Sentiment Analysis Using DistilBERT on IMDb Dataset

ABSTRACT:

This paper investigates the effectiveness of DistilBERT, a compact variant of the BERT architecture, for sentiment analysis on the IMDb movie review dataset. Given the computational constraints of working with a large model, we leverage DistilBERT's efficiency to classify reviews as positive or negative with limited training resources. Our approach includes tokenization, model training, and evaluation using a subset of the IMDb dataset, while key performance metrics—accuracy, F1 score, and ROC-AUC—are reported to assess model efficacy. Experimental results indicate moderate performance, with the model achieving an accuracy of 59.4% and an AUC-ROC of 0.6032. Analysis of misclassified reviews highlights areas for improvement, particularly in handling nuanced or lengthy texts. This study demonstrates DistilBERT's potential for sentiment classification tasks under constrained computational environments and suggests further enhancements for improved sentiment polarity detection.

I. INTRODUCTION:

Sentiment analysis, the computational task of identifying and categorizing opinions expressed in text, has become an essential tool for various applications in natural language processing (NLP). From understanding customer feedback to gauging public opinion on social media, sentiment analysis provides actionable insights by categorizing text as expressing positive, negative, or neutral sentiment. With the emergence of powerful transformer models, significant advancements have been made in sentiment classification tasks; however, the large model sizes often present computational challenges, especially in resource-constrained environments.

This report explores the implementation of a sentiment analysis model using DistilBERT, a lightweight version of the BERT (Bidirectional Encoder Representations from Transformers) model, which retains much of BERT's accuracy while being more computationally efficient. DistilBERT's design reduces the number of parameters and training time, making it a practical choice for applications where computational resources are limited.

Our study focuses on applying DistilBERT to the IMDb movie review dataset, a benchmark dataset commonly used for sentiment classification research. Given that the full dataset consists of 50,000 reviews, we used a subset to expedite training and reduce computational demands. This subset-based approach not only demonstrates the feasibility of deploying transformer models in smaller environments but also highlights how well DistilBERT can perform sentiment classification when trained on a limited dataset.

The objectives of this report are as follows:

To assess the sentiment classification capabilities of DistilBERT on the IMDb dataset.

To analyze model performance through key metrics such as accuracy, F1 score, and ROC-AUC.

To identify areas for improvement by examining common patterns in misclassified instances.

The structure of this report includes a discussion of the methodology and data preprocessing steps, followed by results that include performance metrics and error analysis. Through this study, we aim to provide insights into the effectiveness of DistilBERT in resource-constrained environments and explore potential enhancements to improve sentiment classification accuracy.

II. BACKGROUND:

Sentiment analysis has seen tremendous growth due to advancements in deep learning and the availability of extensive text datasets. Traditionally, sentiment classification relied on feature-based methods, such as Support Vector Machines (SVM) and Naive Bayes classifiers, using bag-of-words or TF-IDF (Term Frequency-Inverse Document Frequency) features. However, these models often struggle to capture the nuances of language, such as context and word order, limiting their effectiveness in complex text analysis tasks.

With the introduction of deep learning architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), significant improvements were observed. These models could capture sequential information and contextual meaning, offering better accuracy for sentiment classification. However, their limitations in handling long-range dependencies and contextual relationships paved the way for transformer-based models, which have since become the standard in NLP tasks.

Transformer models like BERT (Bidirectional Encoder Representations from Transformers) leverage self-attention mechanisms to capture dependencies over long sequences, allowing for a bidirectional understanding of language. BERT, however, is a large model with considerable computational demands. This limitation sparked interest in developing smaller, more efficient transformer models. DistilBERT, a distilled version of BERT, was introduced to address this challenge. It achieves a 40% reduction in model size and a 60% increase in speed, while retaining 97% of BERT's performance on certain NLP benchmarks. This reduction is achieved through a knowledge distillation process that transfers the knowledge from a larger model (BERT) to a smaller model (DistilBERT), making it a suitable choice for tasks with limited resources.

The IMDb dataset, a popular benchmark for sentiment analysis, consists of movie reviews labeled as positive or negative. This dataset is widely used to evaluate the performance of sentiment classifiers due to its linguistic variety and the presence of both straightforward and subtle sentiments. Previous work has demonstrated strong performance on the IMDb dataset using larger models like BERT and GPT. However, applying DistilBERT, especially with a limited subset of data, presents a unique challenge and allows us to investigate how well a distilled model can perform under resource constraints.

In this report, we aim to assess DistilBERT's effectiveness on the IMDb sentiment classification task using a reduced dataset. We will analyze the model's performance, report on metrics such as accuracy and F1 score, and identify areas for potential improvement. This exploration will provide insights into the trade-offs between model size and performance and offer practical considerations for deploying transformer-based models in environments where computational efficiency is critical.

