DistilBERT vs MiniLM for AG News Topic Classification: Accuracy–Efficiency Trade-offs in Fine-Tuning

Shervin Iranaghideh

Department of Mathematics and Computer Science, Islamic Azad University Science and Research Branch

Abstract- This paper evaluates the effectiveness and efficiency of fine-tuning small pretrained transformer models for news topics classification. Using the AG News dataset, we fine-tune and compare DistilBERT (distilbert-base-uncased) and MiniLM (microsoft/MiniLM-L12-H384-uncased) under the same training configuration. DistilBERT and MiniLM achieve very similar classification performance on AG News (94.67% vs 94.41% accuracy; 0.9468 vs 0.9440 macro-F1). Although MiniLM slightly underperforms DistilBERT in absolute accuracy (−0.26 percentage points), it uses roughly half as many parameters (33.36M vs 66.96M) and exhibits better efficiency. These results show that compact transformers can provide competitive performance on topic classification tasks and that MiniLM offers a favorable accuracy–efficiency trade-off for resource-constrained environments.

Index Terms- Accuracy–efficiency trade-off, AG News, DistilBERT, fine-tuning, MiniLM, natural language processing (NLP), news topic classification, pretrained language models, text classification, transformer models.

I. Introduction

Text classification is a core task in natural language processing (NLP) with direct applications in news organization, content recommendation, trend monitoring, and information retrieval. Among text classification settings, topic classification is especially practical because it enables large collections of articles to be automatically grouped into meaningful categories. Modern approaches commonly rely on pretrained transformer language models, which can be adapted to downstream tasks through fine-tuning and often deliver strong accuracy with limited task-specific feature engineering.

Despite their high performance, transformer models vary widely in computational cost and memory footprint. In many real-world or resource-constrained environments—such as free cloud notebooks, edge devices, or applications with strict latency budgets—model size and efficiency can be as important as raw accuracy. This motivates the use of compact pretrained transformers that aim to preserve performance while reducing the number of parameters and improving practical deployability.

In this work, we study the accuracy–efficiency trade-off of fine-tuning two small pretrained transformer models on the AG News benchmark, a widely used dataset for four-class news topic classification (World, Sports, Business, and Sci/Tech). We fine-tune DistilBERT (distilbert-base-uncased) and MiniLM (microsoft/MiniLM-L12-H384-uncased) using the same training configuration to ensure a fair comparison. Our results show that the two models achieve very similar predictive performance: DistilBERT reaches 94.67% accuracy (Macro-F1 0.9468) while MiniLM reaches 94.41% accuracy (Macro-F1 0.9440). Given this small difference, the comparison is primarily driven by efficiency: MiniLM uses 33.36M parameters, roughly half of DistilBERT's 66.96M parameters, suggesting a substantially better parameter–performance trade-off for applications where memory and model size matter.

II. Dataset and Problem Definition

2.1 Dataset: AG News

This study uses the AG News benchmark for news topic classification. The dataset consists of English-language news articles that are labeled into four topic categories: World, Sports, Business, and Sci/Tech. Following the commonly used benchmark setup, the dataset is provided with a predefined split of 120,000 training samples and 7,600 test samples (1,900 samples per class in each split). (Zhang et al., 2015)

In our experiments, the dataset was loaded using the Hugging Face Datasets library, which provides a standardized implementation of the AG News dataset and its train/test split. (Lhoest et al., 2021)

2.2 Problem Definition

We formulate AG News topic classification as a supervised multi-class text classification problem. Let x denote an input news text and $y \in \{0,1,2,3\}$ denote its class label corresponding to the four categories. The objective is to learn a classifier $f_\theta(x)$ that predicts the correct topic label for each input text.
For evaluation, we report:

- Accuracy, measuring the fraction of correctly classified test instances;
- Macro-averaged F1 (Macro-F1), which computes the unweighted mean of per-class F1 scores and is appropriate for multi-class settings where balanced performance across classes is desired.

2.3 Preprocessing and Label Mapping

The dataset contains a text field and an integer label for each example. We map the numeric labels to the topic names as 0 → World, 1 → Sports, 2 → Business and 3 → Sci/Tech. We apply minimal preprocessing and rely primarily on each model's tokenizer. Each news text is tokenized using the corresponding pretrained tokenizer and truncated to a maximum sequence length of 128 tokens to ensure a consistent input size across models and to control memory usage during training.

III. Methodology

3.1 Overview

This study compares two compact pretrained transformer encoders—DistilBERT (distilbert-base-uncased) and MiniLM (microsoft/MiniLM-L12-H384-uncased)—for four-class news topic classification on the AG News dataset. Both models are fine-tuned under the same experimental conditions to ensure a fair comparison, and performance is evaluated on the standard test split using accuracy and macro-averaged F1 score.

