Knowledge Distillation from Monolingual to Multilingual Models for Intelligent and Interpretable Multilingual Emotion Detection

Yuqi Wang[1,2], Zimu Wang[1,2], Nijia Han[1], Wei Wang[1,†], Qi Chen[1], Haiyang Zhang[1], Yushan Pan[1], Anh Nguyen[2]
[1]Xi'an Jiaotong Liverpool University [2]University of Liverpool
{yuqi.wang17,zimu.wang19,nijia.han23}@student.xjtlu.edu.cn
{wei.wang03,qi.chen02,haiyang.zhang,yushan.pan}@xjtlu.edu.cn
anh.nguyen@liverpool.ac.uk

Abstract

Emotion detection from text is a crucial task in understanding natural language with wide-ranging applications. Existing approaches for multilingual emotion detection from text face challenges with data scarcity across many languages and a lack of interpretability. We propose a novel method that leverages both monolingual and multilingual pre-trained language models to improve performance and interpretability. Our approach involves 1) training a high-performing English monolingual model in parallel with a multilingual model and 2) using knowledge distillation to transfer the emotion detection capabilities from the monolingual teacher to the multilingual student model. Experiments on a multilingual dataset demonstrate significant performance gains for refined multilingual models like XLM-RoBERTa and E5 after distillation. Furthermore, our approach enhances interpretability by enabling better identification of emotion-trigger words. Our work presents a promising direction for building accurate, robust and explainable multilingual emotion detection systems.

1 Introduction

Emotion detection, a sub-category of sentiment analysis, is the process of computationally identifying, extracting and categorising the emotion expressed in text. This granular analysis of affective states, including joy, sadness, fear, and anger, represents a crucial task in natural language understanding (NLU) that has garnered substantial research attention for several decades due to its wide range of applications. The growth of social media platforms, such as Facebook and Twitter, has led to an increasing trend of individuals sharing their emotions, thoughts and experiences through short snippets of text in posts, tweets, comments and captions. Consequently, a vast volume of user-generated data enriched with emotional content has been generated, highlighting the immense value of automating the detection and analysis of underlying emotions.

Despite the significant progress made in emotion detection from text, there remain two key challenges that hinder the widespread applicability and trustworthiness of these systems: 1) the majority of existing work focused on developing models for high-resource languages like English, where large task-specific datasets for emotion detection, such as CANCEREMO (Sosea and Caragea, 2020) and EmoNet (Abdul-Mageed and Ungar, 2017), are readily available for training and fine-tuning. On the other hand, there is not enough training data for many minority languages, such as French and Dutch. This data scarcity poses a significant obstacle in building accurate and robust emotion detection models that can cater to the linguistic and cultural diversity present across different communities; 2) while current models excel at overall emotion classification, they often lack the ability to provide explanations or insights into the specific linguistic cues that triggered the detected emotions. Many previous studies have primarily concentrated on maximising the overall accuracy metrics (Wang et al., 2021; Wang and Gan, 2023), overlooking the importance of interpretability and rationale extraction.

Recent advances in multilingual pre-trained language models (multilingual PLM) present a promising direction. These models, such as mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), and mDeberta (He et al., 2021), are pre-trained on large volumes of unlabeled text from multiple languages in an unsupervised manner, allowing them to capture rich cross-lingual representations that can be effectively transferred to downstream tasks like emotion detection. They play a crucial role in enabling cross-lingual trans-

†Corresponding author.

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 470–475

fer and facilitating joint training across different languages (Ruder et al., 2019). Furthermore, the transformer-based architecture with self-attention mechanisms can potentially provide interpretability benefits (Chefer et al., 2021), allowing us to analyse the important trigger words that contribute to the predictions. However, the task-specific performance of a fine-tuned multilingual PLM on the multilingual data may not be as comparable to that of separately trained monolingual language models evaluated on the data from their respective languages (Lothritz et al., 2021; Wu and Dredze, 2020). To address the above-mentioned issue, in this work, we propose a novel approach that combines the strengths of both monolingual and multilingual models for improved emotion detection performance and interpretability across diverse languages.

