11-711 Advanced NLP Assignment 4 Report

Emotion Classification in Chinese Text Domain

Team 11

Yiteng Mu (yitengm@andrew.cmu.edu)

Sennan Cen (sennanc@andrew.cmu.edu)

Xiaonan Sun (xiaonan2@andrew.cmu.edu)

Table of Contents

1. Introduction

Our research in natural language processing (NLP) focuses on emotion prediction and interpretation, which is crucial for empathetic AI development and identifying harmful online behaviors. Utilizing datasets like GoEmotions with detailed emotion taxonomy, our goal is to understand emotional expressions for Chinese.

In the initial phase of our research, we focused on analyzing and modeling the GoEmotions dataset to construct a baseline emotion detection classification model. Our subsequent efforts were directed towards adapting these text-based emotion detection methodologies for the Chinese language. To this end, we selected the CPED dataset, introduced by Chen et al. in 2022, to refine our baseline model. This phase involved innovative approaches in tokenization and adaptation to the linguistic nuances of Chinese text.

2. Literature Review

2.1. Background of Emotion Detection in NLP

Emotion detection from text has been a popular and challenging domain within NLP. Historically, the field began with sentiment analysis, primarily classifying texts into positive, negative, and neutral categories. However, this approach showed limitations in addressing multiple and detailed emotion categories. Two foundational theoretical models have guided much of the emotion detection research: Ekman's model, which identifies six basic emotions, including happiness, sadness, anger, fear, disgust, and surprise (Ekman, 1992), and Plutchik's wheel of emotions, which suggests a more complex framework with eight primary emotions and various intensities and combinations (Plutchik, 1980).

2.2. Models Utilization in Text Classification

The field of emotion detection has witnessed the adoption of various models, ranging from Supervised Machine Learning techniques like Naïve Bayes and SVM to advanced Deep Learning approaches. For example, early methods utilizing SVM models with text classification processes, including tokenization, stop word removal, stemming, and extracting probabilistic metrics as features (Zainuddin & Selamat, 2014). With the rise of Deep Learning, neural network architectures like Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) gained prominence for their ability to capture sequential information and context in text. More recently, Transformer models, namely Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT), have revolutionized NLP tasks. These models benefit from pre-training on large datasets, enabling them to discover hidden patterns and understand contextual meanings in human language more effectively.

2.3. Datasets with Emotions

The evolution of datasets in emotion-focused research has been significant. Early datasets, such as the ISEAR project, laid the groundwork for emotion detection studies (Scherer & Wallbott, 1994). However, previous emotion-related datasets might look inferior in the size of the dataset and number of classification labels in comparison to contemporary ones. Currently, datasets like GoEmotions (Demszky et al., 2020), EmoNet (Abdul-Mageed & Ungar, 2017), and CrowdFlower (Bostan & Klinger, 2018) offer larger sizes and more diverse emotion labels.

Name	**Descriptions**
GoEmotions	Provided over 58k annotation texts with 27 emotion labels (and neutral)
EmoNet	Provided 24 distinct emotion categories
CrowdFlower	Provided 14 labels

Figure 2.3.1. Contemporary Emotion-Related Datasets with Descriptions

Importantly, the advancement in dataset creation extends beyond English. The XED dataset, inspired by Plutchik's wheel of emotions, offers annotations in more than 30 languages, derived primarily from movie subtitles (Öhman et al., 2020). Additionally, in the Chinese language research domain, Chen et al. (2022) introduced a comprehensive emotional dialogue dataset, featuring over 90,000 annotated sentences extracted from Chinese television series.

Nevertheless, potential biases in dataset sources can impact model performance and accuracy. For instance, the texts collected in GoEmotions were mainly from Reddit, which may potentially introduce biases due to factors such as platforms and annotators (Demszky et al., 2020).

