XLM-Muriel at SemEval-2025 Task 11: Hard Parameter Sharing for Multi-lingual Multi-label Emotion Detection

Pouya Hosseinzadeh PhD Student
Mohammad Mehdi Ebadzadeh Full Professor
Hossein Zeinali Assistant Professor
Computer Engineering Department
Amirkabir University of Technology
Tehran, Iran
{pouya.hosseinzadeh}, {ebadzadeh}, {hzeinali}@aut.ac.ir

Abstract

SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection Track A addresses multilingual, multi-label emotion classification across 28 languages, including many low-resource varieties. We propose a hard-parameter-sharing architecture built on XLM-RoBERTa-Large, with lightweight, language-specific classification heads, and a two-stage training regimen that first freezes the shared encoder and then fine-tunes it. On the BRIGHTER dataset, our model achieves macro F1-scores up to 0.84 on high-resource languages and maintains robust performance (0.63 macro F1) even on severely imbalanced low-resource languages. We will analyze our results and discuss limitations and potential strength point of our solution that could be leveraged in future work to improve results on similar tasks.

1 Introduction

Emotions are at once an everyday experience and an elusive phenomenon. Although we routinely express and regulate our feelings, they remain intricate and subtle, often defying clear articulation. Language itself is employed in remarkably nuanced ways to convey these internal states, as noted by previous research (Conneau et al., 2020; Deng and Ren, 2020; Zhang et al., 2020a). Moreover, individuals differ greatly in both how they perceive and display their emotions—even among those sharing similar cultural or social backgrounds—making it impossible to determine someone's true emotional state with complete certainty based solely on their words. Emotion recognition, therefore, is not a singular task but rather a collection of related challenges. It includes identifying the speaker's emotional state, discerning the sentiment a piece of text conveys, and even gauging the emotional response it triggers in a reader. Our focus here is on perceived emotion: inferring the emotion that the majority would attribute to the speaker based on a brief sentence or text snippet. This task deliberately excludes determining the reader's emotional reaction, the emotion of another person mentioned, or the speaker's actual feeling—since the latter remains indeterminate from limited text. This distinction is critical, as factors like cultural context, individual differences, and the inherent limitations of textual communication often cause perceived emotions to differ from the speaker's real emotional state (Samy et al., 2018; Zhang et al., 2020b; Ameer et al., 2020). In this paper, we describe our system developed for Track A of the task, which is designed to determine multi-label classifications for a single text snippet across all 28 languages (Muhammad et al., 2025b). Our system leverages a shared multilingual encoder coupled with language-specific classification heads, enabling it to capture both universal and language-dependent emotional cues (Caruana, 1997; Ghosh et al., 2022; Lin et al., 2022). By utilizing advanced transformer models (such as XLM-RoBERTa) as the backbone, our approach not only processes input text efficiently but also generates predictions for multiple emotion classes—including anger, disgust, fear, joy, sadness, and surprise—for each language simultaneously. This robust design addresses the inherent challenges of multilingual emotion recognition, offering a comprehensive solution that can be adapted to varied linguistic contexts.

2 Related works

Research on multi-label emotion detection has accelerated in recent years. (Deng and Ren, 2020) use emotion-specific feature extractors and label correlation graphs; (Ghosh et al., 2022) propose a multitask framework for depression, sentiment, and emotion; (Lin et al., 2022) leverage adversarial multi-task learning for label dependencies. More recent soft-sharing architectures and

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 314–318
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

mixture-of-experts models (Liu et al., 2024; Fan et al., 2025) dynamically allocate capacity per language, yielding gains on typologically distant languages at the cost of increased compute. However, these methods can overfit low-resource languages when class distributions are highly skewed and have not been evaluated in a truly massively multilingual, multi-label setting.

