DUT_IR at SemEval-2025 Task 11: Enhancing Multi-Label Emotion Classification with an Ensemble of Pre-trained Language Models and Large Language Models

Chao Liu, Junliang Liu, Tengxiao Lv, Huayang Li, Tao Zeng, Ling Luo*, Yuanyuan Sun, Hongfei Lin
School of Computer Science and Technology, Dalian University of Technology, China
{liuchao2464687308, 2958442668, tengxiaolv,
3170511502,ztdx}@mail.dlut.edu.cn
{lingluo, syuan, hflin}@dlut.edu.cn

Abstract

In this work, we tackle the challenge of multi-label emotion classification, where a sentence can simultaneously express multiple emotions. This task is particularly difficult due to the overlapping nature of emotions and the limited context available in short texts. To address these challenges, we propose an ensemble approach that integrates Pre-trained Language Models (BERT-based models) and Large Language Models, each capturing distinct emotional cues within the text. The predictions from these models are aggregated through a voting mechanism, enhancing classification accuracy. Additionally, we incorporate threshold optimization and class weighting techniques to mitigate class imbalance. Our method demonstrates substantial improvements over baseline models. Our approach ranked 3rd out of 90 on the English leaderboard and exhibited strong performance in English in SemEval-2025 Task 11 Track A.

1 Introduction

Emotion classification is crucial in various natural language processing (NLP) applications, including customer feedback analysis, mental health monitoring, and social media sentiment tracking. Unlike traditional sentiment analysis, which categorizes text into positive, negative, or neutral sentiments, multi-label emotion classification is more complex, as a single sentence can express multiple emotions, such as joy, anger, and sadness (Strapparava and Mihalcea, 2008), as shown in Figure 1.This complexity arises from the subjective nature of emotions, their overlapping characteristics, and the ambiguity in short texts.

Although transformer-based models, particularly BERT and its variants, have shown promising results in capturing semantic features and contextual dependencies (Vaswani, 2017), challenges persist, including class imbalance, difficulties in distinguishing subtle emotional expressions, and the need for better generalization across languages (Conneau, 2019).

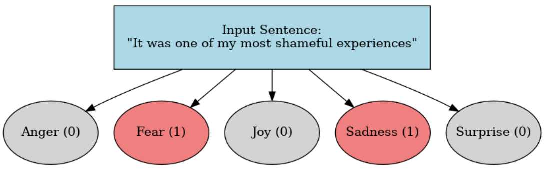

Figure 1: Example of the Multi-Label Emotion Classification task

In this study, we focus on multi-label emotion classification as defined in SemEval-2025 Task 11 Track A (Muhammad et al., 2025a), which aims to evaluate NLP systems' ability to identify multiple emotions in a given text. We propose an ensemble approach, integrating multiple BERT-based pre-trained language models (PLMs) (such as BERT, RoBERTa (Liu et al., 2019), and other variants) along with large language models (LLMs) to capture diverse emotional cues (Brown et al., 2020). The predictions from these models are aggregated using a voting mechanism, which enhances robustness and accuracy. By leveraging both pretrained transformers and LLMs, our approach effectively captures the complex and overlapping nature of emotions, improving the generalization across varied emotional expressions.

In addition to the ensemble strategy, we incorporate threshold optimization and class weighting to address class imbalance and improve decision boundaries. These techniques ensure that underrepresented emotions are adequately considered, leading to significant performance improvements over baseline models and enhancing our system's effectiveness in multi-label emotion classification.

[*] Corresponding Author

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 116–121
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

2 Related Work

The fundamental challenge in multi-label emotion classification lies in detecting non-exclusive emotional states within textual expressions. Early methodologies predominantly employed lexicon-based systems combined with statistical classifiers like SVMs (Mohammad and Turney, 2013), utilizing hand-engineered features such as emotion-word counts and syntactic patterns. While effective for coarse-grained analysis, these approaches exhibited limitations in handling three critical aspects: (1) contextual polysemy in emotional lexicons (e.g., "cold" indicating either temperature or emotional detachment), (2) compositional semantics in multi-emotion expressions, and (3) cross-lingual generalizability.

