoBERTa-large has the best performance. We averaged the run of 5 different seeds.

We also use an XLM-EMO trained on all the languages plus the 80% training data also used for the language-specific model.

Results in Table 2 show that XLM-EMO is competitive in the zero-shot settings. Still, language-specific models beat both the zero-shot and the model with additional training data.[7] On English data, XLM-EMO Trained seems to show better performance than the language-specific model, but this is probably because in language-specific datasets some English data might still be present.

3.3 Comparison with Available Models

We compare how XLM-EMO (large) behaves against out-of-training data to better understand if it generalizes well in other domains. In this test, we use other models to see how they perform in comparison with our XLM-EMO.

As datasets, we use the MultiEmotion Italian dataset (ME) (Sprugnoli, 2020) that contains YouTube and Facebook comments annotated with emotions (we collect only the comments with emotions that overlap with ours) and the EmoEvent dataset (EE) in English and Spanish (Plaza del Arco et al., 2020).[8] For both datasets we filtered out only the text that has been annotated with one of the labels we also use.

Respectively, as language-specific competitors (LS-EMO), we use the FEEL-IT (Bianchi et al., 2021a) as found on HuggingFace[9] and EmoNet Abdul-Mageed and Ungar (2017) as found on GitHub [10]. In addition, we also compare with the multilingual baseline Universal Joy (UJ) (Lamprinidis et al., 2021), using their *combi* model that has been trained on 6 languages (English, Spanish, Portuguese, Tagalog, Indonesian, and Chinese); note that, Italian has not been seen by the UJ model during training.

EmoNet and UJ predict additional emotions. To be as a fair as possible, we filter out the missing emotions from the predicted logits so that both models predict only *joy, anger, sadness, and fear*. The results in Table 3 show that XLM-EMO is the best performing model.

4 Limitations

Unfortunately, we have not been able to find datasets for emotions detection in any of the African Languages. Moreover, automatic translation tools do not often cover African languages or

[7] Similar conclusions have been reached by Nozza et al. (2020).

[8] We could not find another Spanish model to test against this data since the Spanish emotion recognition model (Pérez et al., 2021a,b) is trained on this data.

[9] https://huggingface.co/MilaNLProc/feel-it-italian-emotion

[10] https://github.com/UBC-NLP/EmoNet

they do not provide reliable evidence of being able to provide those translations with a certain level of quality. We reached out to members of our community to understand if there was any work that we were not aware of but we did not find any. Further iterations of this resource might want to focus on those languages.

5 Conclusion

In this short paper, we propose XLM-EMO, a novel resource for emotion detection. The model shows stable performance across 19 languages and it is competitive in a zero-shot setting, supporting its usage in low-resource contexts. We plan to enrich this model with more languages as soon as we find them so that we can continually improve these results and offer better methods to the community.

Acknowledgements

This project has partially received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR), and by Fondazione Cariplo (grant No. 2020-4288, MONICA). Federico Bianchi, Debora Nozza, and Dirk Hovy are members of the MilaNLP group, and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

Ethical Considerations

There is still a mismatch in the adoption of the methods we release and our understanding of them (Bianchi and Hovy, 2021). We are releasing a resource for multi-lingual emotion detection, but any list of language resources runs the risk of being (mis)interpreted as exhaustive, with languages included being regarded as more important than those that are not. We would like to emphatically state that this is not the case here: we tried to include as many languages as possible to allow for a wide comparison and provide a basis for further research. Any omission should not be read as a value judgment.

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.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. XLM-T: A multilingual language model toolkit for Twitter. *arXiv preprint arXiv:2104.12250*.

Federico Bianchi and Dirk Hovy. 2021. On the gap between adoption and understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3895–3901, Online. Association for Computational Linguistics.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021a. FEEL-IT: Emotion and sentiment classification for the Italian language. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 76–83, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

Alexandra Ciobotaru and Liviu P. Dinu. 2021. RED: A novel dataset for Romanian emotion detection from tweets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 291–300, Held Online. INCOMA Ltd.

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.

Diogo Cortiz, Jefferson O. Silva, Newton Calegari, Ana Luísa Freitas, Ana Angélica Soares, Carolina Botelho, Gabriel Gaudencio Rêgo, Waldir Sampaio, and Paulo Sergio Boggio. 2021. A weak supervised dataset of fine-grained emotions in Portuguese. *Symposium in Information and Human Language Technology*.

