11-711 Advanced NLP Project3 Report

Team 11

Yiteng (Terrance) Mu (yitengm@andrew.cmu.edu)

Sennan (Sean) Cen (sennanc@andrew.cmu.edu)

Xiaonan (Nancy) Sun (xiaonan2@andrew.cmu.edu)

1. Introduction

1.1. Background

The precise prediction and interpretation of emotions in language is a pivotal advancement in natural language processing (NLP). This capability has a wide range of scientific and practical applications, from engineering empathetic AI in chatbots to the identification and mitigation of harmful online behavior. The progress in this area is fundamentally linked to the availability of extensive, well-annotated datasets that encompass a diverse range of emotions. Such datasets not only aid in refining algorithms but also enhance the overall understanding of emotional expression in language, marking a critical step forward in both scientific and practical domains of NLP.

1.2. Task: Reproducing and Enhancement of Emotion Classification Models

We identified the importance of accurately discerning and interpreting emotions in textual data. In Project 3, our task involved reproducing the model put forward by Demszky et al. (2020), analyzing its performance, and pinpointing areas that hold potential for further improvement. Central to this task is the exploration and utilization of the GoEmotions dataset's detailed emotion taxonomy for refined emotion classification. The GoEmotions dataset consists of 58,000 English Reddit comments, each classified into one of 27 nuanced emotion categories or designated as neutral. The dataset's extensive scope and the detailed categorization of emotions enables advanced emotion classification methodologies.

2. Baseline Model Details

2.1. Train-Test Split

Our GoEmotion dataset includes around 54,000 English Reddit comments categorized into various emotion categories. It has 43,410 training data, 5,426 validation data, and 5,427 test data.

2.2. Model Architecture and Configuration

We have observed that the model given in the tutorial, PRADO, seems somewhat insufficient for the task. We employed the RobertaConfig from the pre-trained 'roberta-base' model, tailored specifically for the 'GoEmotions' fine-tuning task.
Key configuration parameters included:

Parameter Name	**Description**
Number of Labels	Corresponding to the number of emotion classes, defining the output layer dimension.
Attention and Hidden States	Both output_attentions and output_hidden_states were set to False, indicating that the model does not output attention weights or hidden states.
Tokenization	Utilizing the RobertaTokenizer from 'roberta-base', with do_lower_case set to False to maintain the original casing of the text.

With the RobertaTokenizer, each input sequence was truncated or padded to a fixed length of 200 tokens. And batch size is set to 32, with shuffling enabled.

Training hyperparameters included:

Hyperparameters Name	**Description**
Learning Rate	Experimentation with 1e-5 and 2e-5 (optimal learning rate = 2e-5).

Epochs	A total of 5 training epochs (optimal epoch = 3).
Weight Decay	Set to 0.0.s
Adam Epsilon (AdamW)	Set to 1e-8.
Gradient Accumulation Steps	Set to 1, implying that gradients are updated after every batch.
Warmup Steps	Approximately 10% of the total training steps, calculated as 0.10 * args['t_total'].

In the optimization stage of our model training, we implemented a parameter grouping strategy for weight decay, where parameters like 'bias' and 'LayerNorm.weight' were excluded from decay. The choice of optimizer was AdamW, configured with the learning rate and epsilon as specified in the arguments (args). Complementing this, a linear learning rate scheduler with a warm-up phase was employed, in alignment with the optimizer settings.

During the training and validation loop, several key steps were followed for each batch. Initially, the gradients of the model parameters were reset to zero. This was followed by the computation of the mean loss and the execution of backpropagation. After updating the model parameters with the optimizer, the learning rate scheduler was invoked to adjust the learning rate. Post-training and validation of each epoch, the average loss was computed to assess model performance. To ensure the preservation of progress, the model state was checkpointed and saved to a specified directory after the completion of each epoch. This comprehensive approach was crucial in maintaining the efficiency and effectiveness of the model training process.

3. Analysis

We proved that a minimum of four epochs is essential for the model to adequately learn from the data (Demszky et al., 2020) by plotting the training loss and validation loss after each epoch. In the following plot, the decreasing trend in the training loss indicates that the model is learning effectively. However, the validation loss's slight increase in the later epochs might suggest the beginning of overfitting.