Currently, pre-trained language models, especially BERT and its variants, have performed well in sentiment multi-label classification tasks. These models effectively capture contextual information through a bidirectional Transformer architecture, improving classification accuracy. Studies have shown that PLMs generally outperform traditional methods and early deep learning models. In multi-label prediction, the binary cross entropy loss function is widely used to deal with the independence of each label (Zhang and Wallace, 2015). At the same time, a weighted loss function is used to adjust the label weights to address the label imbalance problem. In addition, some studies have further improved the classification effect by modeling the dependencies between labels through graph neural networks (GNNs) or conditional random fields (CRFs) (Tenenboim et al., 2009). In general, PLMs perform significantly better than traditional methods in this task and have achieved good results on multiple standard datasets.

In the task of sentiment multi-label classification, large language models have performed well, especially in capturing the complex sentiment in text and the relationship between labels. LLMs usually perform label prediction through generative or sequence-to-sequence (Seq2Seq) methods, and mine the pre-trained knowledge of the model by designing appropriate prompts. In addition, similar to PLMs, LLMs also use weighted loss functions to solve the label imbalance problem and combine multi-task learning to further improve the classification effect (Raffel et al., 2020). Although LLMs have achieved excellent results in sentiment multi-label classification, their huge computational requirements remain a challenge.

3 System Overview

As shown in Figure 2, our proposed system is composed of two main stages. In the first stage, we train and fine-tune three transformer-based models, BERT, RoBERTa, and DeBERTa (He et al., 2020), employing strategies such as automatic threshold search, class weight allocation, and data augmentation to address challenges like data imbalance and overfitting. Additionally, we explore advanced large models, including Qwen2.5 (Yang et al., 2024) and Llama3.1, to further enhance performance. In the second stage, we improve model robustness and accuracy by integrating predictions from multiple models (RoBERTa, DeBERTa, Qwen2.5 and Llama3.1), using a hard voting strategy and cross-validation, ensuring better generalization and complementary feature learning.

3.1 Model Architecture

Pre-trained language models (PLMs): Two Transformer-based models have been fine-tuned as sequence classifiers: RoBERTa and DeBERTa. RoBERTa is a pretrained language model based on the Transformer architecture, introduced by Meta AI. As an enhanced version of BERT, RoBERTa significantly improves performance through strategies such as improved training methods, expanded data, and increased computational resources. DeBERTa, developed by Microsoft Research , introduces two key innovations on top of BERT: the disentangled attention mechanism and the enhanced mask decoder. These improvements make DeBERTa particularly suitable for tasks requiring a precise understanding of contextual relationships, such as sentiment analysis and multi-hop reading comprehension.

Large language models (LLMs): In recent years, large language models have demonstrated impressive capabilities in tackling various NLP tasks. Motivated by these advances, we adopted two state-of-the-art LLMs, Qwen2.5 and Llama3.1, to construct our sequence classifier. We begin by pre-processing our data set using the tokenizers designed for each model. Next, we fine-tune both Qwen2.5 and Llama3.1 on the training subset of our data to adapt them for the specific classification task. Once fine-tuned, the models are applied to the test data to generate predictions. Finally, we evalu-

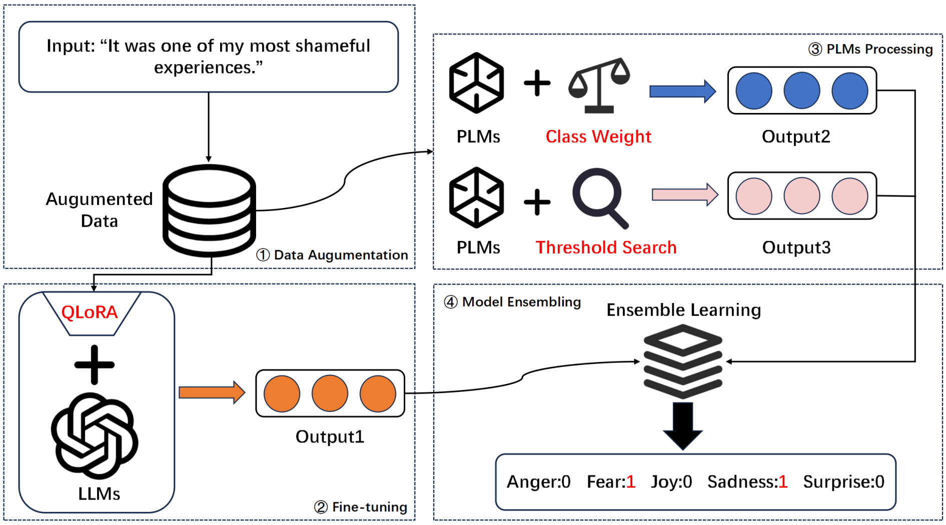

Figure 2: Overview of our system.

ate the performance of these models by comparing the predictions with the true labels.

