Dean&Francis

ISSN 2959-6157

An Empirical Comparison of BERT and Lightweight Variants for IMDb Sentiment Classification

Xuanyu Wu[1, *]

[1]School of Economics and Management, Beijing Jiaotong University, Weihai, 264400, China
*Corresponding author: 23711071@bjtu.edu.cn

Abstract:

Pre-trained transformer models are now the most common choices for sentiment analysis and other text classification tasks. However, their large number of parameters and high inference cost make it hard to deploy them in limited resource settings. To solve this problem, several lightweight variants have been introduced, this paper presents a controlled empirical study of four transformer-based models - BERT-base, DistilBERT, TinyBERT, and ALBERT-base - on the IMDb movie review sentiment classification task. All models are fine-tuned under the same conditions. Besides standard metrics, total and per-sample inference time were measured as well. To better understand stability, the test set is divided into short and long review subsets based on word count and compare model performance across subsets. Results show an accuracy-efficiency trade-off. BERT-base achieves the highest test accuracy (86.3%), followed by ALBERT (85.7%), while DistilBERT is lower (85.2%) but offers about 2 times faster inference. TinyBERT is the fastest but has lower accuracy (81.3%). Across these models, performance on short reviews is higher than on long reviews. The performance drop from short to long texts is more severe for the smallest model, TinyBERT, finding that highly compressed models struggle more with long and complex texts. This paper discusses the implications of findings for selecting sentiment analysis models under different application requirements.

Keywords: Sentiment analysis; Transformer-based models; Model compression; IMDb movie reviews.

1. Introduction

Online platforms such as websites, forums, and social media gathered large amounts of user-generated content containing opinions, attitudes, and emotions. Identifying the sentiment expressed in such text by computer - positive, negative, or neutral - has become a key task in natural language processing (NLP). This task, known as sentiment analysis or opinion mining, has been applied to product reviews, movie reviews, political discussions, and many other domains [1, 2].

Early work on sentiment analysis treated it as a text classification problem, documents were represented with bag-of-words or n-gram features, and standard classifiers such as Naive Bayes, maximum entropy, or support vector machines (SVMs) were trained to distinguish positive and negative documents [1-3]. These methods achieved strong performance on datasets such as IMDb movie reviews and helped define the basic task settings and evaluation metrics for later research [3].

In recent years, large pre-trained language models have changed the way doing sentiment analysis and many other NLP tasks a lot. BERT is a bidirectional transformer model pre-trained on large unlabeled corpora with masked language modeling and next sentence prediction objectives and fine-tuned on downstream tasks with relatively small and labeled datasets [4]. Fine-tuned BERT models generally perform better than previous methods based on static word embeddings and convolutional architectures [1, 2, 4]. However, BERT-base contains around 110 million parameters, leading to high memory usage and slow inference. To make pre-trained models more practical in real life deployments, several lighter variants have been created. DistilBERT uses knowledge distillation to compress BERT into a smaller student model [5]. TinyBERT introduces a multi-stage distillation framework that aligns not only the logits but also hidden states and attention maps between teacher and student [6]. ALBERT reduces parameters through embedding factorization and cross-layer parameter sharing, allowing deeper models with fewer parameters [7]. These models aim to maintain most of BERT's accuracy while being smaller and faster.

Despite the popularity of these models, there is still limited systematic comparison of BERT-base and its lightweight variants on the same task under strictly controlled conditions. In particular, the following questions are the main focus of this paper.

Firstly, it examines how classification accuracy changes when BERT-base is replaced by DistilBERT, TinyBERT or ALBERT for document-level sentiment classification on the IMDb dataset. Secondly, it compares these models in terms of inference efficiency on the full 25,000-review test set, considering both total and per-sample time that used. Thirdly, it analyzes whether input length affects the models in different ways, with attention to whether smaller, heavily compressed models degrade more on long and complex reviews than on shorter texts.

To address these questions, this paper conducts a controlled empirical comparison of BERT-base, DistilBERT, TinyBERT, and ALBERT-base on the IMDb movie review dataset. All models are fine-tuned on the same reduced training set and evaluated with the same preprocessing, hyper-parameters, and hardware. The paper measures performance metrics and inference time, and the paper further divides the test set into short and long review subsets to analyze the influence of text length.