3. Methodology

3.1. Datasets Overview

3.1.1. GoEmotions Dataset Overview

Same dataset as Assignment 3. It is developed by Google, and serves as a benchmark for emotion classification in English, featuring 28 distinct emotion labels. It is commonly employed in NLP projects focusing on sentiment analysis and emotion recognition.

3.1.2. CPED Dataset Overview

A Chinese dataset primarily focused on emotional expressions in Chinese. It contains approximately 90K training samples, 10K validation samples, and 10K test samples. While its primary source is Chinese TV series, it categorizes emotions differently compared to the GoEmotions dataset.

3.2. Dataset Augmentation

3.2.1. Language Translation

Utilizing the Google Translate API, we translated all Chinese utterances into English. This step was crucial for aligning the CPED dataset with the English-centric GoEmotions dataset. The translated dataset served as a basis for fine-tuning the RoBERTa model, developed in our Assignment 3. And the performance will serve as the baseline. In parallel, a RoBERTa model trained on Chinese data was used for processing the original Chinese text. This will be talked about in detail in Section 4.

3.2.2. Label Mapping

As the CPED dataset only contains 13 emotion labels and not all labels are in the GoEmotions label set. We need to do label mapping and transforming to make them consistent.

The project leveraged GPT models for classifying broad categories like "positive-other" and "negative-other" into specific GoEmotions labels. This is considered to be cost and time efficient, due to the time limit in this project. While more rigorous human labeling would be expected in the future.

Manual mapping was also employed for those emotions in CPED that did not directly align with GoEmotions labels, such as mapping "worried" in CPED to ["nervousness", "disapproval"] in GoEmotions.

3.2.3. Text Combination

Considering the nature of the dataset and special features in Chinese, some of the data is highly context-related. The utterances in the CPED dataset are mostly colloquial, therefore they are shorter in length and could not easily detect the emotion from a single fragment of utterances without the contextualized understanding. We chose to combine the utterances from multiple rows into one when they shared identical speakers and emotion labels. Aiming to assist the model to better comprehend the context.

3.2.4. CPED Statistics

On 8.1. Appendix A, we list some charts to illustrate the CPED statistics (e.g. label distribution, sentiment distribution, emotion correlation matrix).

4. Implementation

4.1. Baseline Model Details

4.1.1. Baseline Model Introduction

At the beginning of the phase in our research, we followed the paper on the GoEmotions dataset specified to implement a BERT base model to perform the emotions detection. To enhance the functionality of prediction, we further selected the 'roberta-base' model, tailored specifically for the 'GoEmotions' fine-tuning task. To further transform the emotion detection techniques to the Chinese text documents, we determined to translate the “Utterance” feature in the CPED dataset and fine-tuned the initial RoBERTa model implemented for the GoEmotions dataset.

4.1.2. Baseline Dataset Transformation

The CPED dataset contains 13 unique emotion labels (with neutral) within the “Emotion” feature that needed to be transformed to keep aligned with the emotions in the GoEmotions dataset. The mapping details are stated in the “Datasets” section above. We utilized the Python package ‘googletrans’ with version 4.0.0-rc1 to translate all rows of utterances in the original CPED dataset into English texts.

Due to the fundamental differences between the GoEmotions and CPED datasets in the data collection sources, which led to difficulties in the prediction of the CPED dataset, we combined the translated utterances into one, as stated in Section 3.2.3. Text Combination.

After removing translation-failed rows, combining the utterances, and mapping the emotion labels, we have 46k training data, 6k validation data, and 16k testing data.

4.1.3. Baseline Model Configuration Parameters

(The model configuration is similar to the implementation on GoEmotions):

Hyperparameters Name	**Description**
Learning Rate (learning_rate)	Experimentation in 1e-5 and 2e-5 (optimal: 2e-5)
Epochs (epochs)	A total of 5 training epochs
Weight Decay(weight_decay)	Set to 1e-5
Adam Epsilon (adam_epsilon)	Set to 1e-8
Gradient Accumulation Steps (gradient_accumulation_steps)	Set to 1, implying that gradients are updated after every batch
Warmup Steps (warmup_steps)	Approximately 10% of the total training steps

Figure 4.1.1. Baseline Configurations

In addition, we used the ‘roberta-base’ tokenizer with a batch size equal to 64. The optimizer was set to AdamW. We also implemented a linear scheduler with the warmup phase which the number of warmup steps is stated above.

