An Evaluation of Compact BERT Models for Emotion Classification

Yichen Zhao
Department of Biostatistics
University of Michigan
Ann Arbor, USA
micozhao@umich.edu

***Abstract*—Compact transformer models have gained increasing attention for their efficiency and good performance, however, their performance trade-offs in emotion classification tasks remain less explored. In this project, I evaluate four compact BERT variants (DistilBERT, BERT-Small, BERT-Mini, and BERT-Tiny) using the dair-ai/emotion dataset. All models are fine-tuned under the same training settings, and their performance is compared in terms of accuracy, macro F1-score, per-class results, and computational efficiency. The results show that DistilBERT achieves the best overall performance and performs consistently well across all emotion categories, while BERT-Small and BERT-Mini shows competitive accuracy with shorter training time and lower inference cost. BERT-Tiny provides the fastest runtime but suffers from classification quality decrease, especially for less frequent emotions. These findings show a clear performance–efficiency trade-off and provide us with model selection advice in real-world applications.**

***Index Terms*—Compact BERT, Emotion Classification, Transformer Models, PyTorch, Fine-tuning**

I. Introduction

Transformer models have become the dominant approach for modern natural language understanding tasks. While BERT-base model achieves strong predictive performance, its 110-million parameter size makes training and fine-tuning computationally costly, particularly on CPU-only environments. DistilBERT offers a faster and lighter alternative, but still contains more than 66 million parameters. In the real world, with the increasing demand for efficient NLP models on laptops and mobile devices, there is a growing need for compact models. In order to understand how much accuracy and robustness are lost when moving from medium-sized models like DistilBERT to much smaller variants such as BERT-Small, BERT-Mini, and BERT-Tiny. However, prior work provides limited evidence on how they compare in terms of accuracy, computational efficiency, and class-level behavior on a multi-class emotion classification task.

Prior research have mainly examined standard or medium-sized transformer models. Using a fine-grained 28-class emotion dataset, Cortiz [1] compared BERT, DistilBERT, RoBERTa, XLNet, and ELECTRA. They found that performance differences across these models were relatively small, while DistilBERT, the smallest model among them, offered the best balance between accuracy and computational efficiency. Yet, their study did not extend to smaller variants such as BERT-Small, BERT-Mini, or BERT-Tiny. Other work by Nityasya et al. [2] conducted a knowledge-distillation study from BERT-base to multiple student models. Their findings indicated that compact models such as BERT-Small and BERT-Mini could preserve much of the teacher model's performance while being smaller and faster, but the smallest BERT-Tiny model showed a large performance drop. Despite the prior work, few studies provided systematic comparison conducted on compact BERT variants specifically for multi-class emotion classification tasks.

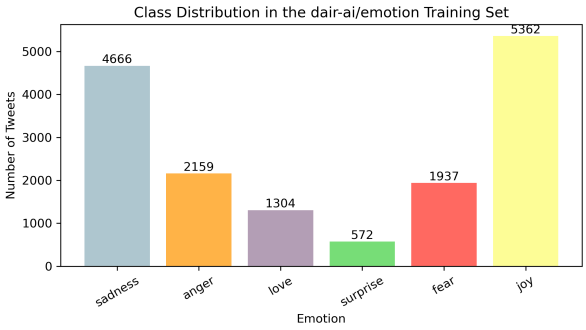

Fig. 1. Class distribution in the dair-ai/emotion training set.

II. Method

A. Data

I conduct all experiments using the dair-ai/emotion dataset [3] on Hugging Face. The dataset contains approximately 20,000 English tweets, each labeled with one of six basic emotions: anger, fear, joy, love, sadness, and surprise. The dataset is provided with predefined training, validation, and test splits in an 8:1:1 ratio. Each instance consists of a short text segment and an integer label that shows its emotion category. The problem can be formulated as a multi-class text classification task, in which the input is a tweet and the output is one of six emotion labels.

To better understand the characteristics of the dataset, I visualize the distribution of emotion labels in the training split (Figure 1). The classes are noticeably imbalanced: joy and sadness are the most frequent emotions, while surprise and love appear considerably less often.

B. Preprocessing and Tokenization

All data is preprocessed using the Hugging Face Transformers and Datasets libraries. For each model, I first apply its corresponding pre-trained tokenizer to convert raw text into token IDs, which will be used as the model inputs during fine-tuning. Then, I set the maximum sequence length to 64 tokens. This is based on the analysis that 99% of samples contain fewer than 57 tokens. I also use dynamic padding through DataCollatorWithPadding, which pads each batch only to the length of its longest sequence. It helps reduce unnecessary computation and improves training efficiency. Finally, I fix a random seed to ensure reproducibility in all experiments.

