STATS 507 Project Proposal

Overview

Nowadays, transformers models like BERT have become the standard for natural language understanding tasks. However, the original BERT-base model contains around 110 million parameters, making it computationally expensive to train and fine-tune. While DistilBERT is known as a lighter and faster variant of BERT, it still has over 66 million parameters. In the real world, with growing demand for NLP on portable devices without GPU access, there's a clear need for truly compact models. However, it remains unclear how much accuracy is lost when moving from medium-sized models like DistilBERT to much smaller variants such as BERT-Small, BERT-Mini, or BERT-Tiny, and which model offers the best trade-off between model size, computational efficiency, and classification accuracy. To solve this problem, this project will systematically fine-tune and compare several compact BERT variants on a multi-class emotion classification task. The goal is to identify which emotions are most challenging for smaller models and provide recommendations for model selection.

I will use the dair-ai/emotion dataset [1] on Hugging Face, which is a dataset of English Twitter messages with six basic emotions: anger, fear, joy, love, sadness, and surprise. For the models, I will fine-tune and evaluate several Pytorch pre-trained Transformer models of different sizes in Hugging Face, including BERT-Tiny, BERT-Mini, BERT-Small, and DistilBERT-base-uncased.

Prior Work

Literature review

The dair-ai/emotion dataset, originally introduced by Saravia et al. [1], contains 20,000 English tweets annotated with six basic emotions—anger, fear, joy, love, sadness, and surprise. In recent years, Transformer-based models have become a dominant approach for emotion classification in NLP, with many studies showing that models such as BERT and DistilBERT achieve strong performance on emotion recognition tasks. Cortiz et al. [2] compared five widely used Transformer architectures—BERT, DistilBERT, RoBERTa, XLNet, and ELECTRA—on a fine-grained 28-class emotion dataset. Their results showed that the performance differences among these medium-sized models were relatively small, while DistilBERT, the smallest model among them, achieved the best trade-offs between accuracy and computational efficiency. However, this study only focused on standard or medium-sized Transformer models and did not evaluate more compact variants such as BERT-Small, BERT-Mini, or BERT-Tiny.

Besides, other research has investigated compact BERT models. Nityasya et al. [3] conducted a comprehensive knowledge-distillation study from BERT-base to various student models, showing that BERT-Small and BERT-Mini can retain much of the teacher's performance on text classification tasks while being smaller and faster,

whereas BERT-Tiny showed a larger drop in accuracy. However, currently there is no systematic comparison of BERT-Tiny, BERT-Mini, BERT-Small, and DistilBERT on emotion classification tasks. This project aims to fill this gap by evaluating the performance–efficiency trade-off of compact BERT variants on a standard emotion classification dataset.

Method

First, I will load and preprocess the dair-ai/emotion dataset with the Hugging Face datasets library, including tokenization and text cleaning. The dataset provides predefined training, validation, and test splits, which I will use for model training, model selection, and final evaluation, respectively. Next, I will fine-tune several compact pre-trained models (BERT-Tiny, BERT-Mini, BERT-Small, and DistilBERT) on the emotion classification task by following a standard Transformer fine-tuning pipeline on Hugging Face. This process includes loading each model's tokenizer, applying tokenization to the dataset, initializing the model with a classification head, and training it using the Trainer API under consistent hyperparameter settings.

Model performance will be evaluated using accuracy, macro precision, macro recall, and macro F1-score. Additionally, per-class F1-scores and confusion matrices will be computed to identify which emotions are particularly challenging for smaller models. These metrics provide the assessment for both overall performance and class-level behavior. To analyze computational efficiency, I will measure parameter count, training time, and inference speed. Together, these metrics will enable a systematic comparison of performance–efficiency trade-offs across compact BERT variants.

Preliminary Results

Data understanding

The dair-ai/emotion dataset contains 20,000 English tweets labeled with six emotion classes: anger, fear, joy, love, sadness, and surprise. As shown in figure 1, the dataset is imbalanced, with joy and sadness appearing more frequently than love and surprise. This imbalance may introduce challenges during model training, as compact models tend to have more difficulty learning and correctly predicting the minority classes. Although no major data quality issues were observed, tweets vary widely in length and often contain informal language, emojis, hashtags, and abbreviations, which may require careful tokenization during the fine-tuning process.

