EMO-NLP at SemEval-2025 Task 11: Multi-label Emotion Detection in Multiple Languages Based on XLMCNN

Jing Li, **Yucheng Xian** and **Xutao Yang***
School of Information Science and Engineering,
Yunnan University, Kunming 650091, China
lijing_fyjo@stu.ynu.edu.cn,
Corresponding author: yangxutao@ynu.edu.cn

Abstract

This paper describes the system implemented by the EMO-NLP team for track A of task 11 in SemEval-2025: Bridging the Gap in Text-Based Emotion Detection. The task focuses on multiple datasets covering 28 languages for multi-label emotion detection. Most of these languages are low-resource languages. To achieve this goal, we propose a multilingual multi-label emotion detection system called XLMCNN, which can perform multi-label emotion detection across multiple languages. To enable emotion detection in various languages, we utilize the pre-trained model XLM-RoberTa-large to obtain embeddings for the text in different languages. Subsequently, we apply a two-dimensional convolutional operation to the embeddings to extract text features, thereby enhancing the accuracy of multi-label emotion detection. Additionally, we assign weights to different emotion labels to mitigate the impact of uneven label distribution. In this task, we focus on nine languages, among which the Amharic language achieves the best performance with our system, ranking 21st out of 45 teams.

1 Introduction

SemEval-2025 task 11[1] consists of three subtasks, each focusing on different aspects. We focus only on track A. Track A aims to conduct multi-label emotion detection in 28 languages, including Amharic, Hausa, German, English, Oromo, Afrikaans, Algerian Arabic, Chinese, Emakhuwa, Hindi, Javanese, Kinyarwanda, Marathi, Moroccan Arabic, Nigerian-Pidgin, Portuguese(Brazilian), Portuguese(Mozambican), Romanian, Russian, Somali, Spanish(Latin American), Sundanese, Swahili, Swedish, Tatar, Tigrinya, Ukrainian, Yoruba, identifying the emotion labels contained in a given sentence of a specific language(Muhammad et al., 2025b). Every sentence in datasets may contain zero, one, or multiple emotions.

Emotion detection is one of the important research directions in the field of natural language processing. Many emotion detection applications exist, such as recommendation systems (Hu et al., 2021) and public opinion monitoring (Boon-Itt and Skunkan, 2020). However, the majority of research on emotion detection has focused on high-resource languages, with relatively little attention given to low-resource languages. The BRIGHTER dataset as well as datasets of Amharic, Oromo, Somali, and Tigrinya languages (Muhammad et al., 2025a; Belay et al., 2025) collected for SemEval-2025 Task 11 includes low-resource languages from Africa, Asia, Latin America, and other regions, providing data support for multi-label emotion detection in low-resource languages.

We focus on the multi-label emotion detection of nine languages in Track A, including Amharic, German, English, Oromo, Russian, Portuguese (Brazilian), Sundanese, Somali, and Tigrinya. For these nine languages, we propose the XLMCNN, a multilingual multi-label emotion detection system, which can directly process preprocessed sentence texts from different languages and detect their sentiment categories. We employ the pre-trained model XLM-RoberTa-large to obtain the vector representations of sentences. Subsequently, we utilize a Convolutional Neural Network (CNN) for feature extraction to enhance the accuracy of sentiment detection. Moreover, when calculating the loss, we assign different weights to different emotion labels to mitigate the adverse effects caused by the imbalance of emotion label count.

2 Related Work

Emotion analysis is an important area in natural language processing, and the methods used in its research have evolved from lexicon-based