3 BRIGHTER dataset

The BRIGHTER dataset(Muhammad et al., 2025a) is a comprehensive collection of multilabeled emotion-annotated texts spanning 28 different languages. Recognizing that most emotion recognition research has focused on high-resource languages, the BRIGHTER dataset addresses this gap by incorporating predominantly low-resource languages from Africa, Asia, Eastern Europe, and Latin America. Each text instance is carefully annotated by fluent speakers from various domains, capturing the subtle and complex ways in which people express emotions. The dataset not only facilitates monolingual and cross-lingual multi-label emotion identification but also supports intensity-level emotion recognition. The data collection and annotation processes for BRIGHTER were designed to overcome the inherent challenges of building high-quality emotion datasets in diverse linguistic settings. By employing rigorous annotation guidelines and leveraging domain expertise, the creators of BRIGHTER have provided a valuable resource that highlights the variability in emotional expression across different cultures and text domains. Experimental results presented in the associated work demonstrate significant performance differences when using or not using large language models, emphasizing the importance of this resource in bridging the gap in text-based emotion recognition research. Ultimately, the BRIGHTER dataset stands as an essential step toward more inclusive and effective emotion recognition solutions in natural language processing.

4 System overview

To tackle the classification nature of this task, we opted for a BERT-family model—specifically, XLM-RoBERTa-Large—as our backbone. We selected XLM-RoBERTa-Large because it outperforms its base variant by 17 percentage points in macro F1 on the BRIGHTER dev set (Table 3), demonstrating superior cross-lingual transfer and richer encoder inductive bias for emotion cues. Our approach supports two potential strategies: training separate models for each language (storing their trained weights for the test phase) or, inspired by recent work on natural language inference, training specialized expert heads on top of a single (Figure 1), shared encoder. Given that emotional expressions exhibit substantial similarity across languages, our system is designed to share semantic representations among all languages through a common encoder while incorporating language-specific classification heads. These expert heads, whose architecture (number of layers and dimensions) is controlled via Python dictionaries and implemented using PyTorch's ModuleDict, adapt to the unique emotion class distributions observed in each language.

5 Experimental setup

Our experimental framework is designed to maximize data utilization under existing hardware constraints. We use a batch size of 1024 and restrict the maximum number of input tokens to 128, ensuring efficient attention mask construction while avoiding information loss from overly lengthy inputs. Initially, the pre-trained model and tokenizer are loaded, and the encoder's weights are frozen so that only the output logits of the language-specific heads are trained using binary cross-entropy (BCEWithLogitsLoss) to handle the multi-label nature of the task. After roughly 10 epochs with a very low learning rate (on the order of 10^{-7}), we fine-tune the encoder along with the specialized heads using a higher learning rate (approximately 10^{-3}) over 3 additional epochs. In this two-step training strategy, gradients are computed solely for one head per epoch while keeping the other heads' parameters unchanged.

6 Results and analysis

Our experimental evaluation employed the model trained with the approach depicted in Figure 1 right, with the results for four models summarized in Table 1. Additionally, a comparative analysis between the XLM-Base and XLM-Large configurations—following the approach in Figure 1 left—is presented in Table 2. The macro F1-scores reported by the competition's evaluation system for all languages indicate that performance improves with larger model sizes, as evidenced by the results in Table 3. However, Table 2 reveals that

