Name: Teo Jin Yee Shawn
Matriculation No.: A0273060N

Abstract
Emotional awareness is a crucial component of mental well-being, yet there are many who struggle to identify their emotions clearly. To explore this domain, this project implements an emotion-classification system that predicts the most likely emotion from a text input. I believe that this model can have potential applications in areas like digital journaling and social media platforms.

Dataset
With practical applications in these platforms in mind, I chose the GoEmotions dataset since it aligns with the more personal and informal language used on such platforms.This dataset contains a large corpus of 58,009 carefully curated comments from Reddit, with human annotations to 27 fine-grained emotion categories or Neutral.

Preprocessing
With the limitation of CPU and GPU in mind, I decided to split the dataset into training, validation and testing dataset of 15000, 3000 and 3000 observations respectively. The 28 categories from the original dataset including the Neutral label were used in the model.

Model
After attempting to train the model on BERT-base and RoBERTa, I found that the model training runtime was prohibitively long in the Google Colab environment. I chose to use DistilBERT-base instead, which retains more than 95% of BERT's performance while having 40% fewer parameters.
The model will find the single label with the highest probability from each input text, based on a probability distribution of the 28 categories. It will then compare with the multi labels in the original dataset. A correct prediction will imply that the predicted label is in the list of actual labels, and a wrong prediction otherwise.

Finetuning
To efficiently fine-tune the model on limited memory, I adopted parameter efficient fine-tuning (PEFT) strategies - specifically Low-Rank Adaptation (LoRA) and Prompt-Tuning. Both techniques significantly reduce the number of trainable parameters while maintaining decent accuracy.

Hyperparameters

Parameter	LoRA	Prompt-Tuning
Learning Rate	1e^-4	1e^-4
No. of Epochs	2	2

Batch Size (Train/Eval)	32/64	32/64
Max Sequence Length	128	128

LoRA used rank = 4, α = 8, and dropout = 0.05 on the query and value projections of the attention layers while prompt-Tuning used 8 virtual tokens initialized randomly. These proportions ensured reasonable generalization while keeping runtime under 20 minutes per model.

Results

Model evaluation was performed on standard metrics: accuracy, precision, recall and F1 score, where the single label Comparing the results of both models, LoRA generally achieved better results as compared to Prompt-tuning. This could be due to the specificity of Prompt-Tuning, which in this context, could fail to perform as well as LoRA.

=== Any-of PRF (LoRA) ===
anyof_accuracy: 0.5250
anyof_precision_micro: 0.5250
anyof_recall_micro: 0.4504
anyof_f1_micro: 0.4848
anyof_precision_macro: 0.4190
anyof_recall_macro: 0.2259
anyof_f1_macro: 0.2420
anyof_f1_samples: 0.4935

=== Any-of PRF (Prompt-Tuning) ===
anyof_accuracy: 0.3420
anyof_precision_micro: 0.3420
anyof_recall_micro: 0.2934
anyof_f1_micro: 0.3158
anyof_precision_macro: 0.0574
anyof_recall_macro: 0.0433
anyof_f1_macro: 0.0298
anyof_f1_samples: 0.3267

To allow for a qualitative evaluation, I gave the model AI-generated inputs of possible social media comments to evaluate the model manually.

LoRA predictions:
• I finally got the job offer I've been waiting for! → excitement
• I'm so thankful that everything turned out okay in the end. → gratitude
• I studied so hard, but I still didn't pass the exam. → disappointment
• I can't believe the delivery was delayed again for no reason! → surprise
• Wow, I didn't expect to see you here after all these years! → surprise

• I’m really nervous about tomorrow’s presentation; my hands are shaking. → disappointment
• That meme literally made me laugh out loud! → neutral
• The food tasted awful and the smell made me feel sick. → sadness
• I might just stay home today and watch some shows. → neutral
• Her dedication and hard work truly inspire everyone around her. → admiration

Prompt-tuned predictions:
• I finally got the job offer I’ve been waiting for! → neutral
• I’m so thankful that everything turned out okay in the end. → neutral
• I studied so hard, but I still didn’t pass the exam. → neutral
• I can’t believe the delivery was delayed again for no reason! → neutral
• Wow, I didn’t expect to see you here after all these years! → admiration
• I’m really nervous about tomorrow’s presentation; my hands are shaking. → neutral
• That meme literally made me laugh out loud! → neutral
• The food tasted awful and the smell made me feel sick. → neutral
• I might just stay home today and watch some shows. → neutral
• Her dedication and hard work truly inspire everyone around her. → admiration

The prompt-tuned predictions are more likely to label the input as neutral, which may explain its poor performance in this dataset, where neutral has a low frequency of occurrence within the 28 categories. However, both fine-tuning strategies performed better than the baseline, where each of the 28 categories have equal chance of being chosen.

Key Takeaways
LoRA provided slightly higher accuracy, precision, recall and F1 score, suggesting that small weight updates in attention layers can better capture emotional cues. DistilBERT also offers an excellent speed-performance trade-off for small-scale model training, which could be used to train for starting hyperparameters before using more computationally-intensive models like BERT or RoBERTa.

Limitations
Only 36.2% of the samples were used out of the full 58000 observations due to resource limits. Training on the full dataset would likely improve model robustness by a significant extent. As the model currently predicts one emotion per input, extending to a multi-label objective could serve greater practical use at the likely expense of poorer performance. There was also class imbalance in the dataset, which could be a reflection of the sentiment likelihood in a random sample on Reddit. This will therefore yield poorer results when using string input from other platforms.

Declaration of Use of AI
I used GPT-5 to produce drafts for the sections of the code (eg. model training, evaluation, etc.), improve model training runtime by adjusting parameters, refine the code, and refine the grammar of my report. I am responsible for the content and quality of the submitted work.