[1] https://github.com/emotion-analysis-project/SemEval2025-task11

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 801–806

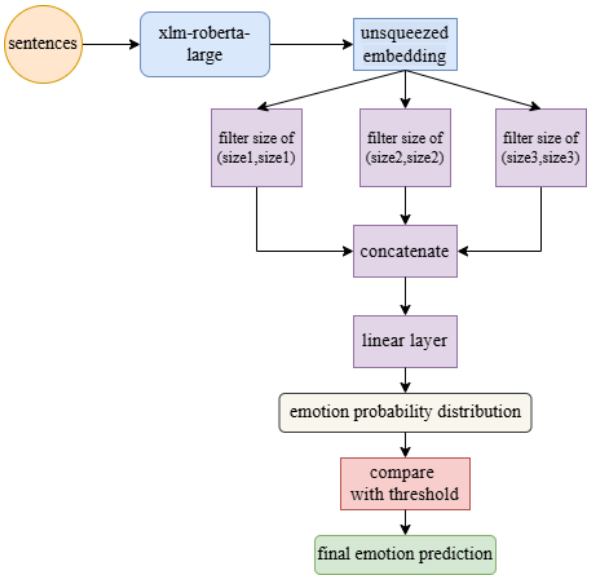

Figure 1: The overall architecture of XLMCNN system

approaches to mainstream approaches, machine learning, and deep learning methods (Medhat et al., 2014). The granularity levels of sentiment analysis research include aspects (Pontiki et al., 2014), documents (Wei et al., 2020), multimodal (Hu et al., 2022), and so on. Research on sentiment polarity has achieved significant success in high-resource languages. However, for multi-label emotion detection, especially in low-resource languages, methods suitable for single-label emotion detection are not applicable due to the co-occurrence of emotion labels (Ahanin et al., 2023). So, more and more researchers have begun to explore new methods for multi-label text emotion detection. In 2014, Jabreel and Moreno (Jabreel and Moreno, 2019) proposed a method combining attention models with Bidirectional Gated Recurrent Units (Bi-GRU) to identify the associations between emotion label and the words in a sentence, thereby achieving multi-label sentiment classification. In 2021, Alhuzali and Ananiadou (Alhuzali and Ananiadou, 2021) used a BERT encoder to make emotion labels and the entire sentence as input to capture the associations between emotion label and all words in the sentence. In 2023, Ameer et al.(Ameer et al., 2023) implemented multi-label emotion detection using RoBERTa and multi-layer attention mechanisms. In 2023, Zahra Ahanin et al.(Ahanin et al., 2023) used a combination of deep learning-based features and human-engineered features to improve the accuracy of multi-label text classification.

3 System Overview

In this section, we provide an overview of the system implementation process and offer an introduction to the details of the system.

3.1 System Architecture

Figure 1 illustrates the overall architecture of our XLMCNN system. The process includes obtaining sentence vectors, performing convolution operations, and predicting the results.

Since the XLM-RoBERTa-large model can process multiple languages and meet the requirements of our experiments, we use this model to obtain vector representations of sentences. We preprocess the input sentences into a form that the model can accept and then use the model's output 'last hidden state ' as the vector representation of the sentences. Since convolutional neural networks can extract localized information, we use it to extract localized emotional information in text. To adapt to the input of a two-dimensional convolutional neural network, we expand the obtained vectors by adding an additional dimension. Then, we perform convolution operations on the expanded vectors using a network with three different convolution kernel sizes. The results of the convolution operations are concatenated along the last dimension. The concatenated results are passed through a fully connected layer as the final classifier. Since this task is for multi-label sentiment classification, where labels are not mutually exclusive but can co-occur, we use the sigmoid function to obtain the final sentiment probability distribution:

$$\hat{y} = \sigma(W_c h + b_c) \tag{1}$$

In the equation, $W_c \in \mathbb{R}^{d_h \times |Y|}$, and $b_c \in \mathbb{R}^{|Y|}$, $|Y|$ represents the number of classes of emotion labels.

3.2 Loss Function

For this multi-label emotion detection task, we use BCEWithLogitsLoss as the loss function. However, after analyzing the number of instances for each emotion label in the dataset, as shown in Figure 2, we found an imbalance among the label counts, which could negatively impact the classification results. Therefore, we employ a weighted BCEWithLogitsLoss to mitigate the impact of the imbalanced label data:

$$Loss = BCEWithLogitsloss(weight = [cw_1, cw_2, \ldots]) \tag{2}$$

language	training	validation	test
amh	2839	710	1774
deu	2082	521	2604
eng	2214	554	2767
orm	2753	689	1721
ptbr	1780	446	2226
rus	2143	536	1000
som	2713	679	1696
sun	739	185	926
tir	2944	737	1840
total	20207	5057	-

Table 1: The sentence numbers in training set, validation set, test set for nine languages.

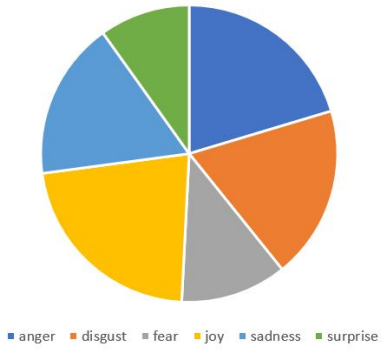

Figure 2: The emotion labels distribution in training set

For the calculation of weights, since the model may overlook emotion labels with fewer instances during training, while the model tends to focus more on labels with a higher number of instances, we use the reciprocal of the proportion of each emotion label in the total number of labels as the weight for that emotion label:

$$cw_i = total/label_i \quad (3)$$

In this equation, cw_i represents the weight of the i-th label, and $total$ represents the total number of emotion label, and $label_i$ represents the number of the i-th emotion label.