4.2. Challenges and Solution

The idea of the baseline model was to translate the Chinese text into English and then utilize an English tokenizer and a pretrained model. However, we encountered issues such as some translations failed to accurately convey the original meaning of the sentences.

Chinese Text	**Machine Translation**	**Problem**
我真的是高看你了	I really look at you high	This expression is typically used to convey disappointment or surprise at someone not meeting the

		expectations that the speaker previously held for them.
腰杆都硬了	The waist is hard	This phrase is often used metaphorically to describe someone who has become more confident, self-assured, or assertive. It implies gaining the strength or courage to stand up for oneself. A better translation would be "growing a backbone".

Table 4.2.1. (Sample) Machine Translation Failure

This led us to reconsider our approach. We are now exploring the possibility of directly employing a Chinese tokenizer and a model trained on Chinese text, to ensure the nuances and specifics of the language are better captured and understood in our analysis.

4.3. Advanced Model details

We explored Chinese pre-trained BERT with Whole Word Masking (Y. Cui et al., 2021), and implemented the RoBERTa model pretrained on Wikipedia Chinese text for the emotion classification task.

4.3.1. Model Architecture

Built on a base RoBERTa model, we added custom additions including additional linear and dropout layers. In addition to that, we customize our cross entropy loss function with class weights for handling imbalanced datasets.

Layers	**Description**
Dropout Layer	0.2, 20% of the neurons are dropped out during training.
Dense Layer	768, transforming the hidden size of RoBERTa to 768 units
Activation Function	ELU (Exponential Linear Unit), introducing non-linearity into the model
Output Projection Layer	Maps the 768-dimensional vector to the number of classes in the task (specified by num_labels).

Figure 4.3.1. Model Architecture (Customized Chinese RoBERTa)

4.3.2. Model Configuration Parameters

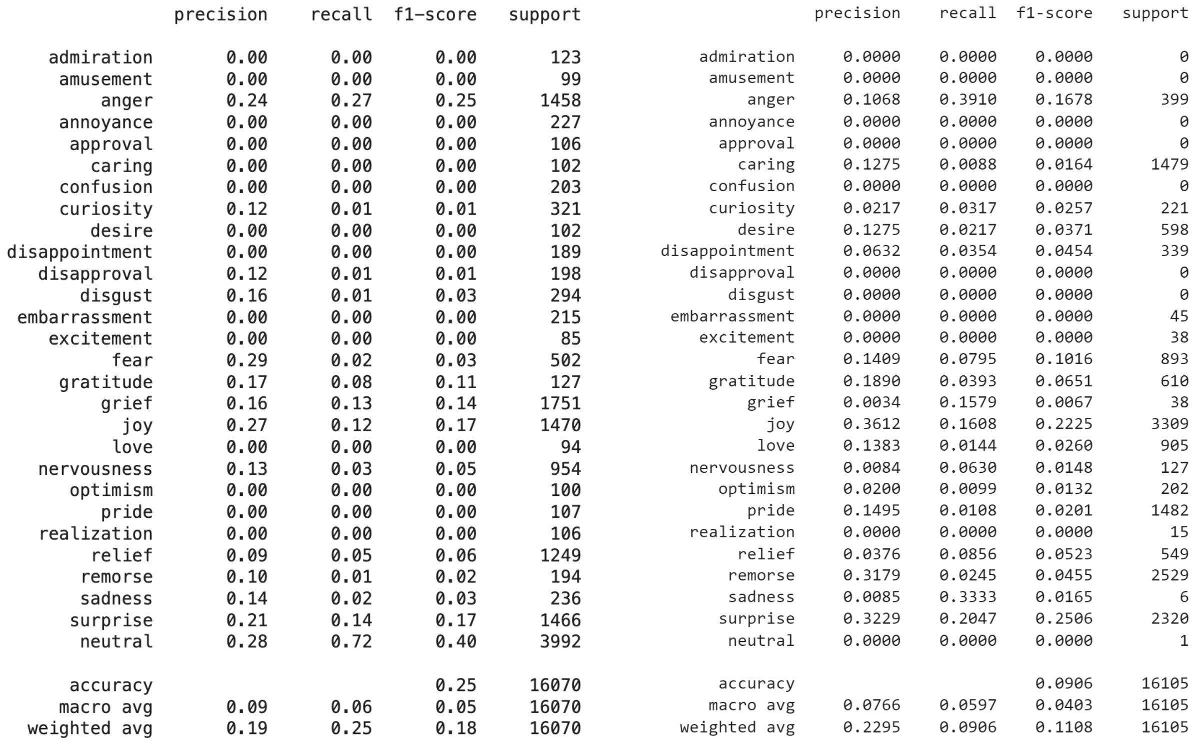
HyperParameters	**Description**
Learning Rate (learning_rate)	2e-5, determining the step size during gradient descent.
Epochs (epochs)	5, defining the number of complete passes through the training dataset.
Dropout Rate (dropout_rate)	0.1, the probability of dropping out neurons during training.

