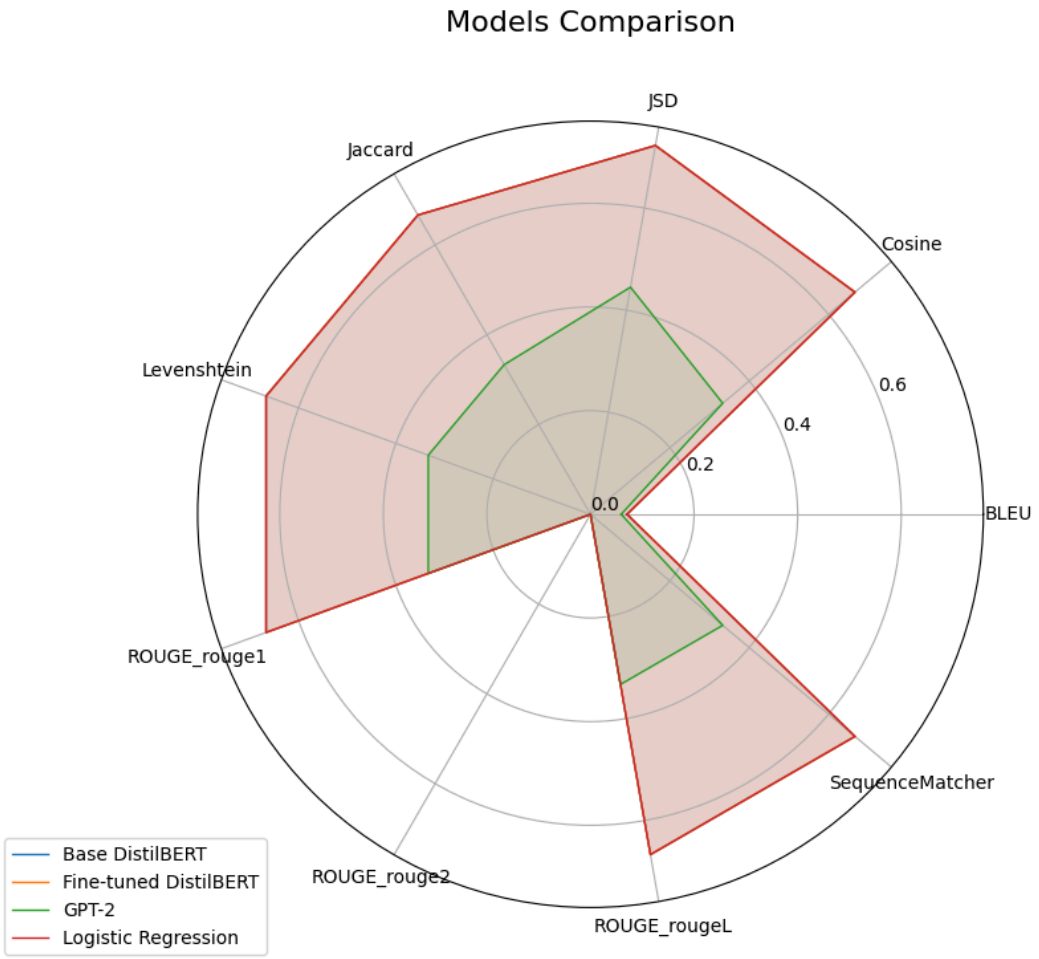
Sentiment Analysis Report

by: Justin Mateo for CSCE580 Project B

AI Test Cases:

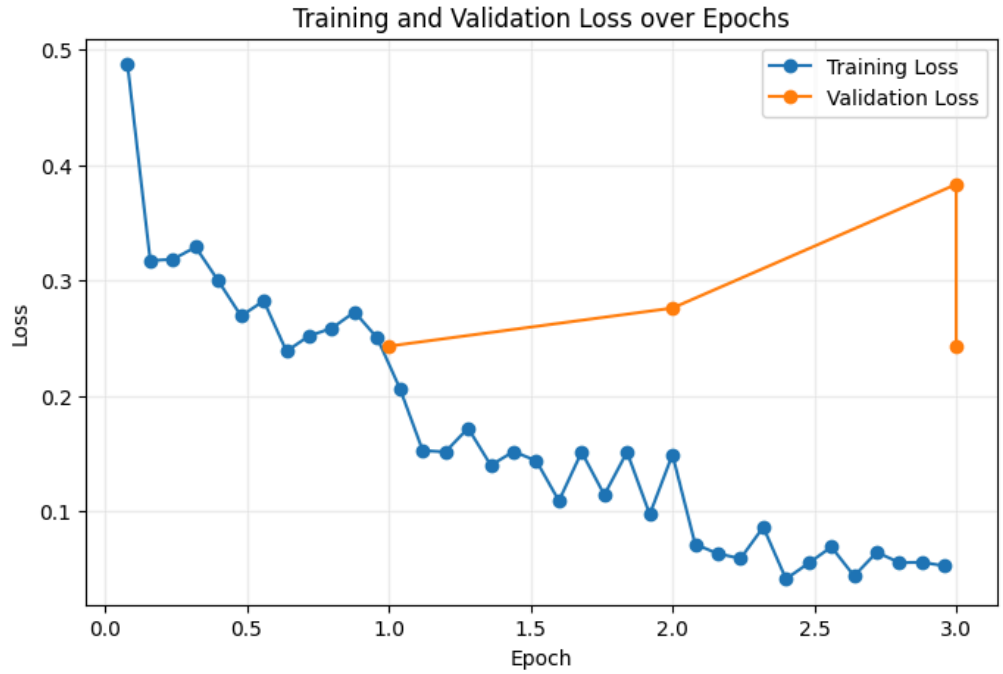


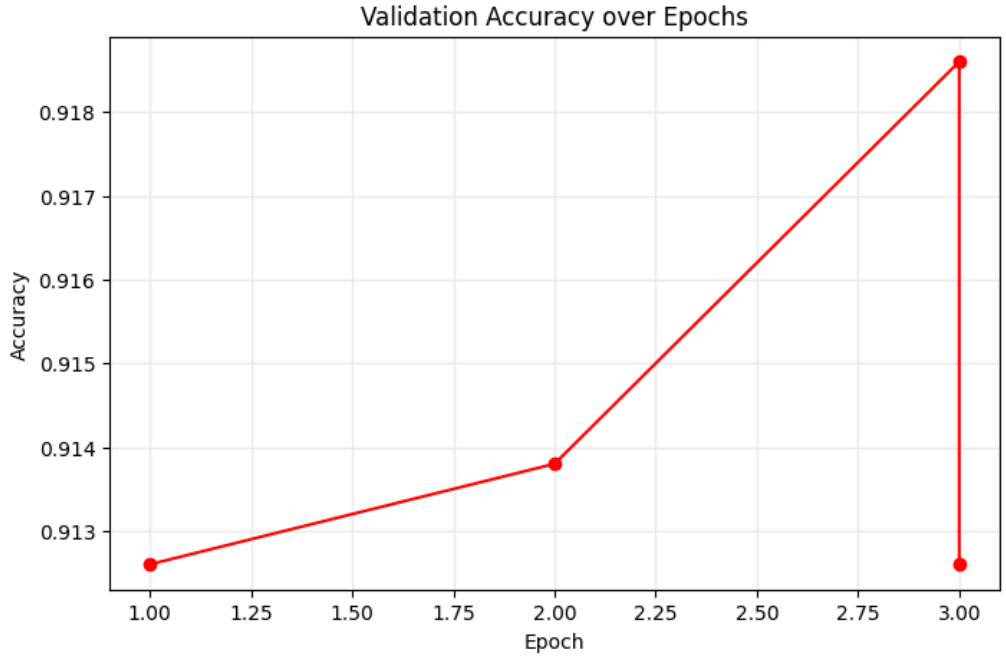
Note: Logistic Regression, and both DistilBERT models perfectly overlap.

Report on results for all the four models (statistical, DistilBERT, finetuned DISTILBERT, GPT) on the three test cases using GAICO. Use statistical model's performance as the baseline:

Based on this graph, both DistilBERTs matched the baseline (logistic regression) across all metrics. GPT-2 was behind in all metrics. So overall, logistic regression and the two DistilBERT models performed the strongest whereas GPT-2 was the only outlier.