2 Methodology

2.1 Parallel Model Adaptation for Emotion Detection

Our proposed approach involves two separate training pipelines leveraging both monolingual and multilingual PLMs. In the first pipeline, we finetune a state-of-the-art English monolingual model (such as RoBERTa (Liu et al., 2019)), denoted as θ_{en}, on the provided English emotion detection training set, enabling the model to capture the linguistic meanings and emotion cues that are specific to the English language; in parallel, we fine-tune a multilingual PLM like XLM-RoBERTa (Conneau et al., 2020), denoted as θ_{mul}, on the same English set. The multilingual PLM, pre-trained on a large corpus of data from various languages, can leverage its cross-lingual representations to learn task-specific patterns from the data in a given language (Liu et al., 2020). By training these two models separately on the same English dataset, we can obtain a high-performance English monolingual model tailored for emotion detection, as well as a multilingual model that has adapted its cross-lingual representations to the task of emotion detection while still retaining its ability to generalise across languages.

2.2 Knowledge Distillation from the Monolingual Model

To further refine and improve the performance of the multilingual model, we propose a knowledge distillation strategy that utilises the high-performing monolingual English model as a teacher. The refinement scheme for the fine-tuned multilingual model is shown in Figure 1. Since the unlabelled data from development set is in various languages, we first translate all the non-English instances into English using neural machine translation systems based on Marian (Junczys-Dowmunt et al., 2018), i.e. $T : X \rightarrow X'$, where $X = \{x_i, x_2, ..., x_{|X|}\}$, standing for the original development set that contains multilingual texts and $X' = \{x'_i, x'_2, ..., x'_{|X'|}\}$, representing the translated English version. We then obtain predictions from both the monolingual model on the translated data X', and the multilingual model on the original data X.

To transfer the ability of the monolingual model to the multilingual model, we compute the Kullback-Leibler (KL) divergence, a non-symmetric loss function, as the consistency loss between their output distribution P and Q on each instance. Importantly, we focus on minimising the consistency loss only when the quality of translated data in X' is suspected to be of good quality, which ensures that we prioritise knowledge transfer from the teacher model on instances where the translation is sufficiently reliable. To achieve this, we introduce a normalised weight $\bar{w}_i$ for each translated instance x'_i and compute the loss with these weights included, i.e.

$$\begin{aligned}\mathcal{L} &= \bar{w} * \mathrm{KL}(P \| Q) \\ &= \sum_{i=1}^{|X|} \bar{w}_i \left[\sum_{j=1}^{k} p(y_j|x'_i, \theta_{en}) \log \frac{p(y_j|x'_i, \theta_{en})}{p(y_j|x_i, \theta_{xlm})} \right]\end{aligned} \tag{1}$$

where $p(y_j|x'_i, \theta_{en})$ and $p(y_j|x_i, \theta_{xlm})$ represent the output probabilities for the j-th category of the monolingual model on the i-th translated English instance and the multilingual model on the i-th original instance, respectively. k is the number of categories for the emotion detection task.

2.3 Translation Quality Weighting

To compute the weight, which reflects the suspected translation quality, we first obtain the predictions of the same multilingual model on both the original data x_i and translated data x'_i. We then calculate the disagreement between these two prediction using the mean squared error (MSE), a