For PLMs, we use it as an encoder and connect it to a classification layer to get the output, while for LLMs, we directly get the classification results of text sentiment in a generative way.

3.2 Automatic Threshold Search

In multi-label classification tasks, the model usually outputs a probability value for each class (for example, a value between 0 and 1 generated after Sigmoid activation (Kingma and Ba, 2014)). Traditional methods usually use a fixed threshold (such as 0.5) to binarize these probabilities into 0/1 labels, but this approach often does not work well when dealing with imbalanced class distribution or differences in confidence distribution (Zhang and Zhou, 2013). In order to solve the imbalanced distribution of class labels mentioned in Section 3.1, we introduced a strategy of setting independent thresholds for each class to improve the credibility of model predictions. Specifically, we traverse a series of candidate thresholds for each class and independently search for the optimal threshold based on its performance on the validation set (Fan and Lin, 2007). This method not only maintains overall prediction accuracy but also significantly improves the model's ability to capture low-frequency classes and complex label relationships, enhancing its robustness and effectiveness in practical applications.

3.3 Class Weight Allocation

To address overfitting in high-frequency classes and the probability shift in low-frequency classes caused by sample imbalance, we not only apply a separate threshold method but also assign class-specific weights in the loss function to ensure the model pays equal attention to all classes during training. After applying class weight allocation, threshold search is no longer used and the threshold defaults to 0.5.Taking the cross entropy loss function as an example, the loss function after introducing weights can be expressed as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{i,c} \cdot \log(p_{i,c}) \tag{1}$$

Among them, w_c is the weight of class c, $y_{i,c}$ is the true label, and $p_{i,c}$ is the predicted label. The calculation method of each class weight w_c is as follow:

$$w_c = \frac{N_{\text{total}}}{N_c} \tag{2}$$

N_{total} is the total number of samples, N_c is the number of samples of class c

3.4 Data Augmentation

Table 1 shows the distribution of 0 and 1 labels for each class in the training set. From the figure, we can clearly see that there is a significant difference in the distribution of 0 and 1 labels in some sentiment classes, which makes the model prone to overfocus on classes with higher sample sizes when predicting, and insufficient attention to low-frequency

Sentiment	Negative	Positive
Anger	2435	333
Fear	1611	1157
Joy	2094	674
Sadness	1890	878
Surprise	1929	839

Table 1: Label distribution of different emotions in the training set.

classes. To address this problem, we tried to perform data enhancement on low-frequency classes. Taking "Anger" as an example, we extracted all "Anger"-labeled texts from the training set and applied simple data augmentation methods, such as synonym replacement, back-translation, and reconstruction using a large language model based on the original text and labels. It is noteworthy that we applied data augmentation strategies to each model.

3.5 Ensemble Learning

In multi-label classification tasks, a single model may not be able to fully capture complex label relationships and semantic features for the following reasons:

- Model bias: Different model architectures (such as BERT and RoBERTa) have different sensitivities when processing text features. For example, BERT is good at capturing bidirectional context, while DeBERTa performs better in decoupling attention mechanisms. A single model may not be sufficient to fully model certain classes (such as low-frequency labels "Anger") or certain specific language expressions (such as irony, metaphor).
- Variance and risk of overfitting: When the amount of training data is limited or there is a lot of noise, a single model is prone to overfitting the distribution of the training set, resulting in decreased generalization ability.
- Feature complementarity: Different models can extract complementary features (for example, word-level features and syntactic structure features). Therefore, by integrating the results of multiple models, multi-dimensional information can be integrated to improve the robustness of the model.

Therefore, we integrate the results of different models through a hard voting strategy (i.e., directly counting the predicted labels of multiple models and selecting the label with the most votes). When the model's output is uncertain (e.g., two votes in favor and two against), the corresponding data is flagged. These ambiguous cases are then re-evaluated by the models. If uncertainty persists after re-inference, a label of 0 or 1 is assigned to the emotion at random with a probability of 50%. Based on previous research, we selected RoBERTa, DeBERTa, Qwen2.5 and Llama3.1 as base models for integration. At the voting stage, we only use the thresholds that were trained for each individual model and do not perform any additional threshold search.

Hyperparameters	PLMs	LLMs
Epochs	10	10
Dropout	0.1	0.05
Optimizer	AdamW	AdamW
Weight Decay	0.001	0.001
Train Batch Size	16	4
Max Input Length	512	512
Learning Rate	2×10^{-5}	1×10^{-4}
Max Output Length	128	128

Table 2: Hyperparameter settings for PLMs and LLMs training.