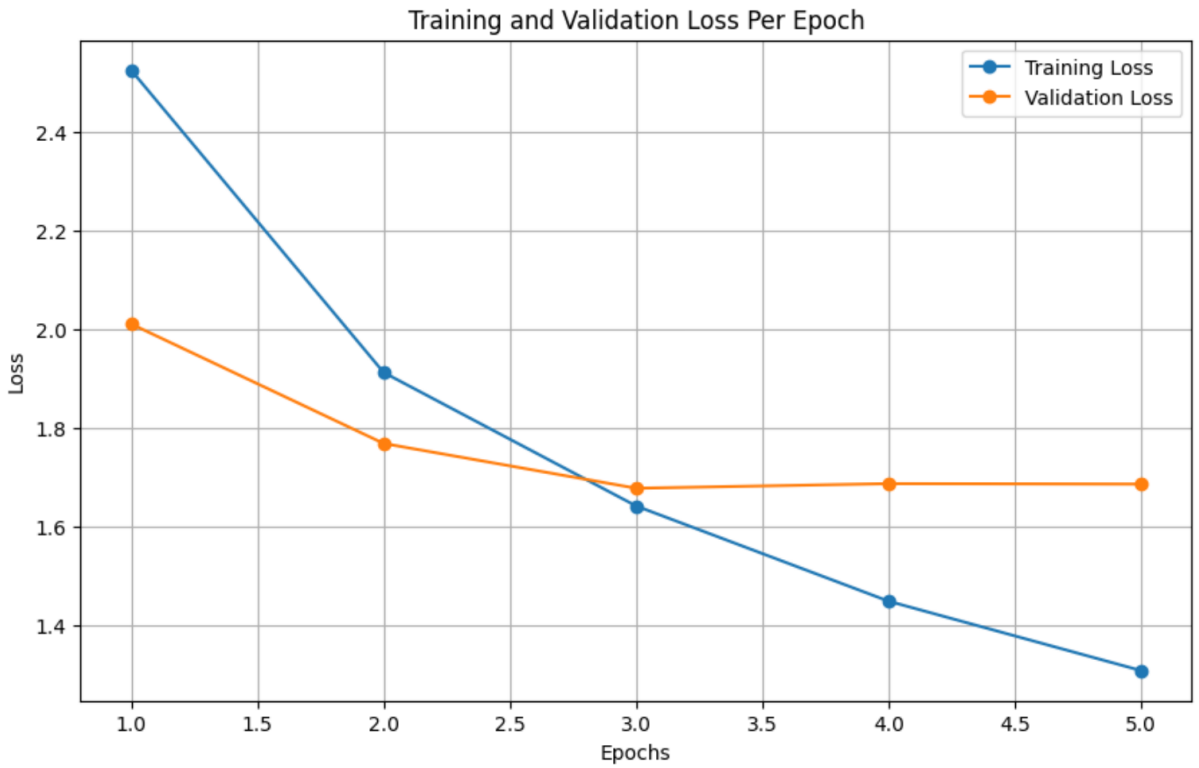

Figure 3.1 Training and Validation Loss Per Epoch

Similar to the approach outlined in the referenced paper, we used precision, recall, and the F1 score as our performance metrics to evaluate the model's performance on new, unseen data. Among these, the F1 score was deemed the most critical indicator. The results of this evaluation are presented below:

```
{'admiration': [0.59, 0.63, 0.61],
 'amusement': [0.71, 0.88, 0.79],
 'anger': [0.44, 0.42, 0.43],
 'annoyance': [0.32, 0.17, 0.22],
 'approval': [0.37, 0.24, 0.29],
 'caring': [0.32, 0.23, 0.27],
 'confusion': [0.31, 0.24, 0.27],
 'curiosity': [0.45, 0.48, 0.46],
 'desire': [0.5, 0.36, 0.42],
 'disappointment': [0.27, 0.13, 0.17],
 'disapproval': [0.26, 0.19, 0.22],
 'disgust': [0.47, 0.4, 0.44],
 'embarrassment': [0.5, 0.13, 0.21],
 'excitement': [0.42, 0.36, 0.39],
 'fear': [0.53, 0.66, 0.59],
 'gratitude': [0.81, 0.84, 0.83],
 'grief': [0.0, 0.0, 0.0],
 'joy': [0.49, 0.52, 0.5],
 'love': [0.61, 0.78, 0.68],
 'nervousness': [0.0, 0.0, 0.0],
 'optimism': [0.5, 0.49, 0.5],
 'pride': [0.0, 0.0, 0.0],
 'realization': [0.35, 0.11, 0.17],
 'relief': [0.0, 0.0, 0.0],
 'remorse': [0.53, 0.67, 0.6],
 'sadness': [0.4, 0.45, 0.43],
 'surprise': [0.41, 0.37, 0.39],
 'neutral': [0.55, 0.7, 0.62]}
```

```
{'admiration': [0.53, 0.83, 0.65],
 'amusement': [0.7, 0.94, 0.8],
 'anger': [0.36, 0.66, 0.47],
 'annoyance': [0.24, 0.63, 0.34],
 'approval': [0.26, 0.57, 0.36],
 'caring': [0.3, 0.56, 0.39],
 'confusion': [0.24, 0.76, 0.37],