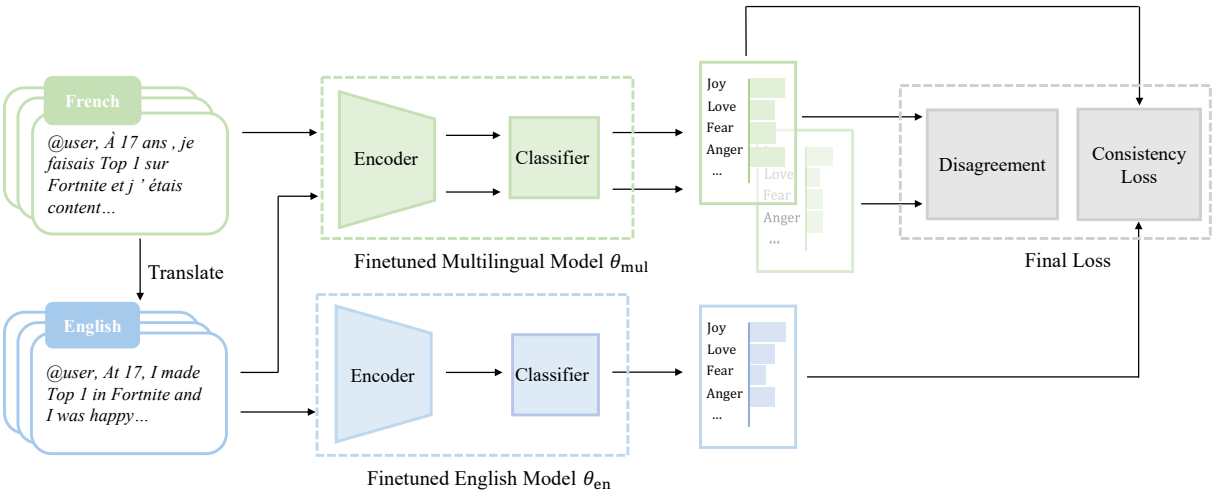

Figure 1: Our proposed refinement scheme for the fine-tuned multilingual model.

symmetric function, i.e.

$$w_i = \sum_{j=1}^{k} \left[p^2\left(y_j|x_i', \theta_{\text{xlm}}\right) - p^2\left(y_j|x_i, \theta_{\text{xlm}}\right) \right] \tag{2}$$

A high MSE value between the predictions of the multilingual on x and x' suggests that the translation quality is likely poor, as the understanding of the model for the original instance and translated instance significantly disagrees, indicating the potential translation flaws or errors, conversely, a low MSE value illustrates the consistency between the original and translated versions, implying a higher translation quality. Therefore, when the disagreement is high, such a translated sample should contribute less to the final loss. To account for this, we normalise the weight as follows:

$$\bar{w}_i = 1 - \frac{w_i - \min(w)}{\max(w) - \min(w)} \tag{3}$$

3 Experimental Results

3.1 Datasets and Shared-task

We utilised the dataset on explainable cross-lingual emotion detection in tweets (EXALT)[1]. This dataset includes annotations for both word triggers and the overall expressed emotion for each instance. The training set is available in English only, while the development set and test set involve five different languages: English, French, Dutch, Russian and Spanish. More details about the shared task can be found in (Maladry et al., 2024).

[1]https://lt3.ugent.be/exalt/

3.2 Baseline

We conducted experiments involving generative models, discriminative models and feature extractors. For the generative models, we employed BLOOM (Le Scao et al., 2023) and BLOOMZ (Muennighoff et al., 2023) with 7.1 billion parameters. Both BLOOM and BLOOMZ are pre-trained on multilingual corpora, and BLOOMZ was further fine-tuned using prompted multitask learning.

For discriminative models, we utilised RoBERTa with the translated samples and XLM-RoBERTa. Additionally, we considered two recent feature exactors, BGE-M3 (Chen et al., 2024), and E5 (Wang et al., 2024), which were pre-trained on the massive multilingual text-pair to extract the cross-lingual features and construct the text embedding for multiple languages. To perform the classification task, we added a fully connected layer as the classifier and applied the average pooling to the text embedding to generate the predictions. The implementation details can be found in the Appendix .1.

3.3 Main Results

We reported the main results of the emotion detection task in Table 1. Based on the overall result, we can see that despite having a larger number of parameters, which generally indicates greater model capacity, generative models such as BLOOM and BLOOMZ did not perform as well as the large discriminative models. This implies that larger models do not necessarily lead to better performance in tasks that require detailed understanding and classification of emotions.