Warmup Steps (warmup_steps)	1000, the number of steps for the learning rate warmup phase.

Figure 4.3.2. Model Configurations (Customized Chinese RoBERTa)

Other hyperparameters can be found in the 8.2. Appendix B.

5. Results



```
                precision    recall  f1-score   support

     admiration       0.00      0.00      0.00       123
      amusement       0.00      0.00      0.00        99
          anger       0.24      0.27      0.25      1458
      annoyance       0.00      0.00      0.00       227
       approval       0.00      0.00      0.00       106
         caring       0.00      0.00      0.00       102
      confusion       0.00      0.00      0.00       203
      curiosity       0.12      0.01      0.01       321
         desire       0.00      0.00      0.00       102
 disappointment       0.00      0.00      0.00       189
    disapproval       0.12      0.01      0.01       198
        disgust       0.16      0.01      0.03       294
  embarrassment       0.00      0.00      0.00       215
     excitement       0.00      0.00      0.00        85
           fear       0.29      0.02      0.03       502
      gratitude       0.17      0.08      0.11       127
          grief       0.16      0.13      0.14      1751
            joy       0.27      0.12      0.17      1470
           love       0.00      0.00      0.00        94
    nervousness       0.13      0.03      0.05       954
       optimism       0.00      0.00      0.00       100
          pride       0.00      0.00      0.00       107
    realization       0.00      0.00      0.00       106
         relief       0.09      0.05      0.06      1249
        remorse       0.10      0.01      0.02       194
        sadness       0.14      0.02      0.03       236
       surprise       0.21      0.14      0.17      1466
        neutral       0.28      0.72      0.40      3992

       accuracy                           0.25     16070
      macro avg       0.09      0.06      0.05     16070
   weighted avg       0.19      0.25      0.18     16070
```

```
                precision    recall  f1-score   support

     admiration     0.0000    0.0000    0.0000         0
      amusement     0.0000    0.0000    0.0000         0
          anger     0.1068    0.3910    0.1678       399
      annoyance     0.0000    0.0000    0.0000         0
       approval     0.0000    0.0000    0.0000         0
         caring     0.1275    0.0088    0.0164      1479
      confusion     0.0000    0.0000    0.0000         0
      curiosity     0.0217    0.0317    0.0257       221
         desire     0.1275    0.0217    0.0371       598
 disappointment     0.0632    0.0354    0.0454       339
    disapproval     0.0000    0.0000    0.0000         0
        disgust     0.0000    0.0000    0.0000         0
  embarrassment     0.0000    0.0000    0.0000        45
     excitement     0.0000    0.0000    0.0000        38
           fear     0.1409    0.0795    0.1016       893
      gratitude     0.1890    0.0393    0.0651       610
          grief     0.0034    0.1579    0.0067        38
            joy     0.3612    0.1608    0.2225      3309
           love     0.1383    0.0144    0.0260       905
    nervousness     0.0084    0.0630    0.0148       127
       optimism     0.0200    0.0099    0.0132       202
          pride     0.1495    0.0108    0.0201      1482
    realization     0.0000    0.0000    0.0000        15
         relief     0.0376    0.0856    0.0523       549
        remorse     0.3179    0.0245    0.0455      2529
        sadness     0.0085    0.3333    0.0165         6
       surprise     0.3229    0.2047    0.2506      2320
        neutral     0.0000    0.0000    0.0000         1

       accuracy                         0.0906     16105
      macro avg     0.0766    0.0597    0.0403     16105
   weighted avg     0.2295    0.0906    0.1108     16105
```

