Questions:

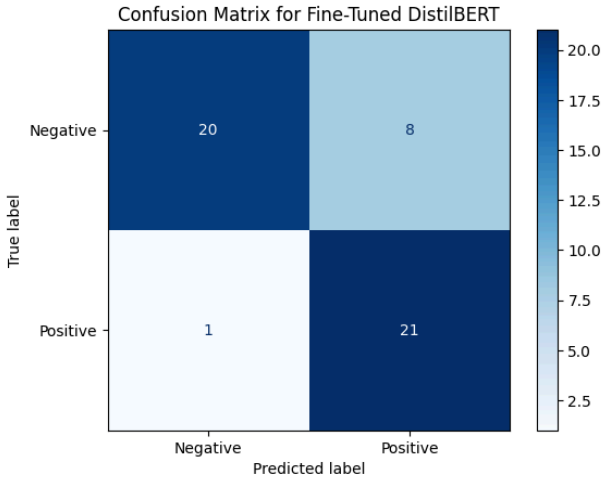
1. What do the accuracy and loss curves tell you about the fine-tuning process?

They show the performance and behavior of the model during the fine-tuning process. The accuracy shows how accurate the model is through the changes over time during training. The loss curve tracks the difference between the model's predictions and the real values. The lower the loss then the more the model has learned.

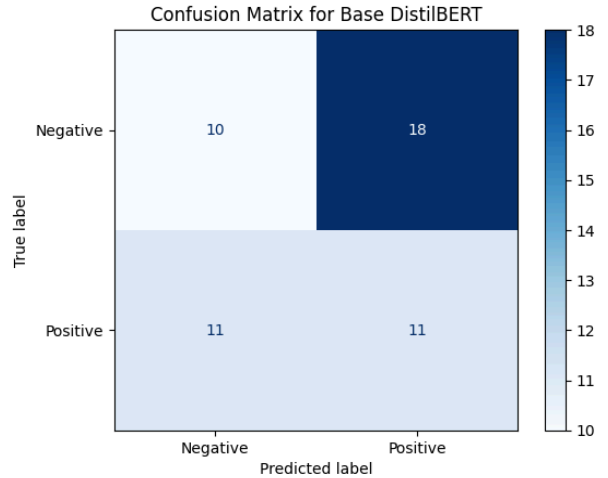
2. How does the fine-tuned DistilBERT model compare to the classical ML model? What advantages or limitations do transformers present over classical algorithms?

The fine-tuned DistilBERT often delivers better performance than the classical ML model. The advantages of transformers is that they are specifically designed to handle sequential and context-dependent data. They also handle generalizations better and more complex patterns than classical ML models. The downsides is that they take a significant amount of resources to train and a large amount of training data to be trained on.

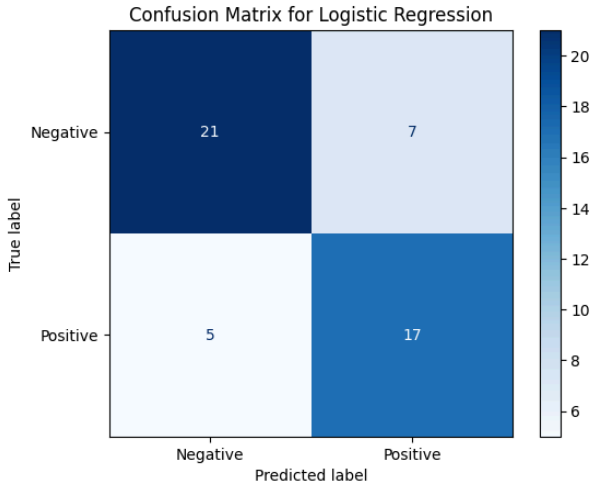
3. What insights can you draw from the confusion matrix? Are there any patterns in the misclassifications?



The fine-tuned confusion matrix performs well in both True Positives (TP = 21) and True Negatives (TN = 20). It also had very few False Positives (FP = 8) and False Negatives (FN = 1).



The base confusion matrix performs much worse than the fine-tuned model with True Positives (TP = 11) and True Negatives (TN = 10). It had a fair number of False Positives (FP =18) and False Negatives (FN = 11).



The Logistic Regression confusion matrix performs better than the base model but worse than the fine tuned model with True Positives (TP = 17) and True Negatives (TN = 21). It had a fair number of False Positives (FP =7) and False Negatives (FN = 5).