III. METHODOLOGY

Dataset and Preprocessing:

IMDb Dataset Overview:

The IMDb dataset, commonly used for sentiment analysis, consists of 50,000 labeled reviews. These reviews are split into two classes: positive and negative sentiments, each comprising 25,000 reviews. The dataset is widely recognized for its balanced distribution and extensive vocabulary, which tests models' ability to understand nuanced language in movie reviews.

For this analysis:

Training set: 2,000 reviews (1,000 positive and 1,000 negative)

Test set: 500 reviews (250 positive and 250 negative)

Data Preprocessing

Using Hugging Face's datasets library and DistilBERT's DistilBertTokenizerFast, the preprocessing pipeline was as follows:

Tokenization: Transformed each review into a sequence of tokens with a maximum length of 512 tokens.

Padding and Truncation: Each review was padded or truncated to the fixed length, ensuring that all inputs had the same dimension, allowing efficient batch processing.

Key Data Features

input_ids: Encoded tokens representing words in the review.

attention_mask: Binary mask indicating whether each token is part of the input text (1) or padding (0).

label: Binary sentiment label for each review (0 = Negative, 1 = Positive).

Model Architecture

The model leveraged DistilBERT, specifically DistilBertForSequenceClassification, which is optimized for binary classification tasks while retaining many advantages of the full BERT model.

DistilBERT Base Model: DistilBERT is 40% smaller and 60% faster than BERT, designed for lightweight deployment and speed with minimal performance trade-offs.

Classification Layer: A dense output layer with a Softmax activation function to classify each review as either positive or negative.

Training Efficiency: DistilBERT's architecture allowed faster training and lower memory consumption, which was ideal given the limited resources.

Training Setup and Hyperparameters

The training configuration was adapted to fit computational constraints, aiming to achieve baseline performance with limited epochs and steps:

Batch Size: 4 for training and evaluation.

Epochs: 1 epoch, with evaluation checkpoints after each epoch.

Learning Rate: 2e-5, optimized for DistilBERT's transformer layers.

Maximum Steps: Limited to 10 steps to demonstrate feasibility without full training.

Note: The configuration of only 1 epoch and 10 steps was a necessary limitation; however, future iterations could extend training to improve generalization and performance.

Model Evaluation and Analysis:

Evaluation metrics included accuracy, precision, recall, F1-score, confusion matrix, ROC-AUC, and an analysis of misclassifications. Each provides insights into the model's strengths, weaknesses, and potential areas for improvement.

A. Classification Report:

A classification report breaks down the performance of the model by class (Negative and Positive):

Metric	Negative (Class 0)	Positive (Class 1)
Precision	0.58	0.63
Recall	0.76	0.42
F1-Score	0.66	0.51

- Accuracy: 59.4%
- F1 Score (Macro): 0.5061

Interpretation:

The higher recall for the negative class (0.76) indicates that the model was better at correctly identifying negative reviews, potentially at the cost of misclassifying positive ones.

Precision was higher for positive reviews, showing that when the model predicted positive sentiment, it was relatively reliable, even if it missed some positive reviews (as seen in recall).

*B. **Confusion Matrix:***

The confusion matrix offers a granular view of model predictions:

	Predicted Negative	Predicted Positive
Actual Negative	190	60
Actual Positive	145	105

Insights:

• True Negatives (190): The model accurately predicted a significant portion of negative reviews.

• False Negatives (145): A substantial number of positive reviews were misclassified as negative.

• False Positives (60): Some negative reviews were incorrectly labeled as positive.

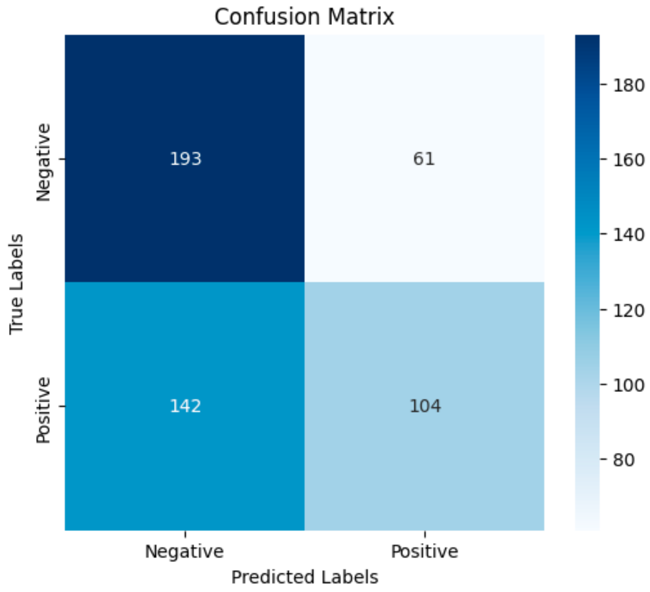

Figure 1

C. *ROC Curve and AUC Score:*

The ROC-AUC curve is a standard for visualizing classification performance. The model achieved an AUC of 0.6032.