Figure 5.1: Performance Comparison
Baseline Model on the Left, Advanced Model on the Right

6. Discussion

This section discusses and compares the performance of our baseline and advanced model.

The baseline model predicted a full spectrum of emotions. Despite its extensive scope, the model displayed limitations, with several labels yielding zero scores in precision, recall, and F1-score, indicating a complete miss in those classifications. Notably, we manually removed some data due to Google translation deficiency, which reflects the challenges of working with machine translation.

Compared with the baseline model, the advanced model had a higher weighted average precision, indicating its enhanced capacity to accurately predict specific classes. This improvement is particularly pronounced in the case of minority labels, such as the "caring" label, where the advanced model not only successfully identified instances but also did so with greater accuracy. Overall, the advanced model is more likely to predict positive labels from the given text. This bias towards positive classifications might suggest that it is more sensitive to nuances within positive sentiment expressions.

7. Conclusion and Future Work

7.1. Conclusion

In this paper, we focused on the task of predicting emotions within Chinese textual content. Initially, we established a baseline approach that involved translating Chinese text into English, followed by the application of an English tokenizer and a pretrained RoBERTa model. To address the issue of inaccurate machine translation, we later investigated the use of a Chinese tokenizer alongside a pretrained model specifically designed for Chinese. Our findings indicate that the proposed customized Chinese RoBERTa model enhances our ability to interpret text-based emotions, in particular marked by its enhanced performance in accurately predicting minority labels.

7.2. Limitations and Future Improvements

7.2.1. Nature of Data Resource

The Chinese Text we used in the CPED dataset is primarily from Chinese TV Shows, and most of them are literate, emotional and dramatic, leading to a higher proportion of negative sentiments in the datasets. Apart from that, most of the utterance is highly context-related, so that there can be different possible emotions given the same sentence. All these deficiencies lead to increased difficulty in emotion classification from pure texts. Datasets from social media, such as Weibo or Red, are expected to mitigate the complexity of Chinese.

7.2.2. Data Translation with Information Loss

When feeding the original RoBERTa Model (English version), Google API doesn't always generate satisfying translations from Chinese. And this critical process will result in information loss and mislabeling of the model. In the future, with time permitted, professional manual translation is expected to retain as much semantic meaning as possible in Chinese.

7.2.3. Emotion Labeling using ChatGPT-4

Due to the time limit, when we perform mapping from original CPED labels to GoEmotions labels, ChatGPT can make mistakes in the process, and that generates more challenges for classification. Manual labeling is expected to better match different emotions.