Accuracy and Loss Curves:

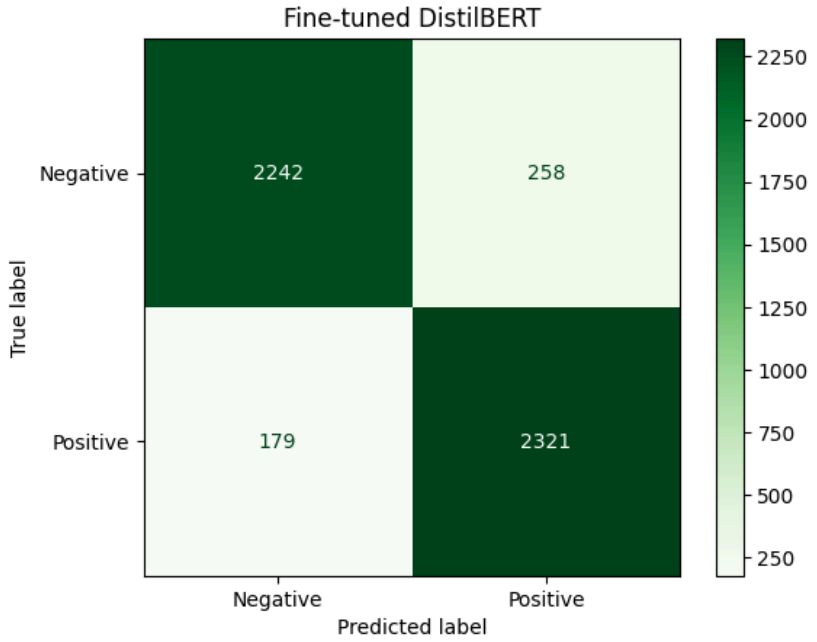


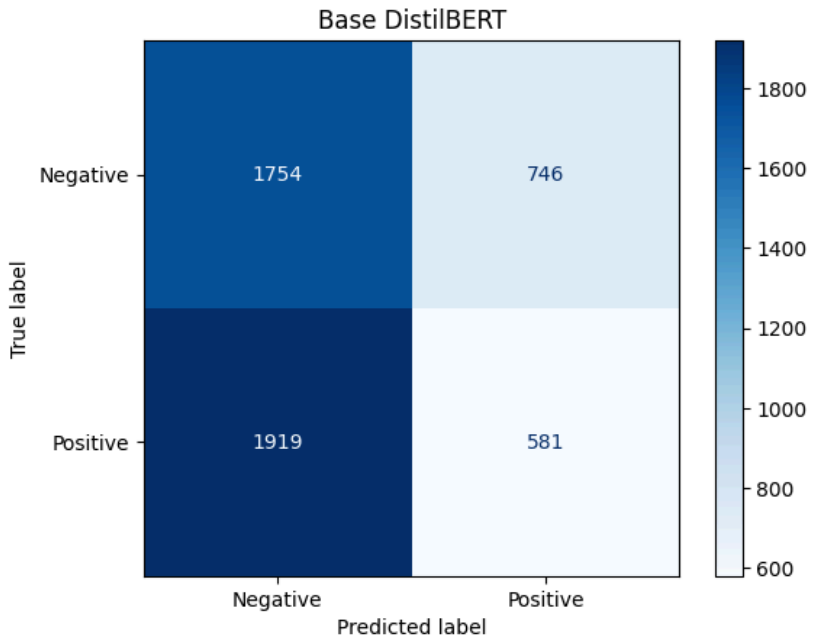


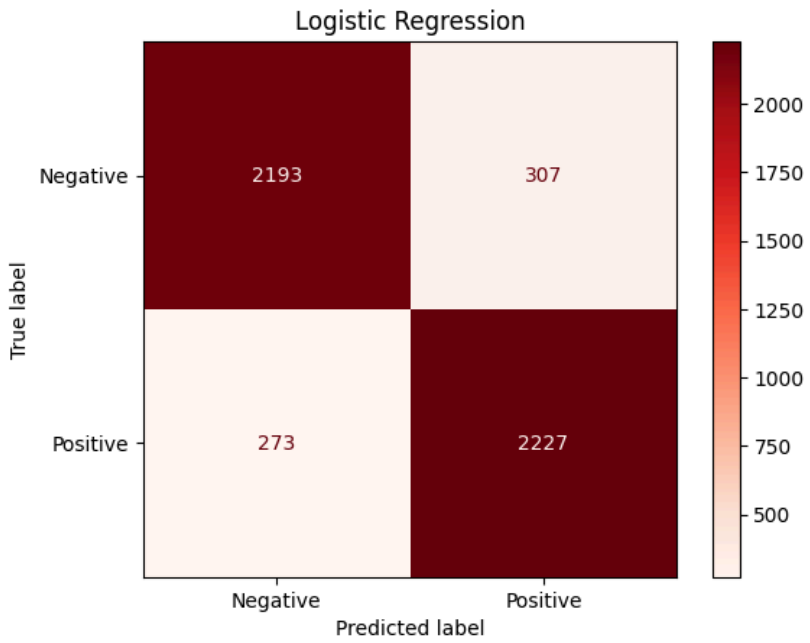
Comment on how the model is learning and any potential overfitting or underfitting:

The model learns well during the initial epochs. This is shown by the decreasing training loss and the increasing validation accuracy over the first two epochs. After one epoch the validation loss diverges while training loss continues to decrease, indicating overfitting.