 'curiosity': [0.4, 0.84, 0.54],
 'desire': [0.43, 0.59, 0.49],
 'disappointment': [0.19, 0.52, 0.28],
 'disapproval': [0.29, 0.61, 0.39],
 'disgust': [0.34, 0.66, 0.45],
 'embarrassment': [0.39, 0.49, 0.43],
 'excitement': [0.26, 0.52, 0.34],
 'fear': [0.46, 0.85, 0.6],
 'gratitude': [0.79, 0.95, 0.86],
 'grief': [0.0, 0.0, 0.0],
 'joy': [0.39, 0.73, 0.51],
 'love': [0.68, 0.92, 0.78],
 'nervousness': [0.28, 0.48, 0.35],
 'optimism': [0.41, 0.69, 0.51],
 'pride': [0.67, 0.25, 0.36],
 'realization': [0.16, 0.29, 0.21],
 'relief': [0.5, 0.09, 0.15],
 'remorse': [0.53, 0.88, 0.66],
 'sadness': [0.38, 0.71, 0.49],
 'surprise': [0.4, 0.66, 0.5],
 'neutral': [0.56, 0.84, 0.68]}
```

Figure 3.2 Comparative Results - Paper (Left) vs. Our Study (Right). The data is presented in the format of 'label':[precision, recall, F1 score] for each category.

We conducted a paired t-test to compare our model's performance against the best-performing model reported in the paper. The results reveal significant differences: while the precision values were comparable (P-value: 0.976), both recall (P-value: 6.814e-11) and F1 scores (P-value: 4.141e-05) showed notable disparities.

Additionally, we created charts to visually compare the performance metrics of our study against those reported in the paper. These charts, which cover eight selected labels, demonstrate that in most instances, our model outperforms the best model described in the paper. For a comprehensive overview, charts covering all labels are provided in the script.