A. Danielewicz-Betz, , H. Kaneda, M. Mozgovoy, M. Purgina, , and and. 2015. Creating English and Japanese Twitter corpora for emotion analysis. *International Journal of Knowledge Engineering-IACSIT*, 1(2):120–124.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Hyo Jin Do and Ho-Jin Choi. 2015. Korean Twitter emotion classification using automatically built emotion lexicons and fine-grained features. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 142–150, Shanghai, China.

Zekeriya Anil Güven, Banu Diri, and Tolgahan Çakaloğlu. 2020. Comparison of n-stage latent dirichlet allocation versus other topic modeling methods for emotion analysis. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 35(4):2135–2145.

Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2019. Emotion recognition for Vietnamese social media text. In *Computational Linguistics - 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11-13, 2019*, volume 1215 of *Communications in Computer and Information Science*, pages 319–333. Springer.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "You sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

Zolkepli Husein. 2018. Malay-dataset, we gather bahasa malaysia corpus!, semi-supervised emotion dataset. https://github.com/huseinzol05/malay-dataset/tree/master/corpus/emotion.

MD. Asif Iqbal, Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H. Sarker. 2022. BEmoC: A corpus for identifying emotion in Bengali texts. *SN Computer Science*, 3(2):135.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring Emotions in the COVID-19 Real World Worry Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal joy a data set and results for classifying emotions across languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, Online. Association for Computational Linguistics.

Fermin Roberto Lapitan, Riza Theresa Batista-Navarro, and Eliezer Albacea. 2016. Crowdsourcing-based annotation of emotions in Filipino and English tweets. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 74–82, Osaka, Japan. The COLING 2016 Organizing Committee.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? Making sense of language-specific BERT models. *arXiv preprint arXiv:2003.02912*.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2017. A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 273–280, Valencia, Spain. Association for Computational Linguistics.

Flor Miriam Plaza del Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France. European Language Resources Association.

Juan Manuel Pérez, Damián A. Furman, Laura Alonso Alemany, and Franco Luque. 2021a. RoBERTuito: a pre-trained language model for social media text in Spanish. *arXiv preprint arXiv:2111.09453*.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021b. pysentimiento: A Python toolkit for

sentiment analysis and socialnlp tasks. *arXiv preprint arXiv:2106.09462*.

Nazanin Sabri, Reyhane Akhavan, and Behnam Bahrak. 2021. EmoPars: A collection of 30K emotion-annotated Persian social media texts. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 167–173, Online. INCOMA Ltd.

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on Indonesian Twitter dataset. In *2018 International Conference on Asian Language Processing, IALP 2018, Bandung, Indonesia, November 15-17, 2018*, pages 90–95. IEEE.

Alexander G. Sboev, Aleksandr Naumov, and Roman B. Rybka. 2020. Data-driven model for emotion detection in Russian texts. In *Proceedings of the 2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence, BICA 2020*, volume 190 of *Procedia Computer Science*, pages 637–642. Elsevier.

Debaditya Shome. 2021. EmoHinD: Fine-grained multi-label emotion recognition from Hindi texts with deep learning. In *12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021*, pages 1–5. IEEE.

Rachele Sprugnoli. 2020. MultiEmotions-It: a new dataset for opinion polarity and emotion analysis for Italian. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Charangan Vasantharajan, Sean Benhur, Prasanna Kumar Kumarasen, Rahul Ponnusamy, Sathiyaraj Thangasamy, Ruba Priyadharshini, Thenmozhi Durairaj, Kanchana Sivanraju, Anbukkarasi Sampath, Bharathi Raja Chakravarthi, and John Phillip McCrae. 2022. Tamilemo: Finegrained emotion detection dataset for tamil. *arXiv preprint arXiv:2202.04725*.

Reyha Verma, Christian von der Weth, Jithin Vachery, and Mohan Kankanhalli. 2020. Identifying worry in Twitter: Beyond emotion analysis. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 72–82, Online. Association for Computational Linguistics.

Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Corpus creation and emotion prediction for Hindi-English code-mixed social media text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 128–135, New Orleans, Louisiana, USA. Association for Computational Linguistics.

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 *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018*, volume 11109 of *Lecture Notes in Computer Science*, pages 429–433. Springer.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Param	Value
Batch Size	64
Warm Up Steps	50
Learning Rate	1e-3
Learning Epochs*	5
Optimizer	AdamW
Betas	0.9 and 0.999
Max Length	100

Table 4: The main parameters we used to run the models. *While epochs are 5, we remark that we are running a step-wise evaluation.

A Training Details

A.1 Parameters

All the models are trained with the same pipeline. We report the shared parameters in Table 4. The only difference can be found in the experiments presented in Section 3.2, the zero-shot tests. Since the language-specific datasets contain less data, we reduced the number of steps for which we run the evaluation and create a checkpoint (i.e, we evaluate every 5 steps).

