Fine-Tuning distilBERT for Enhanced Sentiment Classification

Sarah Ling
Markville Secondary School, Ontario, L3P 7P5, Unionville, Canada

Abstract:This research examines the fine-tuning of the DistilBERT model for sentiment classification using the IMDB dataset of 50,000 movie reviews. Sentiment analysis is vital in natural language processing (NLP), providing insights into emotions and opinions within textual data. We compare the fine-tuned DistilBERT and LLaMA 3 models, focusing on their ability to classify reviews as positive or negative. Through few-shot training on the dataset, our findings reveal that while LLaMA 3 8B excels in capturing complex sentiments, DistilBERT-base-uncased offers a more efficient solution for simpler tasks. The results underscore the effectiveness of fine-tuning. This paper contributes to optimizing sentiment analysis models and suggests future research directions, including hybrid models and advanced training techniques for improved performance across diverse contexts.

Keywords: Sentiment Classification; Fine-Tuning; Natural Language Processing; Large Language Models; Text Classification; Machine Learning; Transformer Models

1. Introduction

Sentiment analysis has been an established area of research in natural language processing (NLP) that studies people's sentiments, opinions, emotions, etc. through computational methods[4][7]. This field has gained signifi- cant interest in both academia and industry fields due to its useful applications in analyzing customer feedback, decision-making, and product creation. Sentiment classification can be defined as the procedure of assigning predetermined sentiment classes (positive, negative) depending on the emotional tone of a message through analyzing text. In NLP this task is widely used to determine the polarity of opinions expressed in text. In recent years, large language models (LLMs) have been popular in various NLP tasks, and a deeper understanding of human emotions through sentiment classification is an important stepping stone towards developing artificial intelligence[1]

Recent work shows that models such as BERT [2] and LLaMA [11] perform well in general sentiment analysis tasks but still struggle with nuanced or structured sentiment tasks, especially when more refined emotional or opinion-based distinctions are required [9]. Despite advancements in LLMs, there are challenges in applying them to complex sentiment tasks, including identifying subtle emotions and handling domain-specific contexts [5, 7]. We propose comparing the fine-tuned DistilBERT and LLaMA 3 models, evaluating their performance on datasets for sentiment analysis. DistilBERT offers computational efficiency, while LLaMA 3 leverages a larger architecture for complex tasks. This allows for a practical assessment of trade-offs between model size and performance.This paper compares the performance of DistilBERT and LLaMA 3 in sentiment analysis using few-shot training. We fine-tune both models on domain-specific datasets, including the IMDB Kaggle movie review dataset. In this research, we fine-tune both models, asking them to classify reviews as positive or negative. We test both models in zero-shot and fine-tuning scenarios to evaluate their generalization across different domains.

1.1 Background Dataset

An IMDB Dataset of 50,000 Movie Reviews is used to train our models. This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. This dataset provides 25,000 highly polar movie reviews for training and 25,000 for testing[6].Large Language Models (LLMs)

A large language model (LLM) is a machine learning model designed to process and generate human language text. Built on transformer architectures [10], LLMs are trained on extensive datasets using deep learning techniques to understand relationships among characters, words, and sentences. They analyze patterns in unstructured data to identify

grammatical rules, semantics, and contextual nuances.

Through probabilistic methods, LLMs predict and generate coherent text without requiring human supervision during training. This enables them to perform a variety of natural language processing tasks, such as summariza- tion, translation, and sentiment analysis. The transformer-based approach significantly enhances their ability to process language efficiently and accurately, making LLMs a powerful tool for advancing AI applications.

1.2 Models

The first model used is the Meta LLaMa 3 8B, which is the next generation of Meta's state-of-the-art open-source large language model[11]. Llama 3 comes in configurations ranging from 8 billion to 70 billion parameters, making it a highly scalable and powerful model capable of processing large amounts of data for diverse applications[3]. It is designed to compete with and surpass existing models in terms of performance across various tasks, including language understanding, coding, reasoning, and more.

The second model used is distilbert-base-uncased which is a distilled version of the BERT base model [8]. DistilBERT is a transformers model, smaller and faster than BERT, which was pretrained on the same corpus in a self-supervised fashion, using the BERT base model as a teacher. The model was trained to return the same probabilities as the BERT base model. DistilBERT has about 66 million parameters, which makes it smaller and more lightweight than the original BERT model and is designed to retain around 97% of BERT's performance while being 60% smaller and faster.

The DistilBERT model was selected for this investigation since it is primarily used for tasks that require transformer models but require fine-tuning. It has been widely used for fine-tuning tasks that use the whole sentence to make decisions, such as sequence classification, token classification, or question answering.

Both models' checkpoints hosted on huggingface are used for the inference.

1.3 Fine-Tuning

Language models are often further trained via a process named fine-tuning. Fine-tuning in machine learning is the process of adapting a pre-trained model for specific tasks or use cases. It has become a fundamental deep learning technique, particularly in the training process of foundation models used for generative AI. Fine-tuning for specific tasks such as interpreting questions and generating responses, or translating text from one language to another are common. In this investigation, we finetune the distilbert-base-uncased and Llama 3 models and compare their performance for the task of sentiment classification.