Patterns in Misclassifications:

Fine-Tuned DistilBERT:

The misclassifications (8 false positives) are more likely to occur when the model confuses a negative class for a positive one. Given that there's only 1 false negative, the fine-tuned model is good at detecting positive samples, showing its effectiveness after task-specific fine-tuning.

Base DistilBERT:

The base model has significant misclassification issues, with a high number of both false positives and false negatives. This indicates the model isn't effectively differentiating between the two classes, most likely due to its lack of fine-tuning for the specific task.

Logistic Regression:

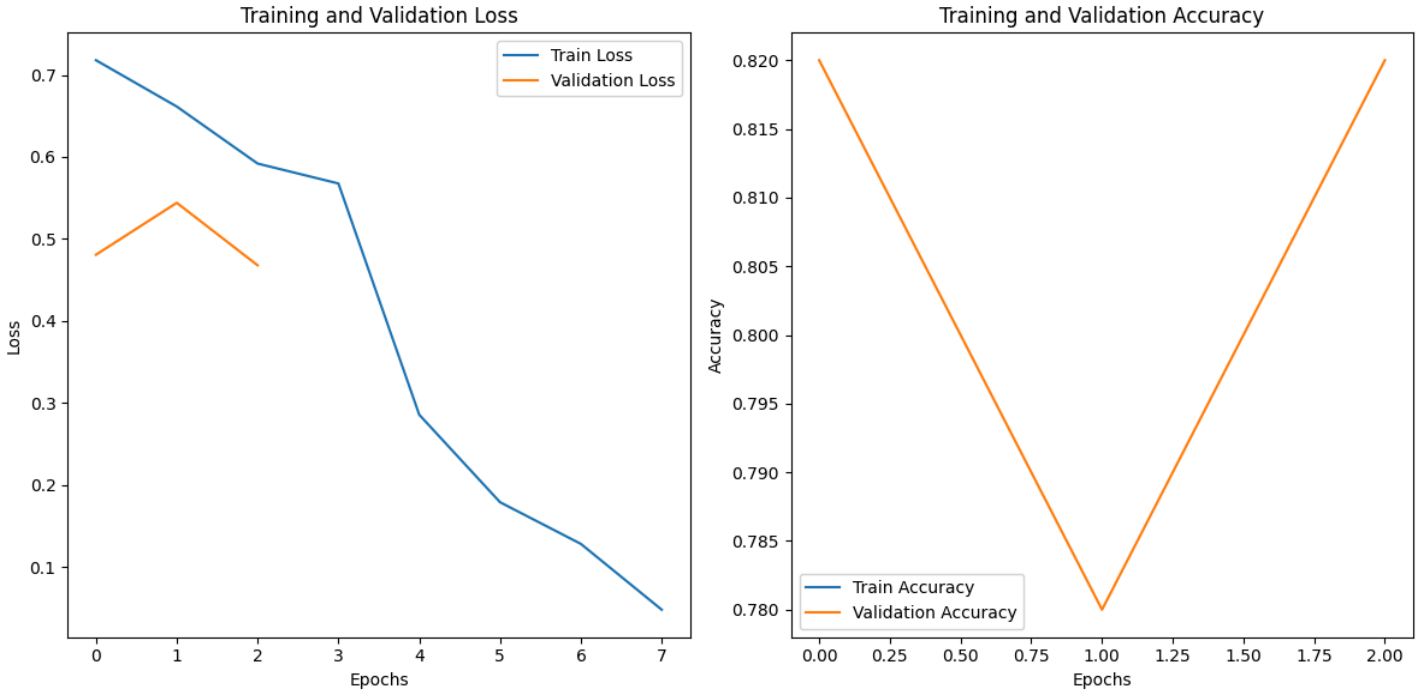
The logistic regression model confuses positive samples with negative ones (5 false negatives), but performs well in predicting the negative class. This suggests that the model might be more conservative in predicting positive sentiment, as it's more likely to assign negative labels.

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

A fine-tuned model is trained on a task-specific dataset whereas the base model is trained on a large, general-purpose set to learn language patterns. Although the base model can understand language solidly, it does not have as much context for the specific scenario that is being worked on.

