Sentiment-Driven Movie Review Classification Using DistilBERT

Kaiyang Luo
Department of Statistics
University of Michigan
Email: kaiyangl@umich.edu

***Abstract*—Sentiment analysis on movie reviews is a fundamental natural language processing (NLP) task with wide applications in opinion mining and recommendation systems. This project investigates sentiment classification using both classical machine-learning models and modern Transformer-based architectures. Specifically, we compare a TF–IDF + Logistic Regression baseline model with a fine-tuned DistilBERT model on the IMDB dataset. Experimental results show that DistilBERT significantly outperforms the baseline in accuracy and F1-score, achieving 0.909 F1 on the IMDB test set. We further conduct error analysis to identify common failure patterns such as sarcasm, contrastive sentiment, and the truncation of long reviews. The study demonstrates both the effectiveness and the limitations of lightweight Transformer models for real-world sentiment classification tasks.**

I. Introduction

Sentiment analysis is one of the most widely studied text-classification tasks, enabling automated systems to understand opinions expressed in natural language. Classical approaches typically rely on TF–IDF or bag-of-words representations combined with machine-learning classifiers. While effective for simpler text, these models struggle with semantic dependencies and contextual meaning.

Recent advances in NLP have been dominated by Transformer-based architectures such as BERT [1] and DistilBERT [2]. These models leverage self-attention and large-scale pretraining to capture linguistic structures far beyond the capacity of traditional methods. The goal of this project is to implement a complete sentiment-classification pipeline using Hugging Face tools and compare a classical baseline with a fine-tuned DistilBERT system.

The primary contributions of this work include: (1) building a reproducible end-to-end pipeline for sentiment classification using Python and Hugging Face, (2) demonstrating the performance improvements of Transformer finetuning over classical ML models on IMDB, and (3) analyzing the types of errors and limitations inherent in the DistilBERT model.

II. Method

A. Dataset

The IMDB dataset consists of 50,000 movie reviews labeled as positive or negative, with a perfectly balanced class distribution. Following common practice, we used the official test set (25,000 reviews) for final evaluation. The remaining 25,000 reviews were split into 90% training and 10% validation. Reviews vary widely in length, with an average of 234 words and a maximum exceeding 2,400 words.

B. Baseline Model: TF–IDF + Logistic Regression

For the classical machine-learning baseline, text was transformed using TF–IDF features with up to 50,000 n-grams (1–2 grams) and English stop-word removal. A Logistic Regression classifier with $C = 1.0$ and `max_iter=200` was trained on the resulting sparse vectors. This baseline provides a strong linear reference model that is fast to train and easy to interpret, but it does not model contextual interactions between words.

C. Transformer Model: DistilBERT Fine-Tuning

We fine-tuned the `distilbert-base-uncased` model using the Hugging Face `Trainer` API. Each review was tokenized with truncation to 216 tokens and dynamic padding. Training was conducted for 2 epochs with a per-device batch size of 16, learning rate of 2×10^{-5}, and weight decay of 0.01. Mixed-precision (fp16) training was used on a single GPU. Validation F1-score was used as the model-selection metric, and the best checkpoint was loaded at the end of training.

D. Evaluation Metrics

We report accuracy, precision, recall, and F1-score for both baseline and fine-tuned models. In addition, we plot confusion matrices and training-loss curves to visualize performance and training dynamics.

III. Results

A. Quantitative Performance

Table I summarizes model performance on the IMDB test set.

TABLE I
Model performance on the IMDB test set.

Model	Acc.	Prec.	Rec.	F1
TF–IDF + LR	0.881	0.88	0.88	0.88
DistilBERT (fine-tuned)	**0.909**	**0.904**	**0.914**	**0.909**

The fine-tuned DistilBERT model improves test accuracy by 2.8 percentage points over the classical baseline and yields a higher F1-score on both classes, demonstrating clear gains from contextualized representations.

B. Confusion Matrix

Fig. 1 shows the confusion matrix for DistilBERT on the 25,000-sample test set. The model correctly classifies 11,294 negative and 11,421 positive reviews, with 1,206 false positives and 1,079 false negatives. Errors are roughly balanced across sentiment classes, with a slight tendency toward false positives.