C. Models

This study evaluates four compact transformer models: BERT-Tiny, BERT-Mini, BERT-Small, and DistilBERT. As can been in Table I, all four are based on the original BERT model but use much smaller configurations in terms of depth (number of layers), width (hidden size and attention heads), and total parameter count. These models range from about 4M parameters to over 66M, offering a wide range for studying performance–efficiency trade-offs. All models are fine-tuned with the same training setup to ensure a fair comparison.

TABLE I
COMPARISONS OF COMPACT BERT VARIANTS

Model	Layers	Hidden size	Attention heads	Parameters (M)
BERT-Tiny	2	128	2	4.39
BERT-Mini	4	256	4	11.17
BERT-Small	4	512	8	28.77
DistilBERT	6	768	12	66.96

D. Fine-Tuning Setup

All four models are fine-tuned using the Hugging Face Trainer API with a shared set of hyperparameters to ensure fair comparison. For each model, I use a learning rate of 2×10^{-5}, a batch size of 8 for both training and evaluation, weight decay of 0.01, and train for 3 epochs on the training split. All experiments are run on a CPU-only laptop, so these settings are chosen to balance between good performance and efficient training time.

During fine-tuning, I evaluate the model on the validation split at the end of each epoch and save a checkpoint after each evaluation. I enable `load_best_model_at_end` so that the final model is automatically selected based on the highest macro F1-score achieved on the validation set across all epochs. As a result, the reported test performance reflects the best-performing checkpoint for each model rather than the final epoch.

E. Evaluation Metrics

Model performance is evaluated using accuracy, macro precision, macro recall, and macro F1-score. Because the dataset exhibits clear class imbalance (Figure 1), macro-averaged metrics are preferred over micro-averaged ones, as they assign equal importance to all emotion categories, including underrepresented classes such as surprise and love.

To further examine class-level behavior, I compute per-class F1-scores and generate confusion matrices. It helps identify which emotions are particularly challenging for smaller models and reveal common misclassification patterns.

In addition to predictive performance, computational efficiency is assessed by measuring parameter count, training time, and evaluation speed (samples processed per second). Together, these metrics make it possible to evaluate how each model's effectiveness relates to its computational cost.

TABLE II
OVERALL PERFORMANCE OF COMPACT BERT VARIANTS.

Model	**Accuracy**	**Macro Precision**	**Macro Recall**	**Macro F1**
BERT-Tiny	0.7655	0.7481	0.5918	0.6090
BERT-Mini	0.9080	0.8715	0.8582	0.8626
BERT-Small	0.9205	0.8721	0.8696	0.8704
DistilBERT	0.9315	0.8968	0.8776	0.8849

III. RESULT

A. Overall Performance Comparison

Table II summarizes the performance of the four compact transformer models on the test set. DistilBERT achieves the highest overall scores, with an accuracy of 0.9315 and a macro F1-score of 0.8849. BERT-Small ranks second (macro F1 = 0.8704), followed closely by BERT-Mini (macro F1 = 0.8626). Both models attain accuracies above 0.90. However, BERT-Tiny shows the lowest performance among the four models, with an accuracy of 0.7655 and a macro F1-score of 0.6090. It constantly ranks last across all evaluation metrics.

B. Class-Level Performance

Then I compare the per-class F1-scores of DistilBERT and BERT-Tiny. As shown in Table III, DistilBERT shows consistently strong performance across all six emotion categories. Sadness (F1 = 0.97) and joy (0.95) are the most accurately recognized emotions, while even less frequent emotions such as love (0.84) and surprise (0.73) have achieved relatively high scores. In contrast, BERT-Tiny struggles on minority classes. While it achieves similar results as DistilBERT on joy (0.85) and sadness (0.82), its F1-score drops sharply for love (0.20) and surprise (0.39). This shows that a reduced model size limits the model's capacity to capture subtle linguistic signals especially for less distinct emotions.

TABLE III
PER-CLASS F1-SCORES FOR DISTILBERT AND BERT-TINY.

Emotion	**DistilBERT F1**	**BERT-Tiny F1**
sadness	0.9716	0.8185
joy	0.9525	0.8535
anger	0.9237	0.7326
fear	0.8926	0.6618
love	0.8401	0.2022
surprise	0.7288	0.3853

These differences are also reflected in the confusion matrices. In BERT-Tiny (Figure 3), love is frequently misclassified

as joy or sadness, and surprise is often predicted as fear. This shows the difficulty in distinguishing subtle emotion changes. In contrast, DistilBERT (Figure 2) shows a much clearer diagonal structure and fewer cross-class confusions. The misclassification between joy and love, as well as fear and surprise, are greatly reduced when using DistilBERT compared to BERT-Tiny. This is also consistent with the high per-class F1-scores we observed in Table III.