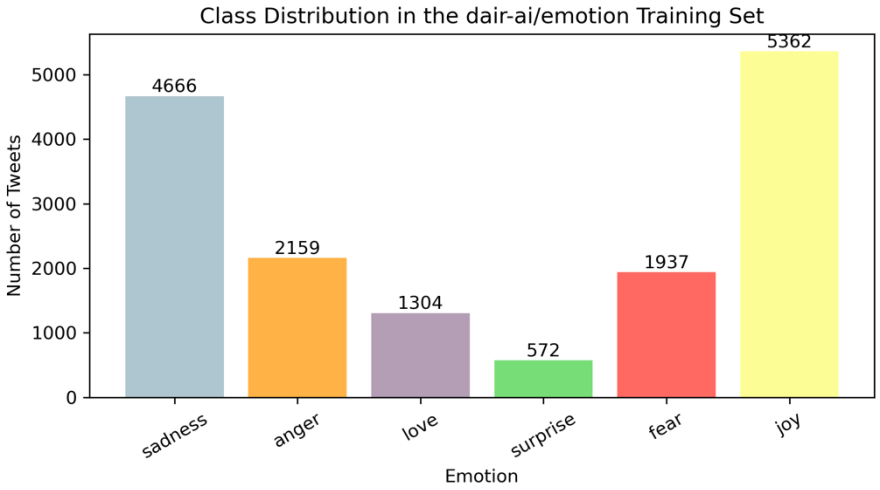

Figure 1: Emotion Distribution

Basic Model

Before fine-tuning, I examined the size of each compact BERT variants. Figure 2 shows the total number of parameters for the four models used in this project. As expected, the parameter counts increases substantially from BERT-Tiny (4.39M) to DistilBERT (66.96M). All the models can be loaded on my laptop without GPU support, though the training time for DistilBERT is the longest.

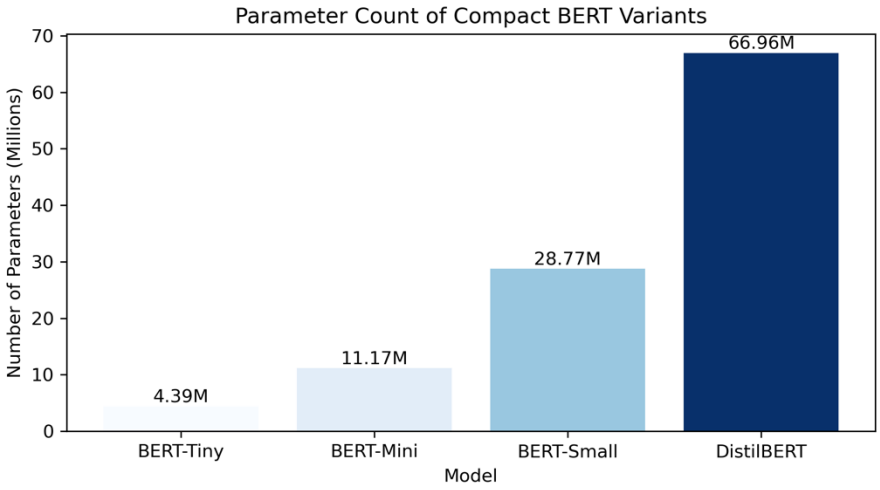

Figure 2: Model Parameters

I have encountered several performance bottlenecks. First, smaller models such as BERT-Tiny and BERT-Mini tended to struggle with capturing small differences between certain emotions such as fear and surprise, leading to a lower per-class accuracy. In addition, the informal and noisy nature of tweets made tokenization less reliable and could cause unstable model performance.

Tools

The main tool from class that I will use is PyTorch, which provides the foundation for

model training and optimization. In addition, I will use Matplotlib and Seaborn to visualize dataset statistics and show the model comparison results. I will explore Hugging Face transformers library for model fine-tuning and Hugging Face Trainer API for managing training and evaluation.

Project Deliverables

A successful project will produce a systematic comparison of compact BERT variants—BERT-Tiny, BERT-Mini, BERT-Small, and DistilBERT on emotion classification tasks. This project will generate fine-tuned models, quantitative evaluation results, and measurements of computational efficiency. The final report will summarize the performance–efficiency trade-offs across the four models. It will also identity which emotions are most difficult for smaller models and provide recommendations for model selection under limited computational resources. Visualizations will be included to support all the findings.

Sub-goals include:
(1) Preprocess and examine the dair-ai/emotion dataset.
(2) Fine-tune four compact BERT variants.
(3) Evaluate classification performance of each model.
(4) Analyze computational efficiency of each model.
(5) Generate visualizations to compare performance and efficiency.
(6) Summarize the overall findings and insights.

Timeline

Week 1-2: Formulate the research question, select a dataset, preprocess the dataset, generate basic visualizations, and load all four compact BERT models.
Week 3-4: Tokenize the dataset, fine-tune the four models, calculate performance and efficiency metrics, and write the project proposal.
Week 5: Complete remaining evaluation work, generate final visualizations, and finish writing the final report summarizing all findings.

References

[1] Saravia, E., Liu, H.-C. T., et al. (2018). CARER. _Proceedings of EMNLP_, 3687-3697. https://www.aclweb.org/anthology/D18-1404/
[2] Cortiz, Diogo. “Exploring Transformers Models for Emotion Recognition: A Comparision of BERT, DistilBERT, RoBERTa, XLNET and ELECTRA.” 2022, pp. 230–34, https://doi.org/10.1145/3562007.3562051.
[3] Nityasya, Made Nindyatama, et al. Which Student Is Best? A Comprehensive Knowledge Distillation Exam for Task-Specific BERT Models. 2022-01, https://doi.org/10.48550/arxiv.2201.00558.