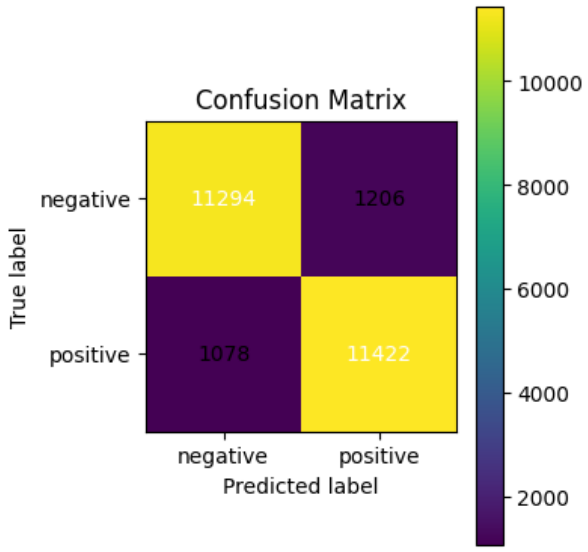

Fig. 1. Confusion matrix of DistilBERT on the IMDB test set.

C. Training Curve

Fig. 2 shows the training-loss trajectory over time. Loss decreases rapidly during the first epoch and continues to improve more slowly in the second epoch, suggesting that the model has not yet fully overfit the training data.

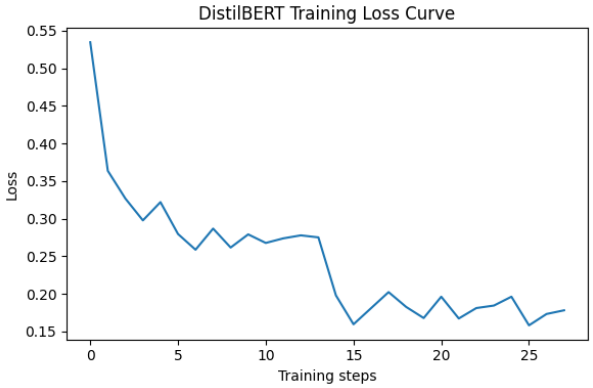

Fig. 2. Training loss curve during DistilBERT fine-tuning.

D. Error Analysis

To better understand the limitations of the fine-tuned DistilBERT model, we inspected the top 20 misclassified test examples with the highest prediction confidence. All examined cases corresponded to false positives, where the model predicted positive sentiment for reviews labeled as negative. This indicates that DistilBERT is overly sensitive to surface-level positive lexical cues.

First, many misclassified reviews contained strongly positive expressions (e.g., *"one of the best Kung fu movies"*, *"pure genius"*, *"I love this show"*) despite the overall sentiment being negative. Because inputs were truncated to 216 tokens, negative content frequently appeared near the end of long reviews and was removed during tokenization. As a result, the model primarily observed positive opening sentences and produced high-confidence positive predictions.

Second, the model struggled with sarcasm and rhetorical exaggeration. Examples such as *"Master P's acting skills make you actually believe he is Italian!"* and *"best movie ever... what poetry!"* were classified as positive even though the reviewer clearly intended a negative evaluation. DistilBERT lacks the pragmatic understanding required to differentiate literal praise from sarcastic criticism.

Third, several errors arose from reviews blending production-related commentary with mild praise, where emotional polarity is subtle or context-dependent. In such cases, isolated positive phrases outweighed the overall negative tone in the model's representation. These findings highlight the model's reliance on local lexical polarity, its limited ability to capture discourse-level sentiment, and the significant impact of input truncation on classification performance.

IV. Conclusion

This project demonstrates that fine-tuning DistilBERT significantly improves sentiment-classification performance compared to a strong TF–IDF + Logistic Regression baseline on the IMDB dataset. While the model performs well overall, detailed error analysis reveals weaknesses related to sarcasm, mixed sentiment, and the truncation of long reviews. Future work may incorporate models with longer context windows (e.g., Longformer), sarcasm-aware training data, or hierarchical document encoders to better capture document-level sentiment.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019.

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