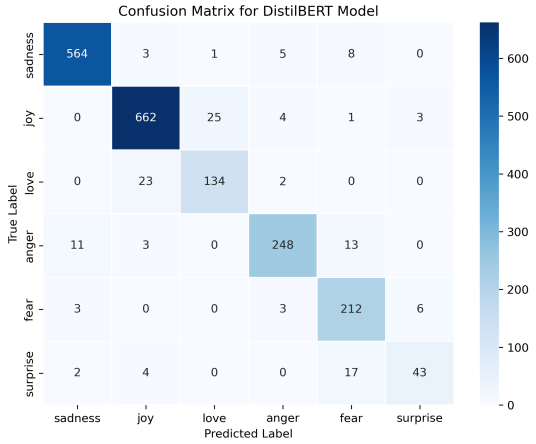

Fig. 2. Confusion matrix for DistilBERT.

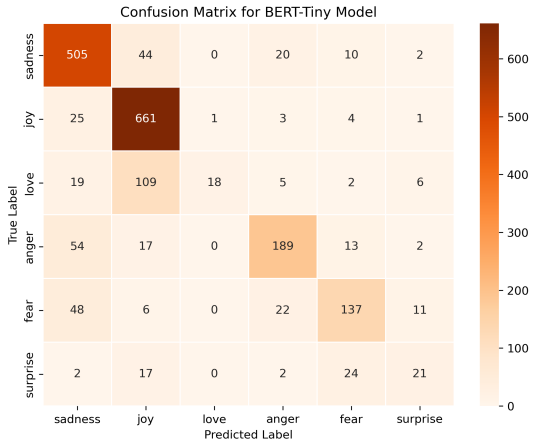

Fig. 3. Confusion matrix for BERT-Tiny.

C. Computational Efficiency

To evaluate the efficiency–performance trade-off across models, I compare their training time and inference speed under the same environment. As can be seen in Table IV, the computational cost increases as model size gets larger. Due to its larger parameter count, DistilBERT, is the slowest to train (12.3 min) and evaluate (6.9 s). BERT-Small shortens the training time by roughly 31% compared to DistilBERT, while they have similar accuracy and macro F1-scores.

The efficiency gap becomes much more obvious for smaller models. BERT-Mini trains about three times faster than DistilBERT, and the evaluation samples go through increases from 288 to 974 samples per second. BERT-Tiny is the fastest model that completes fine-tuning in only 1.4 minutes. It also achieves the fastest inference rate which is over 9 times faster than DistilBERT. However, this efficiency comes at the cost of significantly lower classification performance, particularly on minority emotion categories.

TABLE IV
COMPUTATIONAL EFFICIENCY COMPARISON.

Model	Train Time (min)	Train Samples/s	Eval Time (s)	Eval Sample
DistilBERT	12.31	64.97	6.94	288.17
BERT-Small	9.42	84.90	4.97	402.48
BERT-Mini	4.08	196.06	2.05	974.42
BERT-Tiny	1.39	574.30	0.71	2827.19

IV. Conclusion

In this study, I conducted a systematic comparison of four compact transformer models (DistilBERT, BERT-Small, BERT-Mini, and BERT-Tiny) on a six-class emotion classification task using the dair-ai/emotion dataset. Among the four models, DistilBERT not only achieves the highest accuracy and macro F1-score, but also performs well across all emotion categories, including the less frequent ones. Medium-sized compact models such as BERT-Small and BERT-Mini show competitive results at shorter training time and lower inference cost. Although small-sized model like BERT-Tiny is extremely fast, it sacrifices predictive performance, particularly for less frequent emotions.

Therefore, in terms of model selection, DistilBERT is the best option when accuracy is the priority. BERT-Mini and BERT-Small are suitable for real-world applications when we only have limited computational resources but still demands high speed and reasonable performance. BERT-Tiny is most suitable for lightweight scenarios where fast speed and small model size matter more than accuracy.

There are also several directions for future work. First, these models could be evaluated on emotion datasets with more classes, so we can examine whether the trends observed here also hold under more complex classification settings. In addition, we can apply the same compact models to other text classification tasks such as sentiment analysis and topic categorization to assess the generality of our findings. Finally, we can apply techniques such as data augmentation, class re-weighting, or knowledge distillation to improve the performance of smaller models like BERT-Tiny.

References

[1] D. Cortiz, "Exploring transformers models for emotion recognition: A comparison of BERT, DistilBERT, RoBERTa, XLNet and ELECTRA," in *Proc. 2022 3rd Int. Conf. Control, Robotics and Intelligent System (CCRIS)*, Virtual Event, China, 2022, pp. 230–234, doi: 10.1145/3562007.3562051.

[2] M. N. Nityasya, H. A. Wibowo, R. Chevi, R. E. Prasojo, and A. F. Aji, "Which student is best? A comprehensive knowledge distillation exam for task-specific BERT models," arXiv preprint arXiv:2201.00558, 2022, doi: 10.48550/arXiv.2201.00558.

[3] Dair-ai, "Emotion dataset," Hugging Face, 2020. [Online]. Available: https://huggingface.co/datasets/dair-ai/emotion