TF-IDF vs. DistilBERT for AG News Classification Under Limited Compute*

Yongxi Chen
Dept. of Applied Statistics
University of Michigan
Ann Arbor, MI, USA
cyongxi@umich.edu
Code repository: https://github.com/cyongxi/STATS-507_Final

***Abstract*—This project compares TF-IDF baselines with lightweight head-only DistilBERT models for AG News classification under limited compute. TF-IDF (word+char) achieves the best performance (0.9237 macro-F1), outperforming both DistilBERT variants. Extending DistilBERT's input length and training steps improves results (0.8726 vs. 0.8504 macro-F1) but the frozen encoder remains a limiting factor. Errors concentrate on ambiguous Business vs. Sci/Tech headlines. Overall, TF-IDF is the most effective option under tight computational budgets.**

***Index Terms*—text classification, TF-IDF, DistilBERT, AG News, pretrained models, evaluation, computational constraints**

I. Introduction

Text classification is a standard benchmark for comparing classical models with pretrained Transformers. While models like BERT offer strong accuracy, they are costly to fine-tune. In low-compute settings, it is unclear whether lightweight pretrained models can match simpler methods.

This project compares strong TF-IDF + LinearSVC baselines with head-only DistilBERT models on the AG News dataset. Two TF-IDF variants (word 1–2-gram and word+character) are evaluated against two DistilBERT settings using frozen encoders (128 tokens/200 steps and 192 tokens/1000 steps). The goal is to assess whether lightweight DistilBERT can outperform efficient TF-IDF baselines under strict computational limits.

II. Dataset

A. AG News Corpus

Experiments use the AG News dataset, a widely used benchmark for topic classification. It contains four news categories: World, Sports, Business, and Sci/Tech. The corpus includes 120,000 training samples and 7,600 validation samples. Each example consists of a short news headline and a brief description.

B. Preprocessing

For TF-IDF models, text is lowercased and tokenized using scikit-learn's default tokenizer. For DistilBERT models, preprocessing is handled by the Hugging Face tokenizer with truncation enabled. Two input lengths are evaluated:

- max_length = 128 (baseline setting)
- max_length = 192 (extended context)

No additional text cleaning, stemming, or stopword removal is applied.

III. Methods

A. TF-IDF Baselines

Two classical models are constructed using TF-IDF features and LinearSVC:

- **Word 1–2-gram TF-IDF**: a standard bag-of-words baseline.
- **Word+character TF-IDF**: combines word n-grams with character 3–5-grams, forming a stronger representation for short news headlines.

Both models are trained on the full dataset using default scikit-learn hyperparameters and no additional tuning.

B. DistilBERT Head-Only Models

To evaluate a lightweight pretrained approach, we use DistilBERT, a 6-layer distilled version of BERT that is 40% smaller and significantly faster than the original model [1]. In our setting, the encoder is fully frozen and only the classification head is trained.

- **v1**: max_length = 128 and 200 training steps.
- **v2**: max_length = 192 and 1000 training steps.

Both use the Hugging Face Trainer with batch size 16, linear learning rate scheduling, and default AdamW optimization. Freezing the encoder reduces training time to under one minute for v1 and under one minute for v2 on a single T4 GPU.

C. Evaluation Metrics

Model performance is assessed on the validation set using:

- **Accuracy**, measuring overall correct predictions.
- **Macro-F1**, which averages F1 across all four classes and provides a balanced measure under class imbalance.

Macro-F1 is emphasized in analysis because it reflects both precision and recall across all categories.

IV. Results

A. Overall Performance

Table I summarizes validation accuracy and macro-F1 for all models. Among classical baselines, the word+character TF-IDF model achieves the strongest performance with a macro-F1 of 0.9237. The head-only DistilBERT models perform competitively but remain below the TF-IDF baselines.

TABLE I
Validation Performance of All Models

Model	Accuracy	Macro-F1
TF-IDF (word 1–2-gram)	0.9222	0.9220
TF-IDF (word+char)	**0.9238**	**0.9237**
DistilBERT v1 (128, 200 steps)	0.8518	0.8504
DistilBERT v2 (192, 1000 steps)	0.8729	0.8726

B. DistilBERT Performance under Compute Constraints

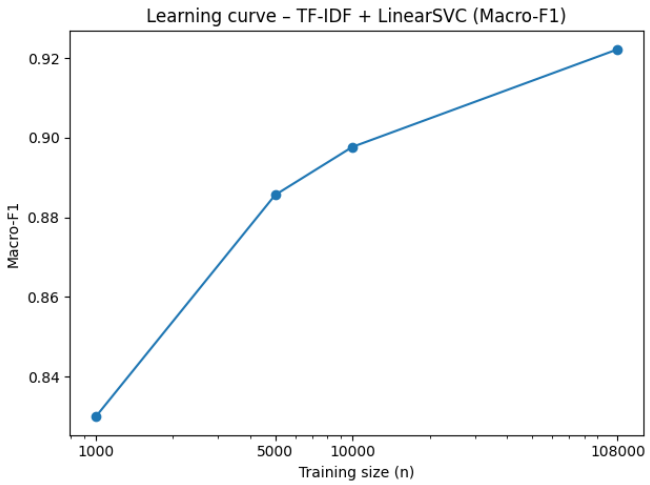

Fig. 1. Learning curve (macro-F1 vs. training size) for TF-IDF + LinearSVC.