The loss we use is weighted with respect to the frequency of each label.

This configuration was obtained after several grid search experiments, we found that one of the parameter that impacts the most the training of *large* configurations of the models is the batch size. Models are trained on a Nvidia GeForce RTX 2080 Ti.

A.2 Pre-processing

We align our pre-processing to the one described in (Barbieri et al., 2021), replacing user tags with

@*user* and links with *http*. For those datasets that had a different pre-processing (e.g., some datasets used @*username* to replace user tags) we applied a normalization procedure to align them with our pre-processing.

PhoBERT Note that the Vietnamese model requires a particular pre-processing pipeline: as suggested by the authors on their own GitHub page, for this specific model we apply segmentation on the Vietnamese text.

B Dataset Details

In general, when a message is annotated with multiple emotions we remove it from the dataset. When a dataset comes with multiple emotions that could overlap (e.g., *joy* and *enthusiasm*), we just select the emotions of our interest and we do not apply any mapping (e.g., treating *enthusiasm* messages as *joy*). This is done to avoid bias in the final collection.

We are going to release also our entire processing pipeline (that is mainly based on data transformations) so that interested researchers can re-run it. Note that all the samplings we do have been run with a fixed seed so that they are reproducible.

Arabic This data come from the Affects In Tweet dataset (Mohammad et al., 2018). We combine train, validation and test in a single dataset but we drop emotions that are not covered by our set of emotions.

Bengali This dataset contains data coming from a different source, such as youtube comments and Facebook posts. We only take the messages with emotions that are part of our set.

English This data come from the Affects In Tweet dataset (Mohammad et al., 2018). We combine train, validation and test in a single dataset but we drop emotions that are not covered by our set of emotions.

Spanish This data come from the Affects In Tweet dataset (Mohammad et al., 2018). We combine train, validation and test in a single dataset but we drop emotions that are not covered by our set of emotions.

Filipino This is one of the languages with a lower amount of data. The number of tweets in Filipino (Lapitan et al., 2016) was already low in the original work (i.e., 647) and the final number is even lower since we removed the emotions that do not overlap with ours.

French For this language, we translated the training data that comes from the Spanish subset of the Affects In Tweet dataset (Mohammad et al., 2018).

German For this language, we translated the training data that comes from the Spanish subset of the Affects In Tweet dataset (Mohammad et al., 2018).

Hindi This dataset comes from a translation of the original GoEmotion dataset (Demszky et al., 2020). We just selected the emotions we are interested in and removed the others. Since this dataset has been translated with Google API we opted for sampling only 2000 examples not to bias the representation too much.

Indonesian We collected this dataset directly from the authors work (Saputri et al., 2018), we dropped the *love* emotions and we mapped *happy* to our emotion *joy*.

Italian This dataset comes from the work of Bianchi et al. (2021a), their labels overlap with ours.

Malyan We were slightly less confident on the quality of the annotations of this dataset and we thus sampled 200 messages for each emotion.

Portuguese This dataset has been collected using a keyword search of terms related to emotions. We focus only on our target emotions and randomly sample a maximum of 1000 tweets. This is done because the keyword used for the emotions are few and we would like to avoid biasing the actual representation.

Romanian This dataset (Ciobotaru and Dinu, 2021) has been collected by scraping Twitter using specific keywords. The emotions considered are 5, where the additional one is *neutral*, which we remove. As our data, we used both the training and the validation data released by the authors.

Russian We mainly focused on Twitter data and from the Russian dataset Sboev et al. (2020) we extract only the data that comes from Twitter. We remove the tweets with *neutral* label.

Tamil The Tamil dataset contains YouTube comments and we use the training dataset described by the authors. We decided to remove the long tail of

messages that have more than 30 tokens to make the dataset more consistent with the other datasets. Our labels are a subset of the labels described in the paper and we take only the messages with those labels.

Turkish The Turkish dataset contains 5 emotions, one of which is *surprise* that was removed from our datasets.

Vietnamese This dataset contains youtube comments and has been manually annotated. We drop the emotions that are not covered in our dataset.

Chinese This dataset comes from the challenge described by (Wang et al., 2018). It contains Chinese messages, some of which contain English words (it is a code-switching dataset).

Korean The Korean dataset contains tweets that we reconstructed using the Twitter API. Since the release of the dataset, most tweets have been deleted or are not available anymore for other reasons. The dataset contains the *Neutral* label that we filter out. The other labels easily map onto ours.