Another noteworthy observation is the signifi-

Models	Development			Testing		
	F1	Precision	Recall	F1	Precision	Recall
Generators						
BLOOM-7b1	49.98	55.58	48.10	49.40	51.74	47.97
BLOOMZ-7b1	47.82	52.44	46.18	49.69	51.65	48.54
Discriminators						
XLM-RoBERTa (base)*	43.29	43.35	44.47	44.76	44.52	46.31
XLM-RoBERTa (large)	50.09	49.78	50.66	53.24	52.77	54.59
RoBERTa (large) w. transl.	52.21	51.57	54.02	-	-	-
Feature Extractors						
E5	49.60	50.09	49.67	54.28	54.05	55.26
BGE-M3	49.67	49.23	50.60	51.51	52.40	51.37
Our Approach						
RoBERTa (large) + XLM-RoBERTa (large)	55.53	**58.80**	54.07	54.54	57.28	53.39
RoBERTa (large) + E5	54.94	56.39	54.42	55.98	56.26	**56.36**
RoBERTa (large) + BGE-M3	53.49	54.63	53.12	51.75	50.73	53.55
Ensemble	**56.01**	56.58	**55.79**	**56.61**	**58.30**	55.73

Table 1: Main results on the development set and testing set for the emotion detection task. * The results are provided by the organiser.

cant performance improvements achieved by the multilingual PLMs after knowledge distillation from the monolingual RoBERTa model. On the development set, all evaluated multilingual models, including XLM-RoBERTa, E5, and BGE-M3, showed significant gains, with an average improvement of 4.87% in terms of the F1 score. The consistent improvements across all multilingual models suggest that the knowledge distillation strategy was effective in transferring the specialised emotion detection capabilities of the monolingual model to the multilingual model. However, the results on the test set were more varied. While XLM-RoBERTa and E5 still demonstrated obvious improvements, the BGE-M3 model only showed a minor increase of 0.24% in performance. This suggests that while the multi-functional pre-training strategy enables BGE-M3 to handle inputs of different granularities, it may have resulted in representations that are less aligned with specific emotion detection tasks and hinder the ability to deal with a large proportion of instances with linguistic phenomena in the test set.

3.4 Explainability of Transformers

To gain insights into the explainability of our models and their ability to identify emotion triggers, we evaluated their zero-shot performance on the binary trigger detection task. Without any explicit training on the fine-tuned emotion detection models, we directly reported the token-level F1 score and mean average precision (MAP) on this task, as presented in Table 2.

Notably, we observed that the trigger detection performances of each model on the development and test sets were basically consistent with their emotion detection performance. Both the refined XLM-RoBERTa and E5 models showed better trigger detection capabilities compared to their original versions. However, the performance of the BGE-M3 model on trigger detection became slightly worse than its original version, which can potentially account for the relatively poor performance gains observed on the emotion detection task on the test set.

Models	Development		Testing	
	F1	MAP	F1	MAP
XLM-R.	33.60	25.11	30.77	24.23
E5	33.03	24.62	30.46	23.78
BGE-M3	33.60	24.68	29.84	23.22
R. + XLM-R. (↑)	33.73	25.46	31.25	24.34
R. + E5 (↑)	33.85	25.02	31.61	24.64
R. + BGE-M3 (↓)	32.97	24.49	29.35	23.03

Table 2: Zero-shot performance on binary trigger detection task as explainable results.

4 Conclusion and Future Work

In this work, we proposed a novel approach that combines the strength of monolingual and multilingual PLM for improved emotion detection performance. Our method involves training a high-performing English monolingual model in parallel with a multilingual model on the same English emotion detection training set. We then employ a knowledge distillation strategy to transfer the

specialised emotion detection capabilities from the monolingual teacher model to refine the multilingual student model. Future work could explore more sophisticated knowledge distillation techniques, as well as employ more accurate and effective translation methods (Na et al., 2024).