This study makes three main contributions. Firstly, it provides a fair comparison of four commonly used transformer-based models on the same sentiment analysis task, using unified data splits and training settings. Secondly, the results show that DistilBERT offers a reasonable balance between accuracy and efficiency, while TinyBERT is much faster but shows a clear drop in performance. Finally, the experiments show that all models perform much better on short reviews than on long ones, and that the smallest model, TinyBERT, suffers the largest performance gap on long texts, which points to a limitation of heavy compression for document-level sentiment analysis.

2. Related Work

2.1 Sentiment analysis and opinion mining

Pang and Lee provided one of the most important surveys of sentiment analysis and opinion mining, summarizing methods for document-level, sentence-level, and aspect-level sentiment classification [1]. They emphasize that sentiment analysis is not just topic classification, because it focuses on the evaluative nature of text and the subjective orientation of the author. Liu offers a monograph that further formalizes the field in terms of opinion targets, opinion holders, and sentiment orientation, and discusses rule-based approaches and supervised machine learning methods [2]. These works also point out that longer opinionated texts can be difficult, because they may contain mixed sentiments and complex discourse structures [1, 2].

A classic empirical study by Pang et al. compares several machine learning methods, including Naive Bayes, maximum entropy, and SVMs, on movie review sentiment classification [3]. Documents are represented with bag-of-words, presence/absence features, and n-grams. The study shows that SVMs with appropriate features can significantly perform better than simple baselines and that tells

ISSN 2959-6157

feature design plays an important role in sentiment classification performance [3]. This paper also set a standard for clearly defined test splits and evaluation metrics, which later work has frequently followed.

2.2 Pre-trained transformer models and lightweight variants

Smairi et al. shows that on the IMDb dataset, the BERT fine-tuned model performed better than the traditional machine learning models [8]. BERT uses a multi-layer bidirectional transformer encoder trained on large unlabeled corpora, and fine-tuning on downstream tasks is done by adding a small classification layer on top of the [CLS] token representation [4]. For sentiment analysis, fine-tuned BERT models have already shown a strong improvements over earlier methods that rely on static word embeddings and recurrent networks [1, 2, 4].

However, the large size of BERT-base makes it hard for applications with limited GPU or CPU resources, to realize on mobile devices. As a result, a series of compression and efficiency-focused models have been introduced, and the method of model compression is recognized [9].

DistilBERT is a distilled version of BERT that reduces the number of layers and parameters while trying to preserve most of BERT's language understanding ability through a distillation objective [5]. The authors report that DistilBERT maintains about 97% of BERT's performance on language understanding benchmarks while being 40% smaller and 60% faster at inference [5].

TinyBERT offers a more detailed knowledge distillation framework. It includes general distillation at the pre-training stage and task-specific distillation at the fine-tuning stage, and it distills knowledge at multiple levels, such as logits, hidden states, and attention matrices [6]. This allows compact models to perform competitively on a variety of natural language understanding tasks [6].

ALBERT takes a different method which is reducing parameters through two main techniques, factorizing the embedding matrix to decouple the size of the hidden layers from the vocabulary embedding dimension and applying cross-layer parameter sharing in the transformer encoder [7]. With these methods, ALBERT greatly reduces the number of parameters compared to standard BERT, while still achieving strong performance on benchmarks such as GLUE and question answering [7].

Although these models have been evaluated on different benchmarks, most previous studies report results either under different training settings or focus on a single model at a time [4-7]. It is difficult to directly answer practical questions such as "How much accuracy do I lose if I switch from BERT-base to DistilBERT or TinyBERT on a real task?" or "Does ALBERT behave differently from DistilBERT on long reviews?". This paper aims to fill this area by comparing BERT-base, DistilBERT, TinyBERT, and ALBERT under the same training regime on the same dataset, with additional analysis of text length.

3. Methodology

3.1 Task definition

This paper discusses document-level sentiment classification on the IMDb movie review dataset. Each review is labeled as either positive or negative. The task is to predict the correct label for each review based on its text content. This setting corresponds to the standard document-level sentiment analysis scenario described in earlier work [1-3].