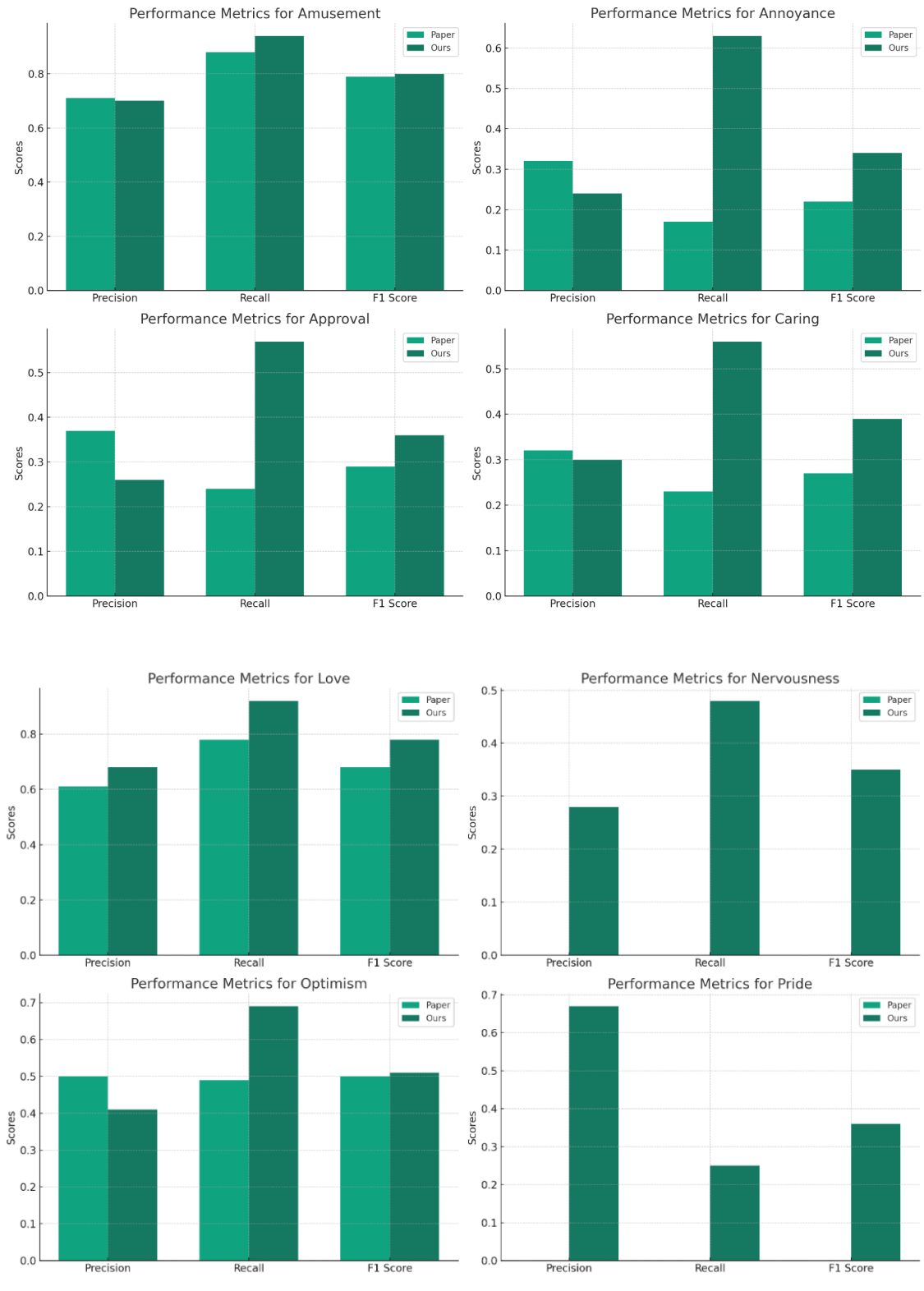

Figure 3.3 Comparative Analysis of Model Performance Across Eight Chosen Emotion Labels

4. Limitation and Further Work

4.1. Limitation

One major limitation we observed, which is also recorded within the paper, is the large potential of current model performance. The baseline models utilized by the paper such as Bert Base Cased and biLSTM obtained the F-1 scores which are normally below 0.5(Demszky, 2020). This model performance could depend on the model architecture, parameter setting, dataset quality, and so on. In this case Bert-base model as a pre-train model might not be trained on a significantly large number of samples to further functionize well in this GoEmotion dataset.

Another limitation that can be seen from the result records is the diverse performance of metrics across different emotion labels. Some emotion labels preserve a good performance while some obtain a value such as 0.0. According to Demszky, this is related to the low frequency of such emotion labels that further negatively affect the model performance(2020).

4.2. Future Work

The GoEmotions dataset, with its extensive range of annotated emotions, offers a promising foundation for various future research and application developments in the field of Natural Language Processing (NLP). Its potential extends well beyond emotion prediction.

In the future, we decide to explore the following aspects:

- Developing more sophisticated models that can understand the nuances and complexities of emotions in text;
- The current training process over the fine-tuned model takes a long time for every single epoch. Examining a less complicated model or with fewer parameters on dataset while retaining the similar or better performance could be a direction to look into.
- Adapting and applying the GoEmotions dataset to other domains, such as social media platforms or literary analysis, to understand how emotion expression varies across different contexts;
- Exploring model's potential in different languages or cultural contexts, potentially through translation and re-annotation efforts.

5. Reference

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