Confusion Matrices:







Discuss misclassifications (e.g., false positives, false negatives) and possible reasons:

Misclassifications were relatively low for fine-tuned DistilBERT and the logistic regression model. Both models had slightly higher false positives, indicating that both were more likely to predict a negative review as positive. One reason for this could be a negative review that used heavy sarcasm. For example: “I loved how the movie dragged out the most boring point of the story for almost an hour, instead of getting to the point.” In this case the model may predict the review to be positive due to the wording of the review.

The pre-trained or base distilBERT had very random results, it was very good at predicting true negatives but also had a very high number of false negatives. This indicates that the base model may be biased towards predicting negative cases. A reason for this could be that the model has a high decision threshold in determining what is a positive review, or due to the lack of training it hasn’t learned what makes a review positive.

Precision, Recall, and F1-Score:

Report and compare the precision, recall, and F1-scores of all models:

Fine tuned DistilBERT performed the best overall, with the classical logistic regression model not too far behind. The base model of DistilBERT performed poorly with GPT-2 performing the worst of the 4 models by a large margin. The difference in scores between the two DistilBERT models indicate that fine-tuning can significantly improve a model’s capability. The performance of the logistic regression model shows that a classical model may be preferred in some situations due to its efficiency as seen in its time complexity, at the cost of some accuracy.

Performance Comparison:

Model	**Precision**	**Recall**	**F1**
Fine-tuned DistilBERT	0.8999	0.9284	0.9139
Pre-trained DistilBERT	0.4378	0.2324	0.3036
GPT-2	0.5070	0.0716	0.1254
Logistic Regression	0.8788	0.8908	0.8847

Time Complexity:

Model	**Training Time**	**Inference Time**
Fine-tuned DistilBERT	21:45	00:35
Pre-trained DistilBERT	N/A	00:36
GPT-2	N/A	02:58
Logistic Regression	~5s	~0s

Discuss and compare:

The training times for the models that required training varied significantly. The time to train the fine-tuned DistilBERT model took 21:45 minutes whereas the logistic regression model only took about 5 seconds to train. The inference times were also significantly different, but relative to each other. The two DistilBERT models took about the same amount of time, with GPT-2 and logistic regression taking almost 3 minutes and no time at all respectively. The most efficient model in terms of time and resources is logistic regression in this comparison. Logistic regression on average achieved 96% of the fine tuned DistilBERT model's performance in the three performance metrics, and took little to no time to train and make inferences.

Questions:

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

The accuracy and loss curves tell you how well a model is learning. Loss curves decreasing steadily indicate that the model is learning well, but a divergence in these curves may indicate overfitting or underfitting. As for the accuracy curve, an increasing curve indicates successful learning by the model. Low accuracy may indicate the model isn't learning effectively.

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 outperformed the classical ML model by a small margin. On average, the logistic regression model achieved 96% of the fine-tuned DistilBERT's performance across the three performance metrics. It also took little to no time to train or produce inferences.

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

The confusion matrices for the classical model and fine-tuned DistilBERT were very similar. They both had a slightly higher number of false positives. As mentioned in my previous answer, one reason could be due to reviews that use heavy sarcasm, where the model is not able to interpret this sarcasm and label a negative review as a positive one.

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

The fine-tuned model outperforms the base model because the training allows the fine-tuned model to adapt to the specific task and dataset being evaluated. While the base model has a general language understanding, the fine-tuned one is specialized specifically for the movie reviews dataset. As a result, it has much higher performance metrics across the board.

5. Which model would you recommend for deployment in a real-world scenario, and why? Consider both performance and efficiency in your answer.

I would recommend the logistic regression model in this case, due to its much higher efficiency yet comparable performance to the fine-tuned DistilBERT model. It achieved 96% of the fine-tuned DistilBERT's performance with significantly lower training and inference times. In a real world scenario the lower overhead costs would make a huge difference in deploying and maintaining such a model.