3.2 Dataset and splits

The IMDb dataset contains 25,000 labeled training reviews and 25,000 labeled test reviews, with a balanced class distribution. To keep training time manageable and to simulate a moderate-data setting, the paper creates a limited training setup as follows.

From the original 25,000 training reviews, the paper randomly samples 4,000 reviews to form the training set. Then selecting 1,000 samples from this set as a validation set used for model selection and early stopping. The original 25,000-review test split is kept unchanged and used as the final test set for all models. All four models are trained and evaluated on exactly the same training, validation, and test partitions.

3.3 Models

As shown in table 1, the paper considers the following four transformer-based models.

Table 1. Introduction of four models.

BERT-base	The standard 12-layer transformer encoder with hidden size 768 and 12 attention heads, commonly used as a baseline in many NLP tasks [4].
DistilBERT	A distilled version of BERT that reduces the number of layers and parameters using knowledge distillation from the BERT teacher model [5].

TinyBERT	A compact model obtained through a multi-stage distillation procedure that transfers knowledge from BERT at both pre-training and task-specific stages [6].
ALBERT-base	A "lite" BERT variant that reduces parameters via embedding factorization and weight sharing across layers [7].

For each model, the paper use the corresponding tokenizer and the standard sequence classification head provided in the Hugging Face Transformers library. The classification head is a linear layer on top of the [CLS] token representation that outputs logits for the two sentiment classes.

3.4 Preprocessing and tokenization

Tokenization is done with the model-specific tokenizer. The maximum sequence length is set to 128 tokens. Sequences longer than 128 tokens are truncated; sequences shorter than 128 tokens are padded to 128 tokens. For each sample the research keeps input_ids, attention_mask, and the sentiment label. These settings are kept identical for all four models to ensure a fair comparison.

3.5 Training setup

As shown in table 2, all models are fine-tuned under the same training configuration.

Table 2. Training elements.

Training samples	4,000 reviews (with 1,000 held out as validation).
Number of epochs	2.
Optimization	AdamW optimizer with a linear learning rate schedule.
Initial learning rate	2e-5.
Loss function	cross-entropy loss on the two sentiment classes.
Other hyper-parameters	(such as batch size and weight decay) They are chosen once and then fixed for all models.

During training, this paper monitor validation accuracy and save the checkpoint with the best validation performance. This checkpoint is then used for all test-set evaluations.

3.6 Length-based subsets: short vs long reviews

To analyze the effect of review length, the paper computes the word count of each IMDb test review and use the empirical 25th and 75th percentiles to define short and long subsets. The 25th percentile is about 133 words and the 75th percentile is about 245 words. Based on this, the short reviews are those with word count ≤ 133 (7,601 samples). Long reviews are those with word count ≥ 245 (7,501 samples).

The paper evaluates each model on the full test set, the short-review subset, and the long-review subset. For each evaluation, the paper also measure total inference time and compute the average time per sample.

3.7 Evaluation metrics

The paper reports the following metrics. Accuracy (Acc), weighted precision, recall, and F1-score, total inference time on the corresponding test set, average inference time per sample (table 3). All metrics are computed in a consistent way across models and subsets.

Table 3. Metrics definition.

Accuracy (Acc)	The proportion of the test samples that are right classified.
Weighted precision	How reliable the model's predictions are, averaging precision over classes while weighting each class by its support.
Weighted recall	How many samples are successfully judged by the model
F1-score	F1-score is the weighted harmonic mean of precision and recall
Total inference time	Total time that the model used to finish all test set.
Average inference time per sample	Time per sample.

ISSN 2959-6157

4. Experiments and Results

4.1 Overall performance on the full test set

Table 4 summarizes the performance of the four models on the full IMDb test set (25,000 reviews) and the inference time.

Table 4. Performance on the full IMDb test set (25,000 reviews).

Model	Test Accuracy	Weighted F1	Total Inference Time (s)	Time per Sample (s)
BERT-base	0.8634	0.8633	885.84	0.0354
ALBERT-base	0.8567	0.8567	1143.60	0.0457
DistilBERT	0.8524	0.8524	434.27	0.0174
TinyBERT	0.8129	0.8128	93.32	0.0037

BERT-base achieves the highest accuracy and F1, as expected for the largest model.

