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 shows marginal difference in absolute accuracy relative to DistilBERT (−0.26 percentage points which is negligible), it uses roughly half as many parameters (33.36M vs 66.96M) and exhibits better efficiency in terms of model size and storage. 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) and 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. Literature Review

2.1 BERT

The language representation model BERT stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT is a Transformer encoder-only architecture designed to learn deep bidirectional contextual representations, meaning each token can attend to both its left and right context in each layer. This bidirectionality is the key conceptual departure from earlier pretraining schemes that were typically unidirectional (left-to-right or right-to-left) or only weakly bidirectional. After pretraining, BERT can be fine-tuned by adding a small task-specific output layer, while keeping the same core architecture across tasks; this makes it a general-purpose language understanding model rather than a task-specific system (Devlin et al., 2019).

Methodologically, BERT is pretrained using two objectives. First, Masked Language Modeling (MLM) randomly masks a subset of input tokens and trains the model to predict the original tokens from the remaining context, enabling true bidirectional conditioning during pretraining. Second, Next Sentence Prediction (NSP) trains the model to predict whether a second segment follows the first in the original text, encouraging representations that capture inter-sentential relationships useful for sentence-pair tasks. The input format uses special tokens (notably [CLS] for an aggregate sequence representation and [SEP] to separate segments), allowing the same model to handle single-sentence and sentence-pair inputs. For downstream tasks, fine-tuning is performed end-to-end: for example, classification often uses the [CLS] representation, while token-level tasks use contextual token embeddings from the encoder (Devlin et al., 2019).

2.2 Knowledge Distillation

Knowledge Distillation (KD), formalized by Hinton et al. (2015), is the general technique of transferring "knowledge" from a large, accurate but expensive model (the teacher) into a smaller, cheaper model (the student). Hinton et al. (2015) did not treat the "knowledge" as an abstract concept, but very concretely as the teacher's full output probability distribution over classes—especially the relative probabilities assigned to incorrect classes, which carry information about class similarities (e.g., which handwritten "2"s look more like "3"s versus "7"s). This is the central motivation: hard one-hot labels tell the student only what is correct, while the teacher's "soft" outputs communicate richer structure learned during training.

For methodology, the core idea is to train the student on a transfer set (which may be labeled or unlabeled) using soft targets produced by the teacher as explained by Hinton et al. (2015). A key practical trick is temperature scaling: the teacher's softmax is evaluated at a higher temperature to produce a "softer" probability distribution (less peaky, more informative), and the student is trained to match those softened outputs (then uses temperature 1 at test time). When labels are available, the student training is improved by combining two signals: (1) matching the teacher's soft targets and (2) also learning from the true hard labels, typically with a smaller weight on the hard-label term so the student still benefits from the teacher's richer supervision. The paper also emphasizes that using soft targets can act as a strong regularizer, because it communicates generalization behavior learned by the teacher, not just training-set memorization.

2.3 DistilBERT: Depth Reduction

DistilBERT is introduced by Sanh et al. (2020) as a general-purpose, compressed version of BERT that is pre-trained with knowledge distillation so it can later be fine-tuned on many downstream tasks in the same way as BERT. The motivation is practical deployment: large pre-trained models are expensive to run under latency, memory, or on-device constraints, so the aim is to keep most of BERT's language understanding while cutting inference cost. Concretely, DistilBERT reduces BERT's size by about 40%, retains about 97% of BERT's GLUE performance, and is reported to be about 60% faster in inference than BERT-base under CPU evaluation.

DistilBERT keeps the same overall Transformer encoder family as BERT, but removes some components and reduces depth. The student model removes token-type embeddings and the pooler, and most importantly cuts the number of Transformer layers in half (Sanh et al. (2020) emphasizes reducing layers as a high-impact lever for compute efficiency). To help optimization and preserve inductive biases from the teacher, the student is initialized from the teacher by taking one layer out of two from the pretrained BERT weights (rather than training the smaller network from scratch).

The key contribution of DistillBERT is performing knowledge distillation during pre-training (not just task-specific distillation after fine-tuning). Sanh et al. (2020) describes a triple loss that combines: (1) a masked language modeling objective (as the supervised pretraining signal), (2) a distillation loss that encourages the student to match the teacher's outputs, and (3) an additional cosine-distance loss to align the directions of student and teacher hidden-state representations. They also follow improved BERT training practices (e.g., dynamic masking) and notably omit the next sentence prediction (NSP) objective during DistilBERT training. The model is pretrained on the same type of corpora used for BERT (English Wikipedia + BookCorpus).

2.4 MiniLM: Deep Self-Attention Distillation

MiniLM stands for "Mini Language Model", Wang et al. (2020) present it as a task-agnostic compression approach for large pre-trained Transformer language models (e.g., BERT). Rather than distilling only final predictions (like masked-token probabilities) or enforcing layer-by-layer matching, MiniLM aims to produce a compact student that can later be fine-tuned normally across tasks while remaining fast and small for deployment. Wang et al. (2020) frames MiniLM as a simple but effective distillation strategy that focuses on the self-attention module—a core computational component of Transformers—and reports that the resulting students can preserve most of the teacher's downstream accuracy with substantially fewer parameters and computations.