An AUC of 0.6032 suggests moderate separation between the two classes, though improvements could be made. A perfect AUC would be 1.0, so this result indicates the model performs slightly better than random chance (AUC = 0.5).

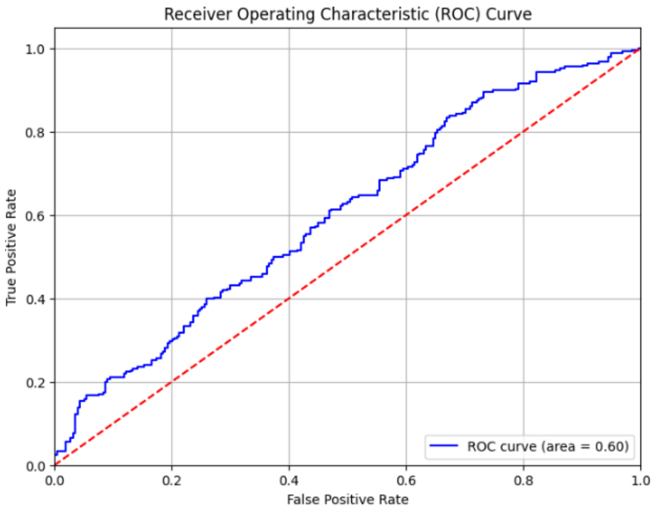

Figure 2

D. *ROC Precision-Recall Curve:*

The Precision-Recall curve is useful for imbalanced data, showing that precision is relatively stable across a range of recall values but decreases at the extremes.

High precision at lower recall suggests that, when confident, the model's predictions are accurate, though it struggles to capture all positive instances. These points to a possible imbalance in recall between classes.

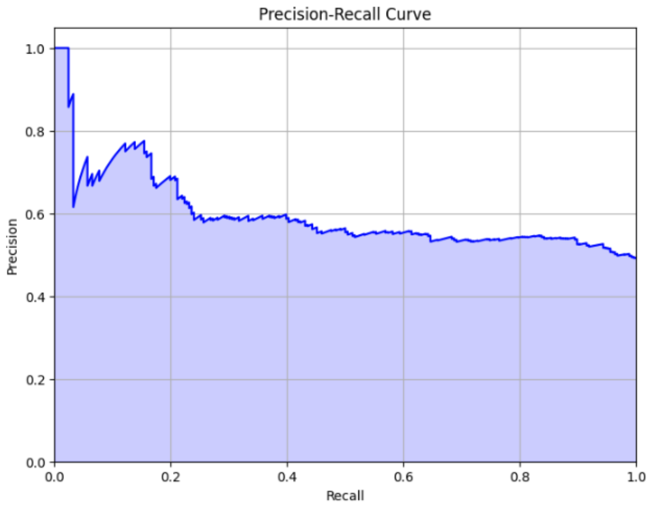

Figure 3

E. *Misclassification Analysis:*

By analyzing misclassified examples, the model showed specific trends:

Longer Reviews: Reviews with more tokens were often misclassified, possibly due to increased complexity.

Ambiguous Language: Reviews with mixed or nuanced language often led to incorrect classifications.

This misclassification pattern may indicate that DistilBERT struggles with highly nuanced sentiment, a common limitation for transformer models on shorter training durations and smaller datasets.

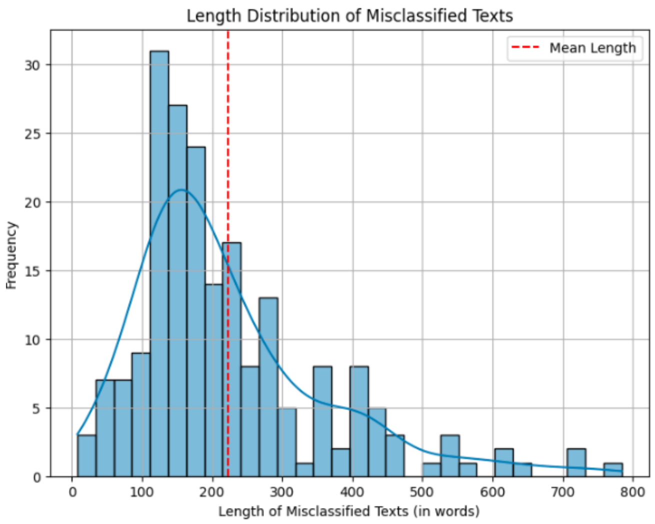

Figure 4

F. *Class Performance Metrics Bar Plot:*

This visualization compares Precision, Recall, and F1 Score for each class ("Negative" and "Positive") in the sentiment analysis model. It provides insights into the model's performance across these key metrics for both classes.