ALBERT-base is slightly behind BERT in accuracy but is slower than BERT in our hardware setting, likely due to implementation and hardware characteristics rather than parameter count alone.

DistilBERT reaches only about 1 percentage point lower accuracy and F1 than BERT but has roughly half the per-sample inference time, which is a great trade-off between performance and efficiency.

TinyBERT is the fastest model in these four, with per-sample inference time around 0.0037 seconds, nearly 9-10 times faster than BERT. However, this comes at the cost of about 5 percentage points lower accuracy and F1.

These results clearly show that smaller models can substantially reduce inference time while keeping accuracy relatively high, but extreme compression (TinyBERT) leads to a noticeable and obvious performance drop on this document-level task.

4.2 Performance on short vs long reviews

To understand how input length affects different models, the paper evaluates each model on the short (≤ 133 words) and long (≥ 245 words) subsets of the test set. Table 5 reports the accuracy and weighted F1 for each subset.

Table 5. Accuracy and F1 on short and long review subsets.

Model	Subset	#Samples	Accuracy	Weighted F1
BERT-base	Short	7,601	0.914	0.914
	Long	7,501	0.797	0.797
ALBERT-base	Short	7,601	0.913	0.912
	Long	7,501	0.785	0.785
DistilBERT	Short	7,601	0.905	0.905
	Long	7,501	0.785	0.785
TinyBERT	Short	7,601	0.875	0.875
	Long	7,501	0.738	0.738

All models perform better on short reviews than on long reviews. The accuracy gap between short and long subsets is roughly 0.11-0.14 for all four models. This confirms that long, multi-sentence reviews are more difficult for sentiment classification, might because they often contain mixed or shifting opinions and more complex discourse structure. The smallest model, TinyBERT, suffers the largest performance drop. While BERT-base and ALBERT-base achieve around 0.91 accuracy on short reviews and around 0.78-0.80 on long reviews, TinyBERT drops from about 0.88 to about 0.74. DistilBERT lies between these extremes but still shows a clear drop.

For short texts, lightweight models are closer to BERT. On short reviews, DistilBERT's accuracy of about 0.905 is quite close to BERT's 0.914, and even TinyBERT's 0.875 may be acceptable in applications where latency is more important than a few percentage points of accuracy. On long reviews, however, the gap between TinyBERT and BERT-base grows larger (around 6 percentage points). These findings suggest that compression hurts performance more on long documents than on short ones, at least for this task. Smaller models have less ability to

capture long-distance dependencies and to integrate information across different sentences, so they find long texts particularly difficult.

5. Discussion

5.1 Accuracy-efficiency trade-offs

The results on the full test set and the length-based subsets clearly illustrate the trade-off between model accuracy and inference efficiency. BERT-base and ALBERT-base provide the best average accuracy but are the slowest models in our experiments. DistilBERT achieves a very competitive accuracy (only about 1 percentage point below BERT-base) while cutting inference time nearly in half, which makes it a strong competitor for applications that require both good performance and fine latency. TinyBERT delivers the highest speed, but its noticeable performance drop - especially on long reviews - means that it is more suited for scenarios where high efficiency is required and some loss in accuracy is acceptable. The reports of Zhou et al. and Fatihah & Ayudhia show TinyBERT is useful in specific cases [10, 11].

These findings are consistent with the design goals of the models. DistilBERT and TinyBERT balance performance and compression through knowledge distillation [5, 6], while ALBERT reduces parameter count through factorization and sharing [7]. Our experiments show how this design choices translate into concrete performance and runtime differences on a standard sentiment analysis dataset.

5.2 The challenge of long reviews

The gap between short and long review performance highlights a key challenge for document-level sentiment analysis. Longer reviews typically include multiple perspectives of the movie, present positive and negative comments and may change sentiment over the course of the text.

This makes it harder to assign a single overall label, even for real readers, and increases the difficulty for automatic models [1, 2]. When the paper additionally truncates reviews to a fixed maximum length (128 tokens), important information near the end of long reviews may be cut off, which can further hurt performance.

The fact that the smallest model (TinyBERT) has the largest drop from short to long reviews suggests that compressed models may have less ability to represent complex sentiment patterns and longer-range dependencies. For applications where long reviews are common, such as in-depth product feedback or forum posts, using a more capable model like BERT-base or ALBERT may be necessary, or alternative architectures that explicitly handle long sequences could be explored in the future.

