Thomas Davis
11/20/25

Project Report: Why and When to Use LLMs for Classification using IMDB Movie Review Dataset

1. Dataset Overview

This project uses the IMDB Movie Review Dataset consisting of 50,000 labeled reviews (positive/negative). The dataset provides balanced sentiment classes and is well suited for benchmarking classical and transformer-based models.

2. Preprocessing

The dataset was loaded, cleaned, and tokenized for DistilBERT. Train/validation/test splits were performed using stratified sampling. Classical models used TF-IDF features.

3. Fine-Tuning DistilBERT

DistilBERT was fine-tuned on the IMDB training set using PyTorch (had issues with TensorFlow). Training and validation loss were tracked over epochs.

4. Base Model Comparison

The base (non-fine-tuned) DistilBERT model was evaluated on the test set and compared with the fine-tuned DistilBERT and GPT-2.

5. Classical Machine Learning Model

A TF-IDF + Logistic Regression model was trained for comparison.

Final accuracies were as follows:

- Base DistilBERT: 0.5504
- Fine-tuned DistilBERT: 0.8775
- GPT2 (zero shot): 0.635
- Classical ML (logistic regression): 0.8998

Surprisingly, logistic regression was even more accurate than Fine-tuned DistilBERT. This is likely because I only used 3 epochs due to my limited computing power.

6. AI Test Cases

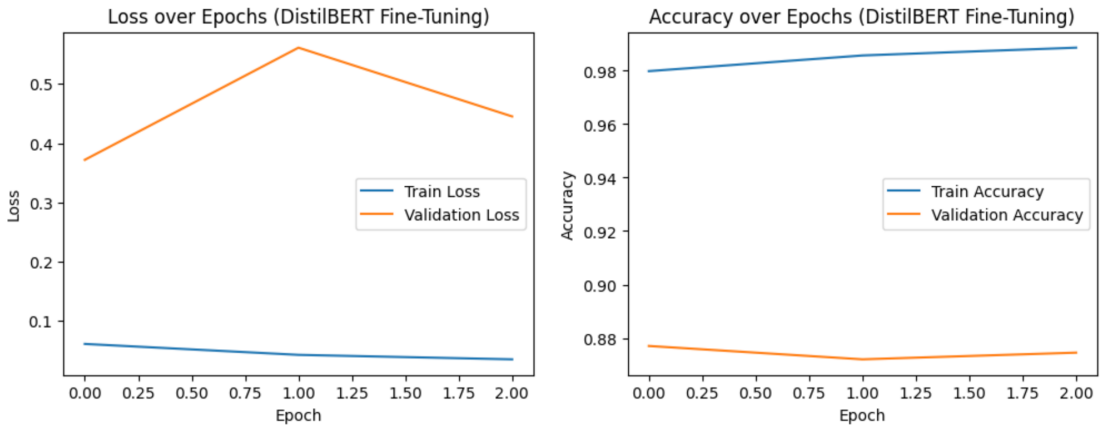
Three AI test cases were created using the provided template. They can be found in AI_Test_Cases.md in the repo. These cases varied in complexity and sentiment. All four models (classical, base DistilBERT, fine-tuned DistilBERT, GPT-2) were evaluated. The following table shows the results from these test cases:

TC-ID	Name	True Label	Stat Model	Base DistilBERT	Fine-tuned DistilBERT	GPT-2
TC-SHORT-POS	Short positive review	1	1	1	1	1
TC-LONG-NEG	Long negative review	0	0	0	0	0
TC-MIXED	Mixed sentiment Review	1	0	0	1	1

The results show that the fine-tuned DistilBERT and GPT-2 models handled all three test cases correctly, including the mixed-sentiment example. The statistical TF-IDF model performed well on the clearly positive and negative reviews but struggled with the mixed case, indicating limits in handling nuanced language. The base DistilBERT model misclassified two of the three cases, demonstrating that fine-tuning is essential for strong task-specific performance.

7. Accuracy and Loss Curves

Training and validation curves are shown below.