What MiniLM is technically is a student Transformer trained to deeply mimic the teacher's self-attention behavior, with two design choices that make it flexible. First, MiniLM distills only from the teacher's last Transformer layer rather than doing strict layer-to-layer alignment; Wang et al. (2020) argues this avoids forcing the student to have the same number of layers as the teacher and avoids searching for an optimal layer mapping. Second, MiniLM introduces a new distillation signal beyond standard attention-map transfer: in addition to matching attention distributions (relationships induced by queries and keys), it also transfers value-relation information (relationships induced by values). This "value relation" is computed in a way that converts vectors into relation matrices of the same size, which allows the student to use different hidden dimensions than the teacher without adding extra projection matrices.

MiniLM's distillation objective therefore combines two complementary signals: (1) attention distribution transfer and (2) self-attention value-relation transfer, both taken from the teacher's last layer to guide the student to mimic how the teacher allocates attention and how it relates token value representations. To further improve distillation—especially when the student is much smaller—Wang et al. (2020) also uses a teacher assistant: an intermediate-size model that is first distilled from the teacher, then used as the teacher for the final small student, which helps bridge the capacity gap and improves smaller-student results.

2.5 Comparison of Architectural Efficiency

DistilBERT and MiniLM are both task-agnostic Transformer compression methods designed to retain much of BERT's transfer performance while reducing inference cost, but they distill different "signals" and make different architectural commitments. DistilBERT compresses BERT primarily by reducing depth (halving the number of encoder layers) and then performing pre-training-time distillation with a triple objective that combines the masked language modeling loss with a teacher–student distillation loss on outputs and a cosine-distance loss aligning hidden representations, using an initialization that copies every other layer from the teacher. This yields a model that is roughly 40% smaller, 60% faster, and retains roughly 97% of BERT's GLUE performance. By contrast, MiniLM argues that stronger compression can come from distilling the self-attention mechanism itself: it transfers the teacher's last-layer attention distributions (query–key relations) and additionally introduces value-relation transfer, avoiding strict layer-by-layer matching and allowing the student to differ in depth and hidden size without extra projection machinery. In Wang et al. (2020) direct comparisons among similarly sized students, MiniLM reports higher average downstream scores than DistilBERT in their tables (e.g., for a 6-layer/768 student, MiniLM's reported average exceeds DistilBERT's under the same parameter budget), and it further shows gains from a teacher-assistant strategy for very small students—highlighting MiniLM's emphasis on attention-centric transfer and flexible student design versus DistilBERT's emphasis on BERT-like structure plus output/representation matching during pretraining.

III. Dataset and Problem Definition

3.1 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)

3.2 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 used two metrics: Accuracy and Macro-averaged F1 (Macro-F1).

3.3 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 to World, 1 to Sports, 2 to Business and 3 to Sci/Tech. Minimal preprocessing was applied and primary usage was 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.

IV. Methodology

4.1 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.

4.2 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.

4.3 Fine-tuning Setup

Each pretrained model was fine-tuned 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.

4.4 Evaluation Metrics

Two standard metrics for multi-class classification are reported:

Accuracy, measuring the fraction of correctly classified test examples.

$$Accuracy = \frac{Correct\ Classifications}{Total\ Classifications}$$

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.

$$Macro\ F1 = \frac{\sum_{i=1}^{n} (F1\ Score)_i}{n}$$

4.5 Model Comparison Criteria

In addition to predictive performance, the emphasis was on model efficiency, since the accuracy difference between the models is small. The important model comparison metric was parameter count (millions of parameters) as an indicator of model size and training time (minutes) measured for the full fine-tuning run.

4.6 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.

V. Experimental Environment

Experiments were conducted on Google Colab running Linux (Linux-6.1.85+ x86_64, glibc 2.35) with Python 3.10.12. Training and evaluation were implemented in PyTorch 2.3.0+cu121 with GPU acceleration enabled. We used the Hugging Face libraries Transformers 4.41.2 and Datasets 2.19.1. The runtime provided an NVIDIA Tesla T4 GPU with 15 GB VRAM (15102 MiB), using NVIDIA driver 535.104.05 (NVIDIA-SMI CUDA version 12.2).

VI. Results

6.1 Overall Performance on AG News

Two fine-tuned small transformer models were evaluated—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 94.67% Accuracy and 0.9468 Macro-F1, and MiniLM obtained 94.41% Accuracy and 0.9440 Macro-F1. The absolute performance gap is negligible (−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)	Model Size (MB)
distilbert-base-uncased	66.96	0.9467	0.9468	13.44	255.4
microsoft/MiniLM-L12-H384-uncased	33.36	0.9441	0.9440	12.89	127.3

6.2 Model Efficiency and Parameter-focused Comparison

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

6.3 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.

6.4 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 having roughly 50% fewer parameters and safetensors file size, making it a practical choice for resource-constrained deployment scenarios.

References

[1] 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.

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

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

[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.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[9] Geoffrey Hinton, Oriol Vinyals, & Jeff Dean. (2015). Distilling the Knowledge in a Neural Network.

[10] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers

[11] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In Advances in Neural Information Processing Systems (NeurIPS).

Authors

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