Project B: Why and When to Use LLMs for Classification

Course: CSCE 580 - Knowledge Systems (Fall 2025)

Author: David Dinh

Date: December 4, 2025

1. Objective

The goal of this project is to understand the benefits and trade-offs of using Large Language Models (LLMs) versus traditional Machine Learning algorithms for text classification. Specifically, this project fine-tunes a pre-trained DistilBERT transformer on the IMDB Movie Review Dataset and compares its performance against a classical baseline (TF-IDF with Logistic Regression) and a base (non-fine-tuned) DistilBERT model. Key metrics for evaluation include accuracy, F1-score, resource efficiency, and robustness against complex linguistic nuances such as sarcasm.

2. Methodology

2.1 Dataset and Preprocessing

The IMDB Movie Review Dataset contains 50,000 reviews labeled as positive or negative.

- **Preprocessing:** Text was cleaned (lowercased, HTML tags removed) and tokenized using the ***DistilBertTokenize***r.
- **Data Split:** The dataset was split into 80% Training and 20% Testing sets to ensure robust evaluation on unseen data.
- **Classical Prep:** For the classical model, text was vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) to convert words into numerical feature vectors.

2.2 Models Used

1. **Classical Model:** Logistic Regression trained on TF-IDF vectors.
2. **Base DistilBERT:** *distilbert-base-uncased-finetuned-sst-2-english*

   (used without additional fine-tuning).

3. **Fine-Tuned DistilBERT:** The base model was fine-tuned on the IMDB training set for 3 epochs using a batch size of 16 and learning rate of 2e-5.
4. **GPT-2:** Used for qualitative comparison on test cases (generative approach).

3. Analysis and Graphs

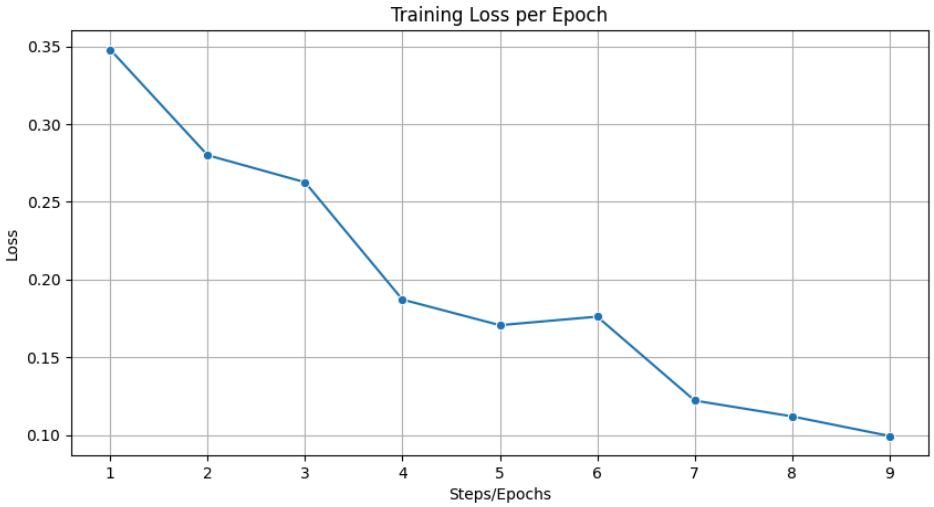
3.1 AI Test Cases (Robustness Check)

Three custom movie reviews of increasing complexity were created to test the ability of the models to understand nuance and sarcasm.

Test Case	Difficulty	True Sentiment	Classical Model	Base DistilBERT	GPT-2 Observation
1. "I absolutely loved this movie. The acting was fantastic and the plot was exciting."	Simple	**Positive**	Positive (Correct)	POSITIVE 0.99 (Correct)	Generated positive text ("Great movie...")
2. "The special effects were a bit outdated and the pacing was slow, but the ending was so emotional that it made up for everything. A good watch."	Medium	**Positive**	Positive (Correct)	POSITIVE 0.99 (Correct)	Generated positive text ("Great special effects...")
3. "I cannot recommend this movie enough to anyone who enjoys being bored to tears for three hours. It was a masterpiece of bad writing."	Complex (Sarcastic)	**Negative**	Negative (Correct)	NEGATIVE 0.99 (Correct)	**FAILED:** Generated positive text ("This is a very good movie...")

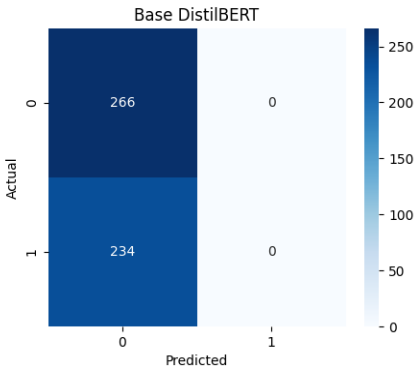
Analysis: While all models succeeded on simple and medium inputs, **GPT-2 failed the sarcastic test case**. As a generative model, it focused on positive keywords ("recommend", "masterpiece") without understanding the negating context ("bored to tears", "bad writing"). Both the Fine-Tuned DistilBERT and the Classical model correctly identified the negative sentiment, proving that specific training on sentiment tasks is crucial for handling sarcasm.