7. Reference

Abdul-Mageed, M., & Ungar L. (2017). Emonet: Fine-grained emotion detection with gated recurrent neural networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume1:Long Papers)*. 718-728. https://aclanthology.org/P17-1067

Bostan, L. A. M., & Klinger R. (2018). An Analysis of Annotated Corpora for Emotion Classification in Text. *Proceedings of the 27th International Conference on Computational Linguistics*. 2104-2119. https://aclanthology.org/C18-1179

Chen, Y., Fan, W., Xing, X., Pang, J., Huang, M., Han, W., ... & Xu, X. (2022). Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.372

Ekman P. (1992). Anargument for basic emotions. *Cognition&Emotion*,6(3-4):169–200

Öhman, E., Pàmies, M., Kajava, K., & Tiedemann., J. (2020). XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. *Proceedings of the 28th International Conference on Computational Linguistics*. 6542-6552. https://aclanthology.org/2020.coling-main.575

Plutchik R. (1980). Ageneralpsychoevolutionary theory of emotion. *Theories of emotion*, pages 3–33. Elsevier.

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology, 67*(1), 55. https://doi.org/10.1037/0022-3514.67.1.55

Y. Cui, W. Che, T. Liu, B. Qin and Z. Yang, "Pre-Training With Whole Word Masking for Chinese BERT," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3504-3514, 2021, doi: 10.1109/TASLP.2021.3124365.

Zainuddin, N., & Selamat, A. (2014). Sentiment analysis using support Vector Machine. *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, 333–334. https://doi.org/10.1109/i4ct.2014.6914200