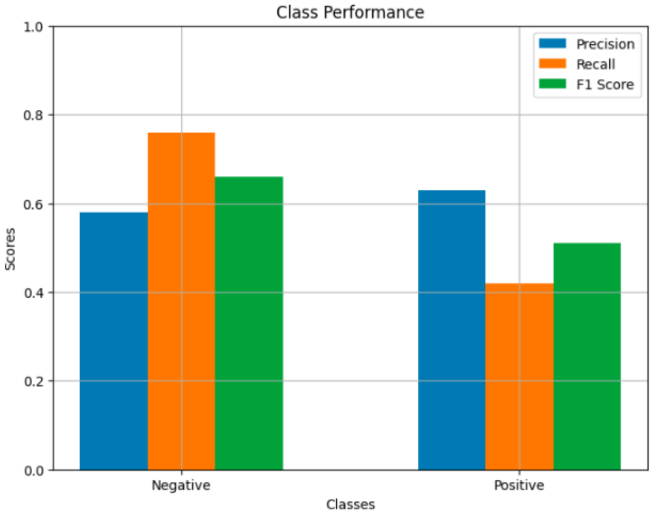

Figure 5

Conclusions and Future Work:

Summary:

The DistilBERT-based sentiment analysis model yielded an accuracy of 59.4% and an AUC score of 0.6032, demonstrating baseline classification capabilities yet leaving room for significant improvement. Performance disparities across classes became apparent, with higher recall observed for the negative class compared to the positive, suggesting a model bias towards identifying negative sentiment. This imbalance could potentially be addressed through techniques such as class weighting, data augmentation, or targeted sampling strategies.

A closer examination of misclassified examples revealed that DistilBERT struggles with complex, nuanced, or ambiguous reviews, a trend that points to the need for further model refinement. The evaluation metrics offered additional insights: the confusion matrix showed a high rate of misclassification for positive reviews; the ROC-AUC curve illustrated a moderate separation between classes, indicating the potential benefit of tuning; and the precision-recall curve suggested that the model maintains relatively high precision but sacrifices recall for positive reviews.

Recommended Next Steps include increasing the number of training epochs to allow the model to better capture sentiment patterns within the dataset. Implementing data augmentation techniques, such as SMOTE or random oversampling of positive reviews, could help balance class representation. Further hyperparameter optimization, including adjustments to learning rates and batch sizes, as well as additional fine-tuning, may yield more robust performance. Finally, adjusting the decision threshold to optimize the precision-recall trade-off could reduce false positives and enhance the model's overall effectiveness.

Future Work:

To fully leverage DistilBERT’s potential for IMDb sentiment analysis, a range of improvements and extensions are recommended. Fine-tuning the model on the complete IMDb dataset with multiple epochs would likely boost accuracy and enhance its ability to generalize across varied review styles. Additionally, employing ensemble methods, which combine DistilBERT with simpler machine learning classifiers, could address errors in complex or ambiguous reviews, increasing model robustness. Applying Explainable AI (XAI) techniques such as SHAP or LIME would offer valuable insights by identifying influential tokens or phrases, thereby improving interpretability and trust in the model’s predictions. In summary, this project highlights DistilBERT’s capabilities in sentiment analysis, although further optimization is necessary to achieve competitive performance levels. The current findings lay a foundation for continued experimentation to refine DistilBERT’s performance on text-based sentiment classification tasks.

References

[1] DistilBERT: A Distilled Version of BERT
Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[2] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT 2019.

[3] IMDB Movie Review Dataset for Sentiment Analysis
Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 142-150).

[4] A Survey on Sentiment Analysis: Techniques and Applications Ravi, K., & Ravi, V. (2015). A survey on sentiment analysis: Techniques and applications. Knowledge-Based Systems, 89, 14-46..

[5] Fine-tuning BERT for Sentiment Analysis
Sun, C., Wang, S., & Zhang, Z. (2019). How to fine-tune BERT for sentiment analysis? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2767-2773).

[6] Understanding the Impact of Data Imbalance in Sentiment Analysis Zhang, Y., & Zhang, D. (2021). Understanding the impact of data imbalance in sentiment analysis. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (pp. 3189-3201)..

[7] Classifying Movie Reviews: Sentiment Analysis and Evaluation of BERT and Other Classifiers Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

[8] A Comprehensive Review on Evaluation Metrics for Sentiment Analysis Jebelli, H., & Tefagh, M. (2019). A comprehensive review on evaluation metrics for sentiment analysis. Journal of King Saud University-Computer and Information Sciences.

[9] Applications of Transformer Models in NLP Tasks
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS 2017).