5.3 Practical implications

From a practical perspective, our results support the following guidelines.

Mobile or real-time short-text applications can often use DistilBERT or even TinyBERT, especially if very low latency is required and a small drop in accuracy is acceptable. Gandhi and Sharma have treated DistilBERT as an "efficient sentiment analysis model" [12]. Batch processing of longer reviews or documents, where accuracy is more important than real-time performance, may benefit from BERT-base or ALBERT-base, even though they are slower. Mixed scenarios could start with DistilBERT as a default choice, because it offers a good balance between accuracy and efficiency in our experiments.

These recommendations are specific to the IMDb task and settings used here, but the general pattern that smaller models are more suitable for short texts and large models handle complex long texts better is likely to hold for many sentiment analysis applications.

6. Conclusion

This paper presented a controlled empirical comparison of four transformer-based models on the IMDb movie review sentiment classification task. All models were fine-tuned under identical training conditions on the same reduced training set and evaluated on the full test set and length-based subsets.

BERT-base achieves the highest accuracy (86.3%), with ALBERT-base slightly behind (85.7%), while DistilBERT reaches 85.2% with faster inference. TinyBERT is much faster than all other models but shows a notable performance drop (81.3% accuracy). All models perform significantly and obviously better on short reviews than on long reviews, confirming that long opinionated documents are more difficult for sentiment classification. The smallest model, TinyBERT, experiences the largest performance gap from short to long reviews, suggesting that heavy compression can make models less robust to long and complex texts.

In future work, it would be valuable to extend this study in several ways. First, experiments could be repeated on additional datasets from different domains and languages to test the generality of the findings. Second, other lightweight models such as MobileBERT or MiniLM could be included in the comparison. Third, alternative architectures specifically designed for long sequences, or hierarchical models that process reviews paragraph

ISSN 2959-6157

by paragraph, could be investigated. Finally, combining model compression techniques with data augmentation or semi-supervised learning may also further improve the accuracy of lightweight models without losing their efficiency advantages.

Although this study provides useful evidence, there are still some limitations. It uses just one English movie review dataset with reduced training set, so the results may not fully generalize to other tasks or settings, and evaluation mainly focuses on accuracy and inference time rather than other factors such as robustness or memory usage.

References

[1] Pang B and Lee L, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, vol. 2, no. 1–2, pp. 1–135, 2008.

[2] Liu B, Sentiment Analysis and Opinion Mining. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.

[3] Pang B, Lee L, and Vaithyanathan S, Thumbs up? Sentiment classification using machine learning techniques, in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79–86.

[4] Devlin J, Chang M W, Lee K, and Toutanova K, BERT: Pre-training of deep bidirectional transformers for language understanding, in Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 4171–4186.

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

[6] Jiao X et al., TinyBERT: Distilling BERT for natural language understanding, arXiv preprint arXiv:1909.10351, 2020.

[7] Lan Z et al., ALBERT: A lite BERT for self-supervised learning of language representations, in Proc. Int. Conf. Learning Representations (ICLR), 2020.

[8] Smairi R, Belguith M, & Salah A B. Fine-tune BERT based on machine learning models for sentiment analysis of IMDb movie reviews. Procedia Computer Science, 2024, 246, 2390–2399.

[9] Gong X, Ying W, Zhong S, and Gong S, Text sentiment analysis based on transformer and augmentation, Frontiers in Psychology, vol. 13, article 906061, 2022.

[10] Zhou P, Zhang Y, and Zhao J, A cross-language attribute-level sentiment analysis approach using TinyBERT and GCN, International Journal of Knowledge Management, vol. 20, no. 1, pp. 1–23, 2024.

[11] Fatihah H A and Ayudhia Z P, TinyBERT-based sentiment analysis for large-scale educational feedback: A case study on Coursera reviews, in Proc. 8th Int. Conf. on Smart Applications, Communications and Networking (SmartNets), 2025.

[12] Gandhi J N and Sharma P, Efficient sentiment classification using DistilBERT for social media data, in Proc. Int. Conf. on Sustainable Innovation in Computing and Engineering (ICSICE-24), 2025.