4 Experiment

In this section, we introduce the dataset of SemEval-2025 task 11 and how we preprocess these data. Additionally, we also show our experiment configurations.

4.1 Dataset and Preprocess

Our experiments focus on nine languages. They are Amharic, German, English, Oromo, Russian, Portuguese (Brazilian), Sundanese, Somali and Tigrinya. For the dataset of these nine languages, each sentence in the English dataset has five emotion labels: anger, fear, joy, sadness, and surprise. For the datasets of the other eight languages, each sentence has six emotion labels, including an additional label of **disgust** compared to the English dataset. For each emotion label, if a sentence contains a particular emotion, it is marked as 1; if it does not contain the emotion, it is marked as 0.

We add the **disgust** emotion label to the English dataset and label it entirely as 0, in order to unify the label categories of the English dataset with those of the other datasets. We divide the training data of each language into training and validation sets at a ratio of 8:2. Then, we combine the training sets of the nine languages and the validation sets of the nine languages separately, resulting in the training dataset and validation dataset used in our experiments. The data distribution of the training set, validation set, and test set are shown in Table 1. Moreover, the datasets are collected from diverse sources, including social media, news, and speeches (Muhammad et al., 2025a), which leads to the inclusion of some noise in the dataset. Specifically, certain sentences contained emojis, punctuation marks, parentheses, extra whitespace, special characters like # , and other similar elements. These types of noise can interfere with the process of emotion detection and classification. Therefore, during the experiment, we remove these types of noise from the dataset.

4.2 Implementation Details

The entire system is implemented on the Kaggle platform, utilizing GPU T4 $\times$ 2. During the experiments, we employ the macro f1 score as the evaluation metric for model performance. By com-

Hyperparameter	value
num class	6
epoch	7
batch size	128
pad size	32
learning rate	$1e^{-5}$
weight decay	$1e^{-2}$
dropout	0.5
threshold	0.5
hidden size	1024
filter size	$3, 4, 5$
number of filter	64

Table 2: The hyperparameter setting during the experiment

paring the average results on the validation sets of the nine languages, we identify the optimal hyperparameters, which are then saved and applied to the test set to assess the model's performance. The XLMCNN employs the 'FacebookAI/xlm-roberta-large'[2] model to generate sentence vectors. Subsequently, it utilizes the two-dimensional Convolutional Neural Network (CNN) for feature extraction. Finally, a linear layer is applied to obtain the probability distribution of emotion classes. The final emotion labels of a sentence are determined by comparing the probability of the emotion class with a predefined threshold. Table 2 shows the specific parameter settings for these three processes. Finally, we use Adam Optimizer (Kingma and Ba, 2015) to optimize the network parameters, and equation 3 implements class weights' setting in the loss function.

5 Result and Analysis

In this section, we present the test set results that are ultimately submitted to the system, comparing the results between the cases of assigning weights to the labels and not assigning weights. Table 3 shows the results for each language on individual emotion labels and the macro f1 score when weights are assigned to the labels and when no weights are assigned to the labels.

Our proposed system, XLMCNN, performs better, and the computed macro f1 scores are higher on Amharic, German, Oromo, Portuguese, Russian, Somali, and Tigrinya than the system that does not assign weights to the labels. Meanwhile, the XLMCNN system also achieves a slightly higher average macro f1 score across these nine languages compared to the system without label weighting. In addition, the emotion labels **fear** and **surprise** have relatively fewer instances than the other labels. It can be observed that using the XLMCNN system, the f1 score for predicting **fear** is slightly higher for half of the nine languages. Similarly, the XLMCNN system achieves a slightly lower f1 score for predicting **surprise** only in English and Sundanese, while it performs better in predicting **surprise** for the other languages. To some extent, this demonstrates that the weighting method in the system is effective in dealing with the negative impact caused by the imbalance of the number of emotion labels. Moreover, both the XLMCNN and the no-label weighting method perform poorly in the Oromo, Sundanese, Somali, and Tigrinya languages, which indicates that our system still needs some improvements for emotion detection of low-resource languages. Finally, regardless of the system used, the prediction of the **fear** emotion in Sundanese, Oromo and Tigrinya performed very poorly. Among the nine languages, our submitted system exceeds the baseline in two languages.

6 Conclusion

In this paper, we employ a method that combines the xlm-roberta-large model with a convolutional neural network while weighting emotion labels to achieve multi-label text emotion detection in multiple languages. The final experimental results indicate that XLMCNN has better overall performance in handling multi-label emotion detection than the method without label weighting. The method we use generally performs across languages, especially Sundanese, Oromo, and Tigrinya, with a very poor prediction for **fear** emotion. In future work, we will explore the issues within the proposed method and improve the performance of XLMCNN in low-resource languages.