2. System Design

2.1 Overview

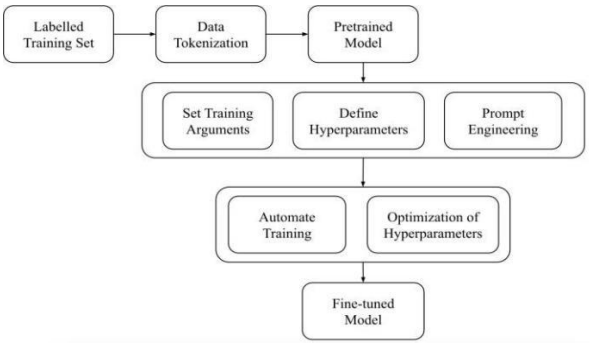

Figure 1. System Overview

Figure 1 provides an outline of the steps taken to finetune a model.

2.2 Data Pre-Processing

The first stage of our process involves preparing the IMDB dataset for use with the transformer model. Since transformer-based models like DistilBERT require tokenized input, we preprocess the data by converting text into a format that the model can interpret.

Tokenization: Each movie review is tokenized using a pre-trained tokenizer from the DistilBERT model. This tokenizer splits text into subword units and converts them into integer token IDs. The tokenizer also handles punctuation, case normalization (lowercasing), and truncation to ensure that the input sequence fits the model's maximum input length.

Padding: To ensure uniform input lengths during batching, tokenized sequences are padded. Padding adds special tokens to shorter sequences so that they match the maximum sequence length in each batch.Once tokenized and padded, the preprocessed dataset is ready for input into the model.

3. Model Definition

The distilBERT model is initialized with weights pre-trained on a large corpus of text data,

allowing it to already understand general language features. We then modify the pre-trained model for the specific task of sentiment classification by:

Adding a Classification Layer: The pre-trained DistilBERT model outputs a contextualized representation for each token in the input sequence. A fully connected layer (classification head) is added on top of the model, with two output nodes corresponding to the two sentiment labels: positive and negative.Load model: We load pre-trained DistilBERT model then add the classification layer with 2 output labels: positive, negative.Label Mapping: A mapping between sentiment labels and numerical IDs is defined (e.g., 0 for negative and 1 for positive). This ensures that predictions made by the model are interpretable.

3.1 Finetuning

The fine-tuning process involves training the pre-trained DistilBERT model on the IMDB dataset while updating its weights to specialize in sentiment analysis. Fine-tuning is a crucial step because it allows the model to transfer its general language understanding to the specific task of classifying movie reviews. We set the seed to a random fixed number to ensure fair comparison and reproductibility. A smaller subset of 3000 reviews was created for training and testing for the distilBERT model. A subset of 100 reviews was used for the LLaMa 3 8B model due to constraints on available RAM for the free version of Google Colab. This smaller subset may affect the finetuning capabilities of the LLaMa 3 8B model since it is not given as much training data as the distilBERT model. This difference in training data will be taken into consideration for comparing the models' performance.

Training: The model is trained using a supervised learning approach. The tokenized IMDB dataset is split into training and evaluation sets. During training, the model learns to minimize the classification error by adjusting its weights via backpropagation. A learning rate scheduler and optimizer (AdamW) are used to ensure smooth convergence and prevent overfitting.

3.2 Training Parameters Were Configured

- output_dir: Directory to save the model.
- num_train_epochs: Number of training epochs, set to 2.
- per_device_train_batch_size: Batch size per device, set to 16. • learning_rate: Set to 2 × 10−5
- weight_decay: Set to 0.01
- optimizer: PagedAdamW with 32-bit precision.

After fine-tuning, the model's performance improved significantly, achieving an overall accuracy of 0.88. This result demonstrates the effectiveness of the fine-tuning process. By the end of the fine-tuning stage, the model is well-adjusted to the task of sentiment classification and can be used to make predictions on unseen movie reviews.

4. Evaluation

This section presents the experimental results obtained from evaluating the model after fine-tuning. The perfor- mance was assessed using precision, recall, and F1-score metrics for each sentiment class (positive and negative) as well as the overall accuracy of the model.

4.1 Performance

The confusion matrixes below provides a breakdown of the model's predictions across all sentiment classes, as shown in Figures 2 and 3. This analysis helps in identifying common misclassifications and understanding the model's strengths and weaknesses.