4 Experimental Setup and Results

4.1 Dataset

We used the BRIGHTER dataset (Muhammad et al., 2025b) provided by the organizer, which contains 28 different languages, and a text segment in the data may be labeled with multiple emotions (anger, sadness, fear, disgust, happiness, surprise) instead of a single emotion class. We participated in this subtask on the English dataset.

4.2 Hyperparameters

Detailed information on the hyperparameter settings of the experiment is shown in Table 2.

4.3 Metrics

The organizer of this evaluation uses the macro F1 score as the main indicator to evaluate the performance of the model. In multi-label classification problems, the macro F1 score is obtained by calculating the F1 score of each class and averaging the F1 scores of all classes. The characteristic of the macro F1 score is that it ignores the difference in the number of samples in each class and gives each

Settings	Macro F1	Anger	Fear	Joy	Sadness	Surprise
RoBERTa	0.750	0.743	0.818	0.667	0.753	0.769
+ threshold search	0.784	0.788	0.800	0.706	**0.833**	0.794
+ class weight	0.783	0.774	0.790	0.772	0.758	**0.820**
+ data augmentation	0.770	0.774	**0.841**	0.724	0.727	0.781
+ ensemble learning	**0.795**	**0.800**	0.774	**0.787**	0.794	0.818

Table 4: Ablation experiment based on RoBERTa.

class the same weight. First, the recall and precision of each class are calculated separately and then the F1 score of each class is obtained based on the harmonic mean of the precision and recall. Finally, the F1 scores of all classes are averaged to obtain the macro F1 score.

5 Results

Table 3 shows the performance of the different base models in this task. Table 4 shows the experimental results based on the RoBERTa model and the improvement methods mentioned in Section 3. It can be clearly seen from the table that the automatic threshold search and class weight allocation strategy significantly enhance the model's attention to low-frequency classes, thereby effectively improving the overall performance. However, the data enhancement method failed to achieve the expected effect and its improvement was limited to a slight improvement. Based on the above experiments, we further integrated the RoBERTa model with the experimental results of adding three improvement methods separately. The experiment shows that this integration strategy significantly improves prediction accuracy, likely because a single model struggles to fully capture complex label relationships and semantic features in text.

Models	Macro F1	Micro F1
BERT	0.724	0.733
RoBERTa	0.750	0.768
DeBERTa	0.739	0.751
Qwen2.5	0.779	0.788
Llama3.1	0.782	0.787

Table 3: Performance of different models on this task.

Finally, we adopted the full model integration (covering language models and large language models) as the ultimate solution of the system, and submitted the prediction result file of the final model on the test set. The official ranking is shown in Table 5, and the system won the third place in the macro F1 indicator.

Rank	Team	Macro F1
1	PAI	0.823
2	NYCU-NLP	0.822
3	**DUT_IR**	**0.812**
4	TeleAI	0.806
5	Pateam	0.805

Table 5: Results of top 5 teams for Task11 Track A English leaderboard on the test set.

6 Conclusion

This paper introduces the system we designed in Track A of Semeval-2025 Task 11, which aims to solve the problem of unbalanced class distribution that is common in multi-class label classification tasks. By combining methods such as automatic threshold search and class weight assignment, we effectively alleviate the model's excessive focus on high-frequency emotions and reduce its tendency to ignore low-frequency emotions. Based on this, we further adopt a model integration strategy to optimize the shortcomings of a single model in capturing complex label relationships and semantic features in text, and significantly improve the robustness and generalization ability of the model. Overall, our system performs outstandingly in the task of multi-label emotion classification, especially on the English test set of Track A, where it achieved an excellent score of third place, verifying the effectiveness and advantages of our method.

7 Acknowledgments

The authors thank the organizers of the SemEval-2025 Task 11. This research was supported by the grants from the National Natural Science Foundation of China (No. 62302076, 62276043).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Rong-En Fan and Chih-Jen Lin. 2007. A study on threshold selection for multi-label classification. *Department of Computer Science, National Taiwan University*, pages 1–23.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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*.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

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. 2025a. 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.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, and Idris Abdulmumin et al. 2025b. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560.

Lena Tenenboim, Lior Rokach, and Bracha Shapira. 2009. Multi-label classification by analyzing labels dependencies. In *Proceedings of the 1st international workshop on learning from multi-label data, Bled, Slovenia*, pages 117–132.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.