Analysis & Discussion

1. What do the accuracy and loss curves tell you about the fine-tuning process?

The refined DistilBERT model learnt satisfactorily across the training epochs, as seen by the training and validation loss curves. The model successfully reduced error on the training data, as evidenced by the training loss rapidly declining with each epoch. By the last epoch, the validation loss also dropped and stabilized, indicating that the model performed well on new data.

As the model approached optimal learning, performance increases began to level off, as seen in the validation accuracy's steady climb across epochs before plateauing. The very narrow difference between training loss and validation loss indicates low overfitting, suggesting that the model learnt significant patterns pertinent to sentiment classification rather than just memorizing the training data. Overall, there are no notable signs of underfitting or overfitting, and the curves indicate a robust training process with strong generalization.

---

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

The fine-tuned DistilBERT model achieved the highest overall accuracy (about 91%), slightly exceeding the traditional Logistic Regression model (around 90%). DistilBERT showed better comprehension of subtle language, such as statements with contrast or complicated feelings (e.g., "the visuals were impressive but the story was confusing"), despite this little difference.

Because transformers like DistilBERT evaluate words in relation to their surrounding context and capture semantic meaning more effectively than TF-IDF-based classical algorithms, they offer deep contextual awareness. They are therefore more equipped to analyze intricate language patterns and nuanced sentiment clues.

However, transformers also have limits. Compared to traditional models, they use much more memory, take longer to train, and require much more processing power. Logistic regression, on the other hand, achieved competitive results and trained very rapidly, making it more economical and efficient, even though its accuracy was marginally lower. Therefore, even if transformers offer better language comprehension, their cost-performance ratio must be carefully considered.

---

3. What insights can you draw from the confusion matrix? Are there any patterns in the misclassifications?

With relatively few misclassifications, the refined DistilBERT model's confusion matrix shows a balanced distribution of genuine positives and true negatives. This indicates that the model shows no significant bias toward either class and performs consistently across both sentiment groups.

The confusion matrix of the base DistilBERT model, on the other hand, shows a significant propensity to predict the negative class, leading to a disproportionately high number of false negatives. This demonstrates that the model was unable to adjust to the IMDB sentiment domain without fine-tuning.

The GPT-2 model regularly misclassified unfavorable ratings as good, demonstrating a blatant optimism bias. This trend suggests that generative language models are unreliable for categorization tasks without task-specific fine-tuning.

In contrast to the fine-tuned transformer, the traditional Logistic Regression model had balanced errors but struggled with complex or mixed-sentiment phrases, suggesting a weakness in interpreting subtle contextual meaning.

---

4. Why might the fine-tuned model outperform the base model?

Large-scale generic corpora are used to pretrain the fundamental DistilBERT model; however, it is not particularly tailored for sentiment classification in the context of movie reviews. Consequently, it lacks specialized knowledge of sentiment-specific patterns and terminology, despite having a general understanding of language structure.

Using labeled IMDB sentiment data, fine-tuning selectively modifies the model's parameters, enabling it to discover domain-relevant correlations between phrases and sentiment polarity. The model can more accurately evaluate contextual meaning, identify negation patterns, and discern subtle emotional cues thanks to this alignment. As a result, the optimized model performs noticeably better than the original model, which does not use task-specific adaptation.

---

5. Which model would you recommend for deployment in a real-world scenario, and why?

The best approach for real-world deployment is determined by the application's priorities.

The optimized DistilBERT model is the best option if accuracy and language comprehension are your main goals. It produced balanced predictions, handled complex sentence structures well, and achieved the highest overall accuracy.

However, the traditional Logistic Regression model is a strong alternative if speed, efficiency, and resource constraints are essential considerations. With much lower processing costs and faster inference times, it achieved almost competitive accuracy.

Given both performance and efficiency considerations, a practical recommendation would be:

- Use **fine-tuned DistilBERT** in high-stakes or customer-facing systems requiring superior accuracy.

- Use **Logistic Regression** in large-scale systems that require rapid, real-time processing with minimal infrastructure cost.

GPT-2 is not recommended for deployment due to its slow inference speed, high resource consumption, and inconsistent performance when used as a zero-shot classifier.

Real-World Implications and Trade-Offs

This experiment demonstrates that although Large Language Models offer remarkable gains in accuracy and semantic comprehension, their operational complexity and computing expense render them inappropriate for every situation. Because of their effectiveness, transparency, and ease of use, classical algorithms remain very important today.

The findings emphasize the importance of choosing models based on scalability, maintainability, and cost-effectiveness, in addition to peak performance. The balance between available resources and predictive capability should inform model selection in real-world applications.