Acknowledgments

All the work in this paper is done in the SemEval-2025 competition. And this work is supported by Scientific Research and Innovation Project of Postgraduates Students in the Academic Degree of Yunnan University (KC-242410495).

[2] https://huggingface.co/

Amharic							
approach	**anger**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**macro f1**
no label weighting	0.6599	0.6944	0.3038	0.6788	0.66	0.529	0.5877
XLMCNN	0.6526	0.7035	0.4124	0.6721	0.6149	0.5556	0.6018
SemEval-basline	-	-	-	-	-	-	**0.6383**
German							
approach	**anger**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**macro f1**
no label weighting	0.7459	0.3955	0.3609	0.6517	0.5866	0.3499	0.5151
XLMCNN	0.7466	0.4815	0.3521	0.6391	0.5868	0.3792	0.5309
SemEval-basline	-	-	-	-	-	-	**0.6423**
English							
approach	**anger**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**macro f1**
no label weighting	0.5141	-	0.8197	0.6883	0.7134	0.6785	0.6828
XLMCNN	0.5078	-	0.8113	0.6709	0.7068	0.6447	0.6683
SemEval-basline	-	-	-	-	-	-	**0.7083**
Oromo							
approach	**anger**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**macro f1**
no label weighting	0.3382	0.2697	0.0303	0.6206	0.2468	0.283	0.2981
XLMCNN	0.3569	0.2744	0.0294	0.6488	0.1237	0.4741	**0.3179**
SemEval-basline	-	-	-	-	-	-	0.1263
Portuguese(Brazil)							
approach	**anger**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**macro f1**
no label weighting	0.6197	0.1443	0.3311	0.6873	0.5703	0.3136	0.4444
XLMCNN	0.6394	0.1284	0.3664	0.683	0.5742	0.3636	**0.4592**
SemEval-basline	-	-	-	-	-	-	0.4257
Russian							
approach	**anger**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**macro f1**
no label weighting	0.7786	0.7019	0.7556	0.8238	0.7133	0.7399	0.7522
XLMCNN	0.7592	0.6948	0.7954	0.8308	0.6973	0.7745	0.7587
SemEval-basline	-	-	-	-	-	-	**0.8377**
Somali							
approach	**anger**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**macro f1**
no label weighting	0.1224	0.2485	0.4895	0.5074	0.4973	0.3121	0.3628
XLMCNN	0.1714	0.273	0.4504	0.5123	0.4858	0.3724	0.3775
SemEval-basline	-	-	-	-	-	-	**0.4593**
Sundanese							
approach	**anger**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**macro f1**
no label weighting	0.1474	0.058	0.0000	0.8225	0.5015	0.3143	0.3073
XLMCNN	0.1263	0.1127	0.0000	0.8268	0.5049	0.1862	0.2928
SemEval-basline	-	-	-	-	-	-	**0.3731**
Tigrinya							
approach	**anger**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**macro f1**
no label weighting	0.075	0.5213	0.0000	0.3923	0.311	0.6211	0.3201
XLMCNN	0.0877	0.5542	0.05	0.363	0.2494	0.6467	0.3252
SemEval-basline	-	-	-	-	-	-	**0.4628**

Table 3: The final result on test set of the nine languages with and without weighting label

References

Z. Ahanin, M. A. Ismail, N. S. S. Singh, and A. AL-Ashmori. 2023. Hybrid feature extraction for multi-label emotion classification in english text messages. *Sustainability*, 15(16):12539.

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1573–1584. Association for Computational Linguistics.

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander F. Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Syst. Appl.*, 213(Part):118534.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Sakun Boon-Itt and Yukolpat Skunkan. 2020. Public perception of the covid-19 pandemic on twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, 6(4):e21978.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7037–7041. IEEE.

Dou Hu, Lingwei Wei, Wei Zhou, Xiaoyong Huai, Zhiqi Fang, and Songlin Hu. 2021. PEN4Rec: Preference Evolution Networks for Session-based Recommendation. In *Proceedings of the 14th International Conference on Knowledge Science, Engineering and Management (KSEM 2021)*, volume 12815 of *Lecture Notes in Computer Science*, pages 504–516, Tokyo, Japan. Springer.

Mohammed Jabreel and Antonio Moreno. 2019. A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, 9(6):1123.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

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, and 29 others. 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, and 2 others. 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.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

Lingwei Wei, Dou Hu, Wei Zhou, Xuehai Tang, Xiaodan Zhang, Xin Wang, Jizhong Han, and Songlin Hu. 2020. Hierarchical interaction networks with rethinking mechanism for document-level sentiment analysis. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part III*, volume 12459 of *Lecture Notes in Computer Science*, pages 633–649. Springer.