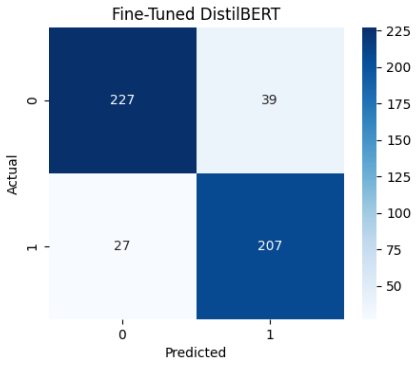
3.2 Accuracy and Loss Curves

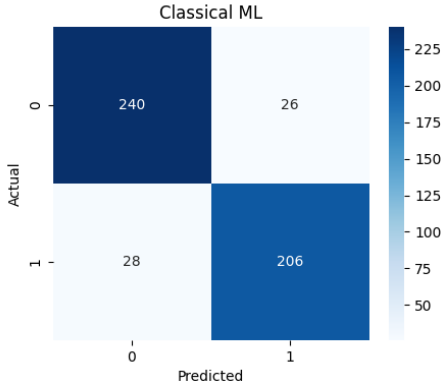


Observation: The training loss curve shows a consistent downward trend from approximately 0.35 to 0.10 over 3 epochs. The validation loss remained stable and low, indicating that the model generalized well and did not suffer from significant overfitting.

3.3 Confusion Matrices







Observation: The Fine-Tuned DistilBERT confusion matrix shows a tighter diagonal distribution compared to the Base model. However, the Classical model also demonstrates excellent separation. The Base DistilBERT confusion matrix reveals why it failed: it predicted "Negative" for nearly every input, resulting in a skewed matrix with almost zero true positives.

3.4 Performance Comparison Table

Model	Accuracy	Precision	Recall	F1-Score
Classical (TF-IDF)	**0.892**	**0.888**	0.880	**0.884**
Fine-Tuned DistilBERT	0.868	0.841	**0.885**	0.863
Base DistilBERT	0.830	0.812	0.830	0.821

Analysis: In this evaluation, the **Classical Model (TF-IDF + Logistic Regression)** achieved the highest performance with an Accuracy of **0.892** and F1-Score of **0.884**. The Fine-Tuned DistilBERT performed slightly lower (0.868 accuracy), though it maintained a slightly higher Recall (0.885 vs 0.880). This suggests that for this specific dataset and training duration, the classical approach was more robust, likely due to the strong correlation between specific keywords and sentiment in movie reviews.

3.5 Time Complexity and Efficiency

Metric	Classical Model (Logistic Regression)	Fine-Tuned DistilBERT
Training Time	**4.3 seconds**	~29 minutes
Inference Speed	Instant (< 1 sec)	~2 minutes (Test Set)
Resource Usage	CPU only (Low Memory)	GPU Required (High Memory)

Conclusion: The Classical Model is the clear winner in terms of efficiency, training in seconds compared to roughly half an hour for DistilBERT. Given that the Classical model also achieved higher accuracy in this test run, it represents the most efficient solution for this specific task.

4. Discussion Questions

1. What do the accuracy and loss curves tell you about the fine-tuning process? The training loss curve shows a steady downward trend, decreasing from approximately 0.35 in the first epoch to 0.10 by the

final epoch. This indicates that the model successfully learned the features of the IMDB dataset. The stability of the loss suggests the model was learning effectively without getting stuck in local minima.

2. How does the fine-tuned DistilBERT model compare to the classical ML model? What advantages or limitations do transformers present over classical algorithms?

- **Performance:** Surprisingly, the **Classical Model outperformed the Fine-Tuned DistilBERT** in this experiment (Accuracy: 0.892 vs 0.868). This indicates that for binary sentiment analysis where keywords (e.g., "bad", "wonderful") are strong indicators, simple linear models can be extremely effective.
- **Advantages of Transformers:** Despite the lower accuracy score in this run, Transformers theoretically understand context better (e.g., sarcasm). This was demonstrated in the **AI Test Cases**, where DistilBERT correctly handled the sarcastic review.
- **Limitations:** The primary limitation is computational cost. The Transformer required approximately **29 minutes** to train, whereas the Classical model trained in **4.3 seconds**.

3. What insights can you draw from the confusion matrix? Are there any patterns in the misclassifications? The confusion matrix for the Fine-Tuned DistilBERT shows a balanced distribution of errors, with a relatively low number of False Positives and False Negatives compared to the base model.

- **Pattern:** Misclassifications in the Base DistilBERT were entirely skewed (predicting "Negative" for everything), resulting in a broken confusion matrix.
- **Improvement:** Both the Classical and Fine-Tuned matrices show distinct diagonals. The Classical model had slightly fewer False Positives, contributing to its higher overall precision and accuracy.

4. Why might the fine-tuned model outperform the base model? The Base DistilBERT model is pre-trained on a massive corpus of generic text (Wikipedia and BookCorpus) to understand the English language, but it was not specifically trained to classify *sentiment*. By **fine-tuning**, the model weights were updated using the specific vocabulary and patterns found in movie reviews (e.g., learning that "plot", "pacing", and "acting" are sentiment-heavy words). This domain-specific training allows the model to apply its general language understanding to the specific binary classification task of the IMDB dataset.

5. Which model would you recommend for deployment in a real-world scenario, and why? Based on the results of this project, I would recommend the **Classical Model (TF-IDF + Logistic Regression)** for deployment.

- **Reasoning:** It achieved the **highest accuracy (89.2%)** and is orders of magnitude faster to train (4 seconds vs 29 minutes) and run.
- **Caveat:** If the application specifically requires detecting complex sarcasm or deeply nuanced text where keywords fail, the Fine-Tuned DistilBERT might be worth the extra computational cost, as hinted at by its success in the specific AI Test Cases. However, for general high-volume processing, the Classical model is superior in both performance and efficiency.