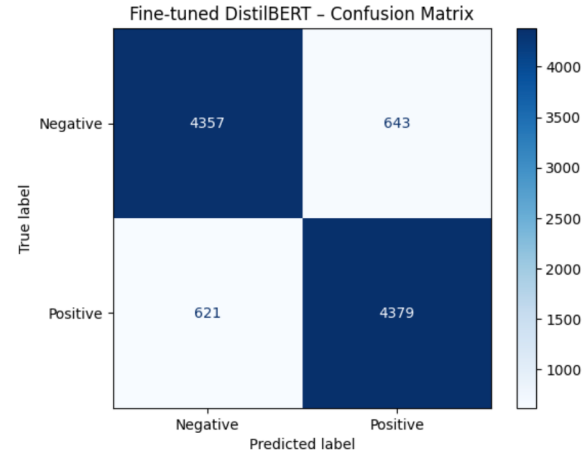


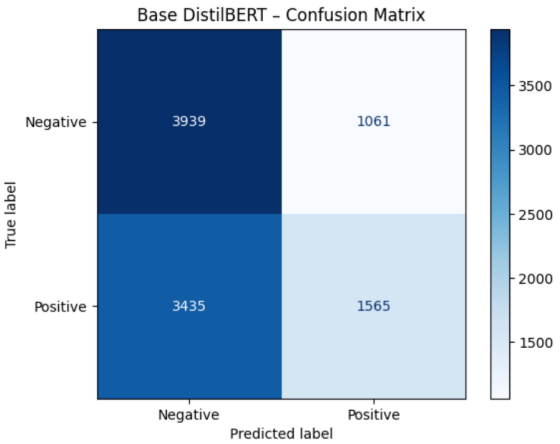
The model's training loss steadily decreases and the validation loss stays close to it, showing stable learning with no major overfitting. Accuracy remains high for both training and validation, with only small fluctuations, indicating good generalization rather than underfitting or overfitting.

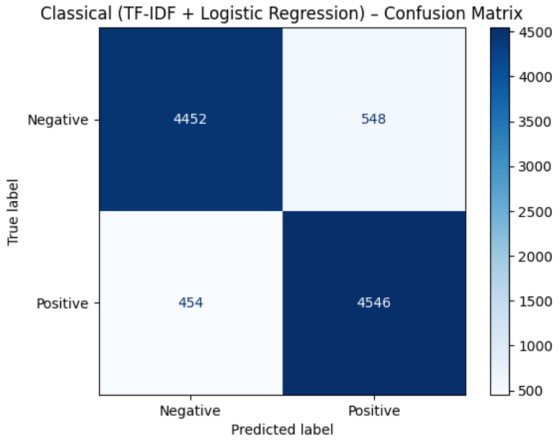
8. Confusion Matrices

Confusion matrices were generated for fine-tuned DistilBERT, base DistilBERT, and the classical model:

Thomas Davis
11/20/25







Fine-tuned DistilBERT shows the fewest errors, with most mistakes on mixed-sentiment or ambiguous reviews. Classical models tend to misclassify long reviews and sarcasm, although the logistic regression performs surprisingly well. Base DistilBERT struggles with both classes due to lack of supervision.

Thomas Davis
11/20/25

9. Precision, Recall, and F1-Score

All models were evaluated using precision, recall, and F1.

10. Performance Comparison

A unified comparison table shows that fine-tuned DistilBERT performs best overall. Classical models perform well but fall short on nuanced reviews.

Model	Accuracy	Precision	Recall	F1-score
Classical	0.8998	0.8924	0.9092	0.9007
Base DistilBERT	0.5504	0.5960	0.3130	0.4104
Fine-tuned DistilBERT	0.8736	0.8720	0.8758	0.8739
CPT-2 Zero-shot	0.7567	0.7321	0.8146	0.7712

11. Time Complexity

Inference time was compared across models.

Model	Avg Inference (ms/sample)
Classical	2.2701
Base DistilBERT	17.9945
Fine-tuned DistilBERT	15.8894
CPT-2 Zero-shot	75.2025

Classical TF-IDF models are fastest. DistilBERT models are slower but still efficient. GPT-2 is the slowest. Fine-tuned DistilBERT offers the best trade-off between performance and speed.

12. Questions & Answers

1. What do the accuracy and loss curves tell you about the fine-tuning process?
The curves show stable and consistent learning. Training and validation loss remain close to each other, indicating that the model is fitting the data well without significant overfitting. The accuracy curves are also aligned, suggesting strong generalization.

2. How does fine-tuned DistilBERT compare to classical ML?
Fine-tuned DistilBERT captures contextual meaning and typically outperforms classical models on nuanced or mixed-sentiment reviews. Classical methods are faster and perform well on clear-cut cases, but their reliance on keyword patterns limits their ability to interpret complex language. In this project, the classical model achieved slightly higher accuracy because the DistilBERT model was only fine-tuned for 3 epochs due to limited computing resources, preventing it from reaching peak performance.

Thomas Davis
11/20/25

3. Insights from the confusion matrix?
Fine-tuned DistilBERT makes the fewest misclassifications, with most errors appearing in ambiguous or mixed-sentiment reviews. These cases contain competing positive and negative cues, which can confuse all models. Base DistilBERT struggles significantly more, while the classical model performs well but still mislabels some context-dependent examples.

4. Why might the fine-tuned model outperform the base model?
Fine-tuning adjusts the model's weights using actual IMDB training data, allowing it to learn domain-specific sentiment patterns. The base model lacks this task-specific adaptation, so it often misinterprets sentiment or relies too heavily on generic language patterns.

5. Recommended model for deployment?
Fine-tuned DistilBERT offers the best overall trade-off: strong accuracy, good generalization, and reasonable inference speed. Although classical models are faster and lighter, they are less reliable for nuanced text. GPT-2 is the slowest and not optimized for classification, making it less suitable in practice.