8. Appendix

8.1. Appendix A. CPED Dataset Statistics

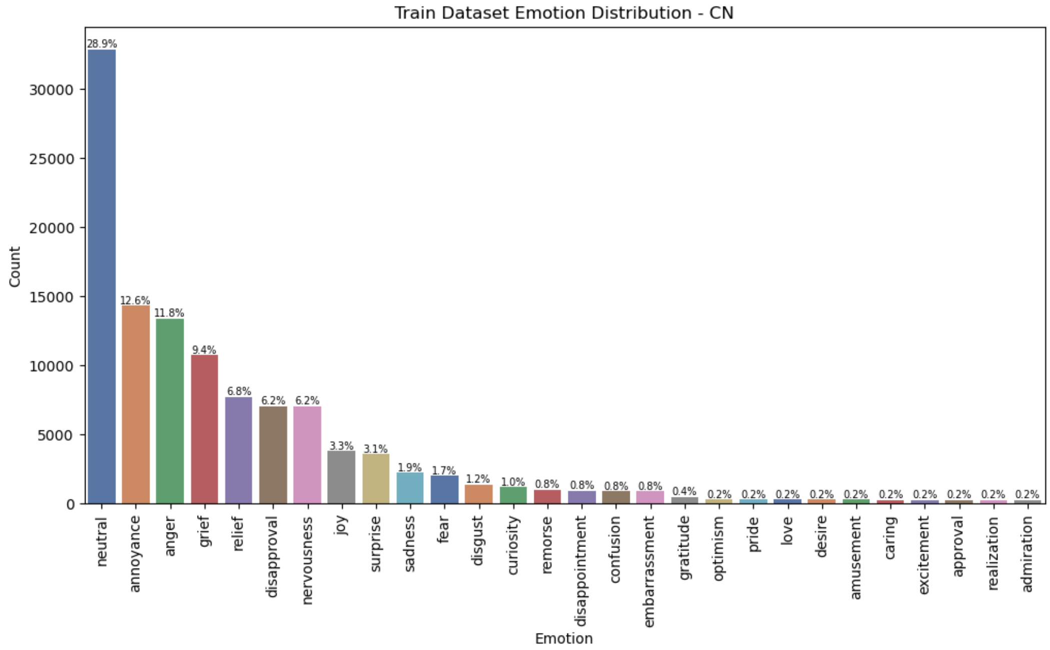


Table 1. Emotion Label Distribution in Training Set - CPED

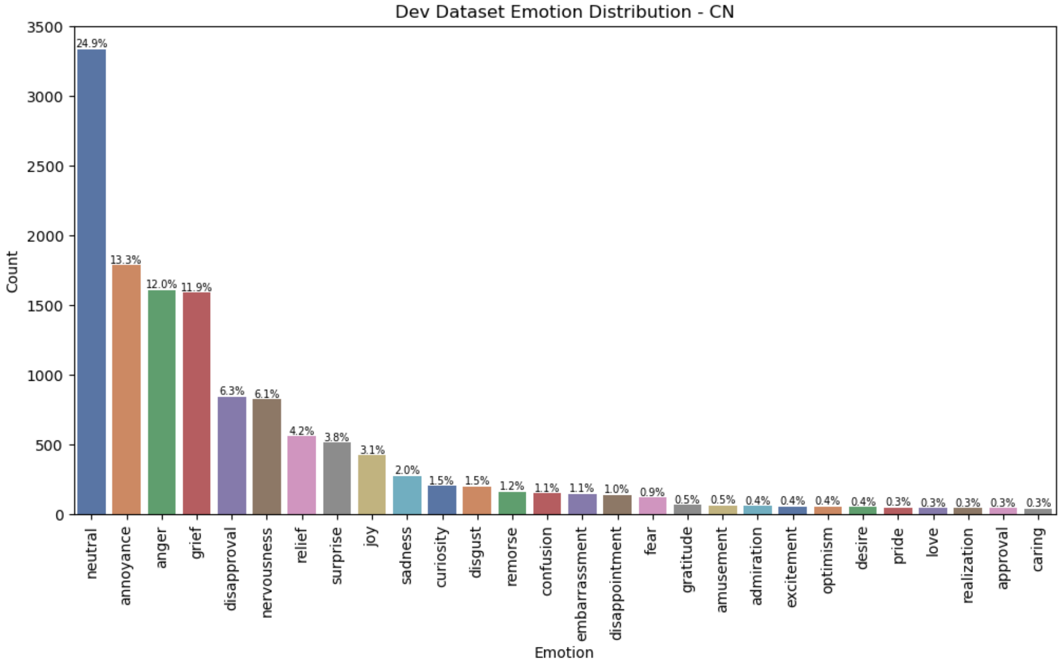


Table 2. Emotion Label Distribution in Validation Set - CPED

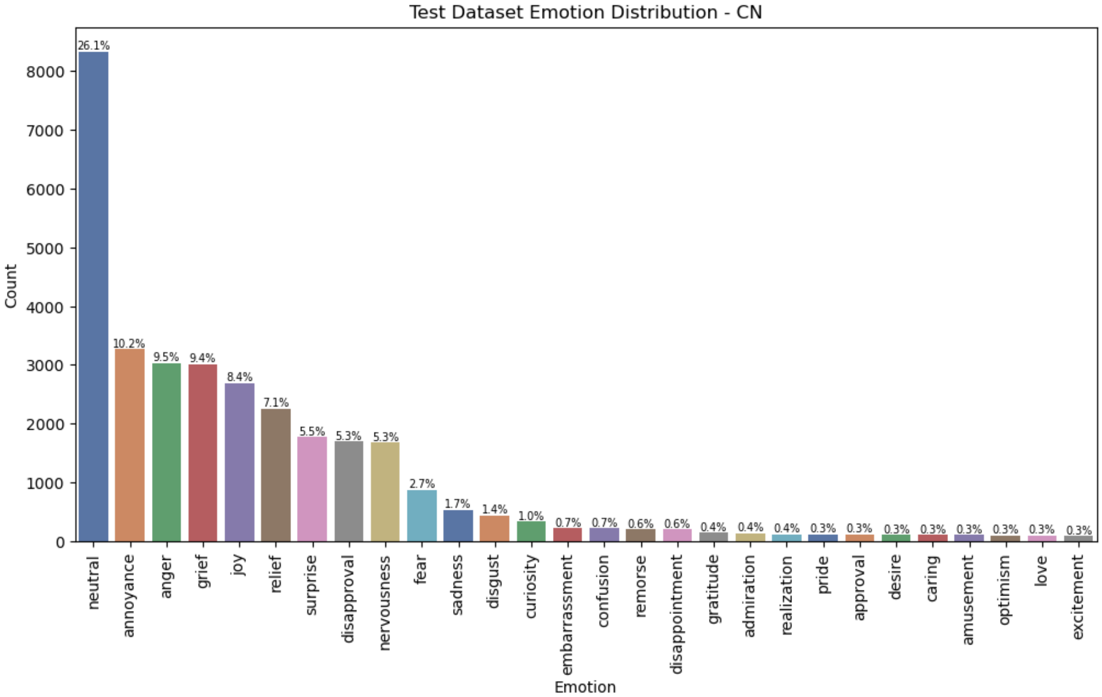


Table 2. Emotion Label Distribution in Test Set - CPED

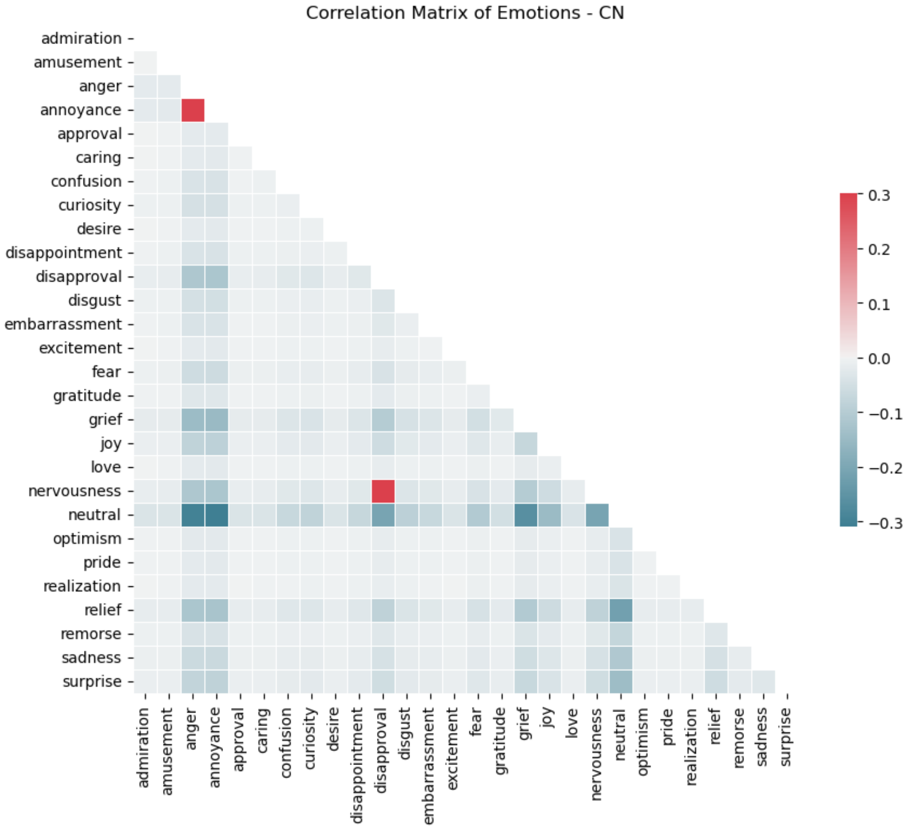


Table 4. Correlation Matrix of Emotions - CPED

8.2. Appendix B. Other Hyperparameters of Model

Other Hyperparameters	**Description**
Weight Decay (weight_decay)	0.01, used for L2 regularization to prevent overfitting.
Adam Epsilon (adam_epsilon)	1e-8, a small number to prevent division by zero in the Adam optimizer.
Maximum Gradient Norm (max_grad_norm)	1.0, for clipping gradients to stabilize training.

Table 5. Other Hyperparameters of Advanced Model (referred to section 4.3)