3.2 Input construction and preprocessing

AG News samples can be represented as a single text field or as separate fields such as title and description depending on the dataset source. In our implementation, the model input is formed as a single sequence by using the dataset's text content (or by concatenating the title and description when available). We apply minimal preprocessing and rely on each model's pretrained tokenizer.

All inputs are tokenized with truncation enabled and a maximum sequence length of 128 tokens, to control memory usage and keep training consistent across models.

3.3 Fine-tuning Setup

We fine-tune each pretrained model by adding a classification head that outputs probabilities over the four topic classes. During training, model parameters are updated end-to-end using supervised learning on the AG News training set.

Both models use the same training configuration: 2 epochs, learning rate: 2×10^{-5}, training batch size 16 (evaluation batch size 64), AdamW optimizer with weight decay 0.01, 10% warmup ratio, mixed-precision (FP16) training, and random seed 42.

3.4 Evaluation Metrics

We report two standard metrics for multi-class classification:

- Accuracy, measuring the fraction of correctly classified test examples.
- Macro-F1, computed as the average of the per-class F1 scores. Macro-F1 is included to reflect balanced performance across categories, even when the dataset is approximately balanced.

3.5 Model Comparison Criteria

In addition to predictive performance, we emphasize model efficiency, since the accuracy difference between the models is small. We compare models using:

- Parameter count (millions of parameters) as an indicator of model size and memory footprint,
- Training time (minutes) measured for the full fine-tuning run,

3.6 Implementation

All experiments are implemented using the Hugging Face Transformers training pipeline with PyTorch backend. Models are fine-tuned using the Trainer API, and checkpoints are saved for later evaluation and reuse.

IV. Experimental Environment

Experiments were conducted on Google Colab running Linux (Linux-6.6.105+ x86_64, glibc 2.35) with Python 3.12.12. Training and evaluation were implemented in PyTorch 2.9.0+cu126 with GPU acceleration enabled. We used the Hugging Face libraries Transformers 4.57.3 and Datasets 4.0.0. The runtime provided an NVIDIA Tesla T4 GPU with 15 GB VRAM (15360 MiB), using NVIDIA driver 550.54.15 (NVIDIA-SMI CUDA version 12.4).

V. Results

5.1 Overall Performance on AG News

We evaluated two fine-tuned compact transformer models—DistilBERT (distilbert-base-uncased) and MiniLM (microsoft/MiniLM-L12-H384-uncased)—on the AG News test set using Accuracy and Macro-F1. Both models achieved strong and very similar performance (table below). DistilBERT obtained the highest scores with 94.67% Accuracy and 0.9468 Macro-F1, while MiniLM reached 94.41% Accuracy and 0.9440 Macro-F1. The absolute performance gap is small (−0.26 percentage points Accuracy and −0.0028 Macro-F1 for MiniLM relative to DistilBERT), indicating that both compact models are effective for news topic classification in this setting.

Comparison of fine-tuned models on AG News				
Model	Parameters (M)	Accuracy	Macro-F1	Train time (min)
distilbert-base-uncased	66.96	0.9467	0.9468	13.44
microsoft/MiniLM-L12-H384-uncased	33.36	0.9441	0.9440	12.89

5.2 Model Efficiency and Parameter-focused Comparison

Because predictive performance is near-parity, efficiency becomes the key differentiator. MiniLM uses 33.36M parameters, approximately half of DistilBERT's 66.96M parameters. Despite this reduction in model size, MiniLM's test performance remains close to DistilBERT. This suggests that MiniLM provides a stronger parameter–performance trade-off, achieving comparable results with significantly fewer parameters.

5.3 Training Time

Training time for both models was similar under the same fine-tuning configuration. DistilBERT required 13.44 minutes, while MiniLM required 12.89 minutes, a difference of 0.55 minutes. Although training time is not dramatically different, MiniLM's lower parameter count makes it more attractive for deployment and storage, even when training speed is comparable.

5.4 Summary of Findings

Overall, both models deliver identical performance on AG News in this study. However, MiniLM achieves nearly the same Accuracy and Macro-F1 while using roughly 50% fewer parameters, making it a practical choice for resource-constrained deployment scenarios.

References

[1] Xiang Zhang, Junbo Zhao, & Yann LeCun. (2016). Character-level Convolutional Networks for Text Classification.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin. (2023). Attention Is All You Need.

[3] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics.

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, & Thomas Wolf. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

[5] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, & Ming Zhou. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers.

[6] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45). Association for Computational Linguistics.

[7] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, & Thomas Wolf. (2021). Datasets: A Community Library for Natural Language Processing.

Authors

Author –Shervin Iranaghideh, shervinіranaghideh@gmail.com,+98 993 8064 299.