Evaluating Lightweight Transformers for Low-Resource Document Classification

A. Researcher, B. Scholar, and C. Analyst

Abstract

We study whether lightweight transformer architectures can match the document-classification accuracy of larger pretrained models when only a few hundred labeled examples are available per class. Across three low-resource corpora, a distilled six-layer encoder reaches within two points of a twelve-layer baseline while using roughly a third of the inference compute. We release our evaluation harness and discuss where the smaller model's errors concentrate.

1. Introduction

Document classification in specialized domains — legal filings, clinical notes, regulatory filings — often has to work with far fewer labeled examples than the benchmarks large language models are tuned against. Practitioners in these settings frequently cannot afford the inference cost of a full-size transformer in a production pipeline, which motivates a direct comparison between compact and large architectures under realistic label budgets.

This paper makes three contributions. First, we construct three low-resource classification corpora with carefully controlled label counts. Second, we benchmark a distilled six-layer encoder against a standard twelve-layer baseline on each corpus. Third, we analyze the error cases where the smaller model diverges from the larger one, and report in Figure 1 how that gap narrows as label count increases.

2. Related Work

Prior work on model distillation has largely evaluated accuracy retention on high-resource benchmarks with tens of thousands of labeled examples per class. Separately, low-resource classification research has focused on data augmentation and few-shot prompting rather than on architecture size directly. We are not aware of prior work that holds the label budget fixed while varying model size specifically to characterize this tradeoff.

3. Methodology

We built three corpora of 2,400 documents each, drawn from public filings in three domains, and held out label counts of 50, 150, and 400 examples per class for training, with the remainder used for evaluation. Both models were fine-tuned with identical hyperparameters, early stopping on a held-out validation split, and five random seeds per configuration. Table 1 summarizes the corpus statistics used in every run.

Figure 1 shows the overall experimental pipeline: corpus construction, label-budget sampling, fine-tuning under five seeds, and held-out evaluation, repeated independently for each model size.

4. Results

At the 400-example label budget, the distilled six-layer encoder reaches within two F1 points of the twelve-layer baseline on all three corpora. The gap widens to roughly six points at the 50-example budget, where the larger model's pretraining appears to matter more. Figure 2 plots F1 score against label count for both model sizes across all three corpora.

5. Discussion

The narrowing gap at larger label budgets suggests that, once enough in-domain supervision is available, the smaller model's reduced capacity stops being the binding constraint. Error analysis shows the six-layer model's mistakes concentrate on documents with boilerplate structure dissimilar to the fine-tuning examples, rather than on any single class.

6. Conclusion

A distilled six-layer encoder is a reasonable default for production document classification once a few hundred labeled examples per class are available, trading a small accuracy gap for a substantial reduction in inference compute. Future work should test whether this holds under distribution shift between training and deployment documents.

Figure 1: Experimental pipeline, from corpus construction through held-out evaluation.

Table 1: Corpus statistics for the three low-resource domains used in this study.

Figure 2: F1 score versus label count per class, compared across the six-layer and twelve-layer models on all three corpora.

References

[1] T. Author and R. Coauthor. Knowledge distillation for efficient text classification. Proceedings of the Workshop on Efficient NLP, 2023.

[2] M. Ling and S. Park. Few-shot document classification in specialized domains. Transactions of the Association for Computational Studies, 2022.

[3] J. Ortiz. Compact transformer architectures: a survey. ACM Computing Surveys, 2021.

[4] H. Novak and P. Singh. Label-efficient fine-tuning for clinical text classification. Journal of Applied Language Processing, 2023.

[5] D. Kim, A. Ferreira, and L. Costa. Measuring inference cost tradeoffs in production NLP systems. Proceedings of Systems for ML, 2022.

[6] E. Brandt. Data augmentation strategies for low-resource text classification. Computational Linguistics Review, 2020.

[7] N. Osei and V. Petrov. Benchmarking encoder size against label budget. Proceedings of the Conference on Empirical Methods, 2024.