Limitations

There are several potential limitations in our work: 1) while a weighting scheme is proposed to account for the translation errors, the quality of the translation system can still significantly impact the knowledge distillation process; 2) the computational complexity involved in training multiple models and performing additional inference steps for weighting and distillation may pose practical limitations.

Acknowledgements

We would like to acknowledge the financial support provided by the Postgraduate Research Scholarship (PGRS) (contract number PGRS-20-06-013) at Xi'an Jiaotong-Liverpool University. Additionally, this research has received partial funding from the Jiangsu Science and Technology Programme (contract number BK20221260) and the Research Development Fund (contract number RDF-22-01-132) at Xi'an Jiaotong-Liverpool University.

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.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Cedric Lothritz, Kevin Allix, Bertrand Lebichot, Lisa Veiber, Tegawendé F Bissyandé, and Jacques Klein. 2021. Comparing multilingual and multiple monolingual models for intent classification and slot filling. In *International Conference on Applications of Natural Language to Information Systems*, pages 367–375. Springer.

Aaron Maladry, Pranaydeep Singh, and Els Lefever. 2024. Findings of the wassa 2024 exalt shared task on explainability for cross-lingual emotion in tweets. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey

Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

Hongbin Na, Zimu Wang, Mieradilijiang Maimaiti, Tong Chen, Wei Wang, Tao Shen, and Ling Chen. 2024. Rethinking human-like translation strategy: Integrating drift-diffusion model with large language models for machine translation. *arXiv preprint arXiv:2402.10699*.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38.

Tiberiu Sosea and Cornelia Caragea. 2020. Cancer-emo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Yuqi Wang, Qi Chen, and Wei Wang. 2021. Multi-task bert for aspect-based sentiment analysis. In *2021 IEEE international conference on smart computing (SMARTCOMP)*, pages 383–385. IEEE.

Zimu Wang and Hong-Seng Gan. 2023. Multi-level adversarial training for stock sentiment prediction. In *2023 IEEE 3rd International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pages 127–134. IEEE.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *ACL 2020*, page 120.

Appendix

.1 Experimental Details

We downloaded all pre-trained models from the Hugging Face repository[2]. The translation system was implemented using machine translation models from the Helsinki-NLP group[3] and a standalone language identification tool (LANGID)[4]. Hardware acceleration was achieved using 2 NVIDIA 3090 GPUs.

For fine-tuning generative models with the provided training data, we employed a parameter-efficient approach using 4-bit quantized low-rank adaptation (QLoRA) (Dettmers et al., 2024). The learning rate was set to 5×10^{-5}, and we used the Alpaca (Taori et al., 2023) template. We showed the prompt for the emotion detection task in Table 3. The batch size was set to 2 per GPU, and the gradient accumulation steps were set to 2. For fine-tuning discriminative models and feature extractors, we utilised the Adam optimiser with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate was set to 5×10^{-6}, and the batch size was set to 16 per GPU.

Below is an instruction that describes a task. Write a response that appropriately completes the request. *### Instruction:* *Please use one word to describe the sentiment expressed in the given tweet.* *Tweet:* `[Tweet]` *### Response:*

Table 3: Prompt used for generative model.

For all models, the epoch count was set to 20. Early stopping was implemented to mitigate the risk of overfitting. In order to achieve this, we further split the provided English data into a training set and a "validation set" with a ratio of 90:10. The best checkpoint on the "validation set" was saved.

In the refinement process, the learning rates for XLM-RoBERTa, E5, and BGE-M3 were set to 1×10^{-7}, 5×10^{-7}, and 7×10^{-7}, respectively, which were chosen from the "validation set" split from the original training set, so no labelled data in languages other than English was used.

For the binary trigger detection, we computed the cosine similarity between the last hidden state of the $< s >$ token and each token from the transformer. We chose a threshold for each model based on the result of the above-mentioned "validation set".

[2]https://huggingface.co/

[3]https://github.com/Helsinki-NLP/Opus-MT

[4]https://pypi.org/project/langid/