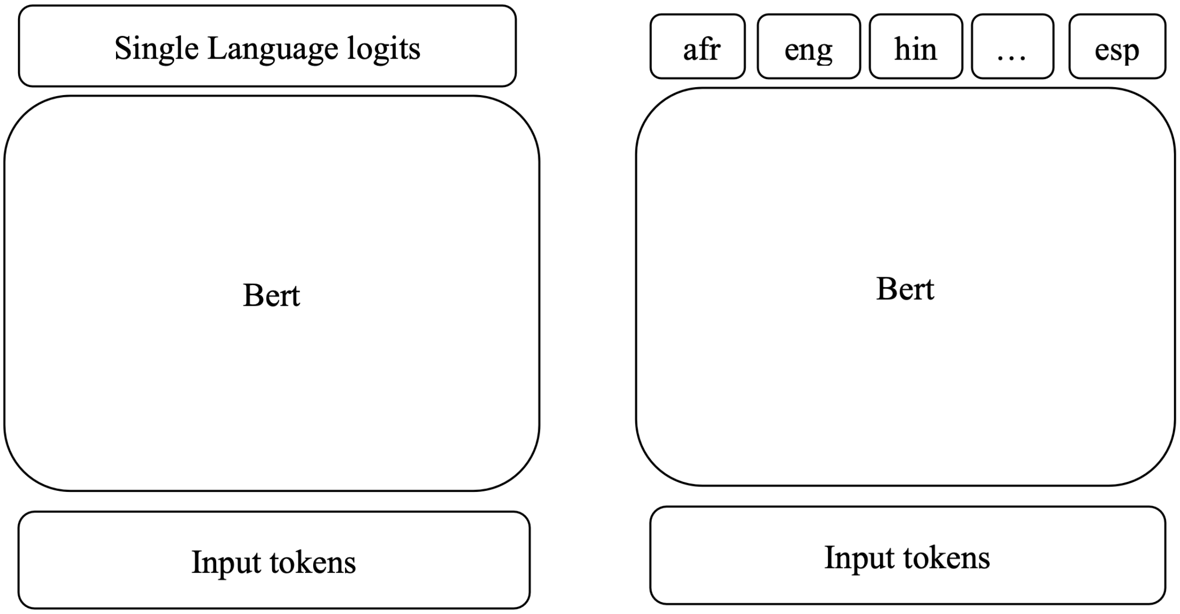

Figure 1: shared encoder with single head (left) for all languages, and seperate head (right) for each language

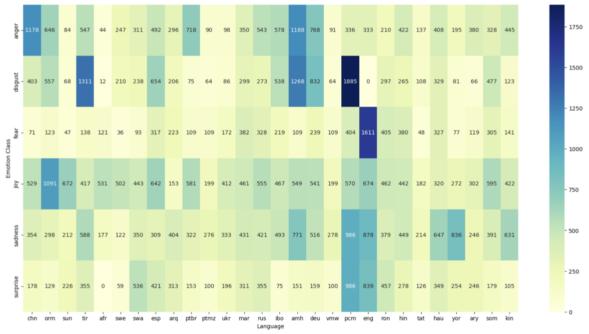

Figure 2: data per class for each language

the training strategy plays an even more crucial role: while training all parameters exclusively on one language may boost accuracy for that specific language, it can concurrently reduce overall model performance across languages. This suggests that a more balanced approach is necessary, either by training certain languages with separate parameters or by adopting an alternative architecture for their corresponding expert heads to avoid adverse effects on languages with greater semantic similarity in emotion expression.

The results of our system are shown in table 1. The system has the best performance on Mar and rus and hin. Regarding the figure 2, in which each class number is depicted for each language, these three languages have the most balanced distribution of data among all classes. As it is obvious from figure 2. Our model even performs rational in cases where the data is unbalanced. For example, without any precaution in one extreme case such eng language the model has the f1-score macro 0.627. The fact that the data distribution for each class in ptmz and vmw are almost the same but the model performance on vmw is almost a third of its performance on ptmz tell us some interesting stories regarding the semantic representation of these languages, because the sole parameter that make difference here is inductive biases which is provided by the encoder itself. This fact gives an intuition that using richer embedding for specific language input might help to increase the performance.