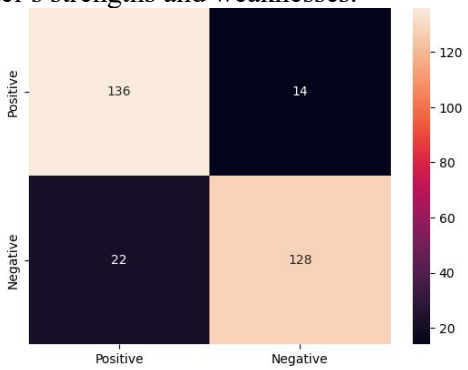

Figure 2. DistilBERT-Base-Uncased

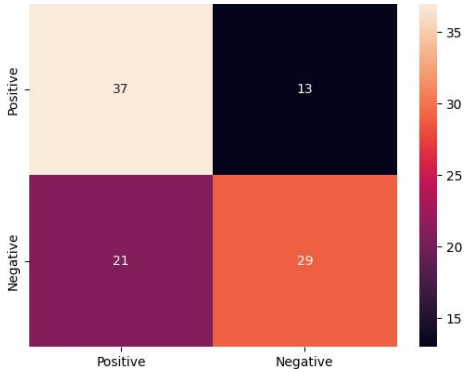

Figure 3. LLaMa-3-8B

The evaluation of the model's performance was based on standard metrics including precision, recall, and F1-score. These metrics were computed for each sentiment class (positive and negative) as well as for the overall model performance. The formulas for these metrics are as follows.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Table 1. Comparison of Model Performance

Model	Accuracy	Precision	Recall	F1 Score
distilBERT-base-uncased	0.88	0.86075	0.90666	0.88311
Meta-Llama-3-8B	0.66	0.63793	0.74000	0.68518

4.2 Latency

DistilBERT-base-uncased is much faster, with low latency suitable for simple real-time applications such as sentiment classification due to its smaller size and efficient architecture. For this task of sentiment classification, it took approximately 2:23min to run each epoch.

Meta-Llama-3-8B provides more powerful capabilities and nuanced understanding but comes with higher

latency, making it more suitable for applications where processing speed is less critical compared to output quality.

5. Discussion

This study compared the performance of DistilBERT and LLaMA 3 on sentiment classification using the IMDB dataset. Our findings indicate that both models have their strengths and weaknesses, and understanding these can provide insights into optimizing sentiment analysis systems in practical applications.

Model Performance: The DistilBERT model had higher accuracy, precision, recall, and F1 score, suggesting that it outperformed LLaMa 3 on the task of sentiment analysis for our IMDB dataset.

Generalization Across Domains: The models performed well on the IMDB dataset but may struggle to generalize to different domains, where sentiment expressions vary significantly. Implementing domain adaptation techniques and fine-tuning on diverse datasets could enhance model robustness. Highly polar movie reviews were used, and both models' ability to detect nuanced sentiments was not tested. Future work could explore specialized training datasets to improve the recognition of subtle sentiments.

Promising avenues for future research include developing hybrid models that combine the strengths of both LLaMA 3 and DistilBERT. Enhancing training techniques, such as few-shot and zero-shot learning, will also be essential for improving performance across diverse contexts. Addressing the limitations identified in this study will be crucial for creating more accurate and efficient sentiment classification systems that effectively meet user needs. These findings underscore the importance of model selection based on application requirements.

Future Work: Several avenues for future exploration remain:

- Leveraging hybrid models combining efficiency and nuanced understanding [5].
- Expanding experiments to include domain-specific datasets to enhance generalization.
- Investigating advanced fine-tuning techniques, such as adapters and LoRA layers, for further optimization [11].

6. Conclusion

In this study, we conduct an evaluation of sentiment classification by comparing the performance of a large language model and a small language model. The DistilBERT model had higher accuracy, precision, recall, and F1 score, suggesting that it outperformed LLaMa 3 on the task of sentiment analysis for our IMDB dataset. These findings suggest that model size alone does not guarantee better performance, emphasizing the importance of selecting the appropriate model for specific tasks. Future work could explore the influence of dataset size, fine-tuning strategies, and domain-specific training to further optimize sentiment classification models.

References

[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 (2023).

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.).Association for Computational Linguistics, Minneapolis, Minnesota,4171–4186. https://doi.org/10.18653/v1/N19-1423

[3]Diefan Lin, Yi Wen, Weishi Wang, and Yan Su. 2024. Enhanced Sentiment Intensity Regression Through LoRA Fine-Tuning on Llama 3.IEEE Access 12 (2024), 108072–108087. https://doi.org/10.1109/ACCESS.2024.3438353

[4] Bing Liu. 2020. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge university press.

[5] Haochen Liu, Sai Krishna Rallabandi, Yijing Wu, Parag Pravin Dakle, and Preethi Raghavan. 2023. Self-training Strategies for Sentiment Analysis: An Empirical Study. arXiv preprint arXiv:2309.08777 (2023).

[6] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). Association for Computational Linguistics, Portland, Oregon, USA, 142–150. https://aclanthology.org/P11-1015Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2023. Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. IEEE Transactions on Affective Computing 14, 1 (2023), 108–132. https://doi.org/10.1109/TAFFC.2020.3038167

[7] V Sanh. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019).

[8] Amira Samy Talaat. 2023. Sentiment analysis classification system using hybrid BERT models. Journal of Big Data 10, 1 (2023), 110.

[9] A Vaswani. 2017. Attention is all you need. Advances in Neural Information Processing Systems(2017).

[10] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. arXiv preprint arXiv:2401.02385 (2024).