Increasing DistilBERT's input length and training steps leads to measurable improvements. Compared with v1, the extended v2 model gains:

- +2.11% accuracy (0.8518 → 0.8729)
- +2.22% macro-F1 (0.8504 → 0.8726)

Despite these gains, both DistilBERT configurations remain below the TF-IDF models, indicating that a frozen encoder limits performance under short training budgets.

C. Confusion Matrix Analysis

The TF-IDF and DistilBERT models display similar error patterns. Both achieve strong performance on *World* and *Sports*, while most misclassifications occur between *Business* and *Sci/Tech*. This reflects real semantic overlap in headlines involving technology companies, markets, and product announcements.

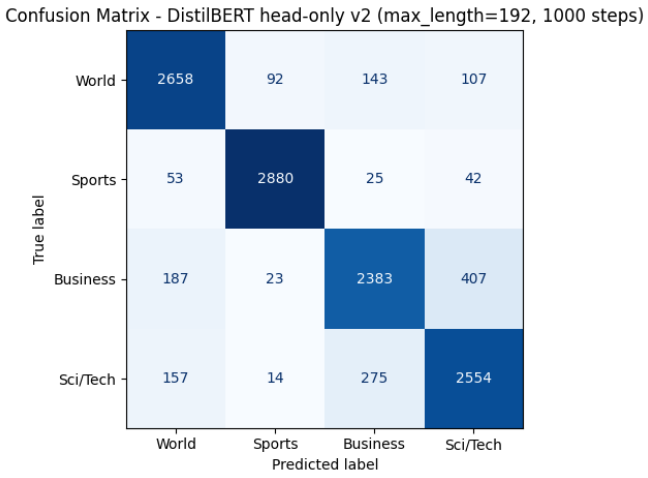

Fig. 2. Confusion matrix for DistilBERT v2 on the validation set.

D. Error Analysis

Manual inspection of misclassified examples shows that many ambiguous headlines contain mixed business and technology cues. Examples include corporate acquisitions of tech companies, regulatory decisions involving major software firms, and news on hardware product transitions. Such cases challenge both TF-IDF and DistilBERT models and help explain the performance ceiling.

V. Discussion

A. Performance Gap Between TF-IDF and DistilBERT

Although pretrained Transformers typically achieve strong performance on text classification tasks, the results in this project show that classical TF-IDF models outperform both head-only DistilBERT variants. Two factors explain this gap.

First, the DistilBERT encoder is fully frozen during training, which prevents the model from adapting its internal representations to the AG News domain. Only the final classification head is updated, severely limiting model capacity. In contrast, TF-IDF + LinearSVC trains all decision parameters directly on the task, enabling fine-grained separation between classes.

Second, even with a longer maximum input length (192 tokens), a portion of AG News samples are truncated. This reduces the benefit of semantic representations learned during pretraining. TF-IDF retains all text information and character-level features, making it more robust to variation in headline length and writing style.

B. Improvements from Increasing Input Length and Training Steps

Comparisons between the two DistilBERT configurations show that both input length and optimization budget significantly affect performance. Increasing the maximum sequence length from 128 to 192 reduces the loss of context during tokenization, while increasing the number of training steps from 200 to 1000 provides the classification head more opportunity to converge. These adjustments improve macro-F1

from 0.8504 (v1) to 0.8726 (v2). However, the frozen encoder remains the primary bottleneck, suggesting that partial fine-tuning of upper Transformer layers would likely yield further gains.

C. Error Patterns: Business vs. Sci/Tech Ambiguity

Error analysis indicates that both TF-IDF and DistilBERT frequently confuse the *Business* and *Sci/Tech* categories. Many headlines about technology companies—such as product announcements, financial reports, and corporate strategic updates—share vocabulary and syntactic structure across the two classes. Such semantic overlap makes these categories inherently difficult even for humans.

The fixed DistilBERT representations further restrict the model's ability to capture subtle contextual distinctions, amplifying these confusions.

D. Implications for Low-Compute NLP Settings

The results highlight an important trade-off relevant to low-resource environments. Classical feature-based pipelines such as TF-IDF + LinearSVC provide competitive performance while remaining computationally lightweight and easy to train. Pretrained Transformer models, when used without fine-tuning, may underperform unless given sufficient input length and optimization steps.

For practitioners working under strict compute budgets (e.g., classroom settings, edge devices, or rapid prototyping), TF-IDF remains a strong and efficient baseline. Fully unlocking the benefits of Transformer architectures requires either partial fine-tuning or more extensive compute resources.

VI. Conclusion

This project compared strong TF-IDF baselines with lightweight head-only DistilBERT models for AG News topic classification under strict computational constraints. The results show that the word+character TF-IDF model achieves the highest performance (0.9237 macro-F1), outperforming both DistilBERT configurations. Increasing DistilBERT's input length and training steps improves its performance (0.8726 vs. 0.8504 macro-F1), but the frozen encoder limits the model's ability to fully adapt to the task.

Error analysis indicates that most misclassifications occur between the Business and Sci/Tech categories, reflecting intrinsic semantic overlap in news headlines. These findings highlight that classical linear models remain strong, efficient, and highly competitive in low-compute settings. To close the performance gap, future work may explore partial fine-tuning of upper Transformer layers, domain-adaptive pretraining, or hybrid architectures that combine TF-IDF signals with transformer embeddings.

References

[1] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019.

[2] HuggingFace, "Transformers — Hugging Face Documentation," Available: https://huggingface.co/docs/transformers/en/index#transformers