7 Conclusion

Our analysis of the training data further demonstrates that the number of instances for a given emotion (i.e., samples labeled with 1) critically influences model performance. Insufficient representation for a specific emotion impedes the model's ability to extract its distinctive features, leading to poor detection—as observed in languages like Emakhuwa and Yoruba, where challenging classes such as fear and disgust are underrepresented. Despite these challenges, our model achieved a highly competitive standing on the development leaderboard. This outcome underscores both the strengths of our approach and the potential for further optimization. Future work should consider strategies such as increasing model size, fine-tuning hyperparameters more effectively, and employing advanced

Afr	Amh	Arq	Ary	Chn	Deu	Eng	Esp	Hau	Hin	Ibo
0.427	0.664	0.461	0.508	0.592	0.583	0.627	0.745	0.587	0.84	0.478
Kin	**Mar**	**Orm**	**Pcm**	**Ptbr**	**Ptmz**	**Ron**	**Ros**	**Som**	**Sun**	**Swa**
0.321	0.842	0.508	0.531	0.5156	0.432	0.658	0.842	0.445	0.44	0.27
Swe	**Tat**	**Tir**	**Ukr**	**Vmw**	**Yor**					
0.58	0.675	0.482	0.618	0.161	0.301					

Table 1: F1-score results for each language

base single head	base multi head	large single head	large multi head
0.35	0.36	0.49	0.53

Table 2: Average F1-score in macro mode for different settings of XLM model in multi single head

techniques like targeted fine-tuning on specific subsets of data. Detailed error analysis in comparison with top-performing models will be instrumental in identifying and addressing the current limitations, ultimately driving our model closer to the top of the leaderboard.

Limitations

One limitation of our approach is the inherent challenge in balancing the shared multilingual encoder with language-specific expert heads. While the shared encoder leverages common semantic features across languages, it might not fully capture the unique linguistic nuances present in each individual language. This can lead to situations where the expert heads for some languages either overfit to the available training data or fail to compensate adequately for the encoder's generic representations. Moreover, the differing amounts of training data across languages can further exacerbate these issues, resulting in inconsistent performance and potentially underrepresenting certain emotional classes in low-resource languages.

Another challenge lies in the training strategy itself. Our two-stage optimization process, which initially focuses on training only the expert heads before fine-tuning the entire model, requires careful tuning of learning rates and may not generalize well to all languages uniformly. In addition, the reliance on pre-trained models such as XLM-RoBERTa-Large introduces biases towards high-resource languages, which might hinder the model's ability to generalize in truly low-resource scenarios. These factors, combined with the complexities of multi-label classification in a diverse multilingual context, suggest that while our approach is promising, there remains significant room for improvement through further architectural innovations and more balanced data collection.

References

Iqra Ameer, Noman Ashraf, Grigori Sidorov, and Helena Gómez Adorno. 2020. Multi-label emotion classification using content-based features in twitter. *Computación y Sistemas*, 24(3):1159–1164.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Jiawen Deng and Fuji Ren. 2020. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Transactions on Affective Computing*, 14(1):475–486.

Haofang Fan, Xiran Hu, and Geng Zhao. 2025. Cross-lingual social misinformation detector based on hierarchical mixture-of-experts adapter. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7253–7265.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation*, 14(1):110–129.

Nankai Lin, Sihui Fu, Xiaotian Lin, and Lianxi Wang. 2022. Multi-label emotion classification based on adversarial multi-task learning. *Information Processing & Management*, 59(6):103097.

Dahuang Liu, Zhenguo Yang, and Zhiwei Guo. 2024. Progressive fusion network with mixture of experts for multimodal sentiment analysis. In *2024 16th International Conference on Advanced Computational Intelligence (ICACI)*, pages 150–157. IEEE.

XLM-Roberta-base	XLM-Roberta-large	infoXLM-large	LaBSE
0.36	0.53	0.48	0.42

Table 3: Average F1-score in macro mode for 4 tested models

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Ahmed E Samy, Samhaa R El-Beltagy, and Ehab Hassanien. 2018. A context integrated model for multi-label emotion detection. *Procedia computer science*, 142:61–71.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020a. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3584–3593.

Xiao Zhang, Wenzhong Li, Haochao Ying, Feng Li, Siyi Tang, and Sanglu Lu. 2020b. Emotion detection in online social networks: a multilabel learning approach. *IEEE Internet of Things Journal*, 7(9):8133–8143.