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

Each model has its uses in a real world scenario, so in different situations I would use a different model. If the situation is short notice, or extremely generalized and with an excessively limited available dataset to train on, it might be better to use the base DistilBERT model since it can run more efficiently without needing a dataset to train on. If the situation permits the time, I would choose to use the fine-tuned model since it would result in the best performance when deployed long term.



Accuracy and Loss Curves:
The model is improving loss and it's neither overfitting or underfitting, which means that it is balanced in its training.

Precision, Recall, and F1-Score:

```
Logistic Regression Metrics:

{'Negative': {'precision': 0.8076923076923077, 'recall': 0.75, 'f1-score':
0.7777777777777778, 'support': 28.0}, 'Positive': {'precision': 0.7083333333333334,
'recall': 0.7727272727272727, 'f1-score': 0.7391304347826086, 'support': 22.0},
'accuracy': 0.76, 'macro avg': {'precision': 0.7580128205128205, 'recall':
0.7613636363636364, 'f1-score': 0.7584541062801933, 'support': 50.0}, 'weighted avg':
{'precision': 0.763974358974359, 'recall': 0.76, 'f1-score': 0.7607729468599034,
'support': 50.0}}

Base DistilBERT Metrics:

{'Negative': {'precision': 0.47619047619047616, 'recall': 0.35714285714285715,
'f1-score': 0.40816326530612246, 'support': 28.0}, 'Positive': {'precision':
0.3793103448275862, 'recall': 0.5, 'f1-score': 0.43137254901960786, 'support': 22.0},
'accuracy': 0.42, 'macro avg': {'precision': 0.4277504105090312, 'recall':
```

0.4285714285714286, 'f1-score': 0.41976790716286516, 'support': 50.0}, 'weighted avg': {'precision': 0.4335632183908046, 'recall': 0.42, 'f1-score': 0.418375350140056, 'support': 50.0}}

Fine-tuned DistilBERT Metrics:

{'Negative': {'precision': 0.9523809523809523, 'recall': 0.7142857142857143, 'f1-score': 0.8163265306122449, 'support': 28.0}, 'Positive': {'precision': 0.7241379310344828, 'recall': 0.9545454545454546, 'f1-score': 0.8235294117647058, 'support': 22.0}, 'accuracy': 0.82, 'macro avg': {'precision': 0.8382594417077176, 'recall': 0.8344155844155845, 'f1-score': 0.8199279711884754, 'support': 50.0}, 'weighted avg': {'precision': 0.8519540229885056, 'recall': 0.82, 'f1-score': 0.8194957983193276, 'support': 50.0}}

Model Comparison:

Model	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Avg)
Logistic Regression	0.76	0.7083	0.7727	0.7391
Base DistilBERT	0.42	0.3793	0.5	0.4314
Fine-tuned DistilBERT	0.82	0.7241	0.9545	0.8235

Time Complexity:

Fine-Tuned: The time taken for training the fine-tuned model was fairly short given the increase in performance boost. Even with training on a small dataset of 200, it was able to have an increased precision of 8% over the next closest (Logistic Regression). The inference time however, is slower than a simpler model like logistic regression, making it take longer as more and more data is processed. It also requires a significant amount of memory and compute power to train.

Base Model: Since it is a base model, there was no training required; however, it does not make up for it in inference time since the model will be similar to the fine tuned model. It will also be less resource-efficient than the fine-tuned model and logistic regression.

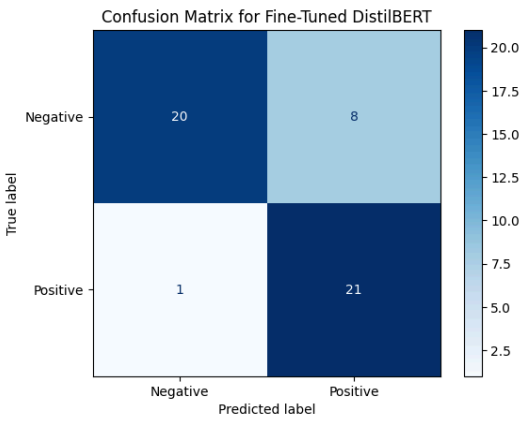
Logistic Regression: Since it is a simpler algorithm, it requires significantly less time training compared to DistilBERT. Its inference time is faster than both DistilBERT models due to this simplicity and is thusly the most resource-efficient model among the three.

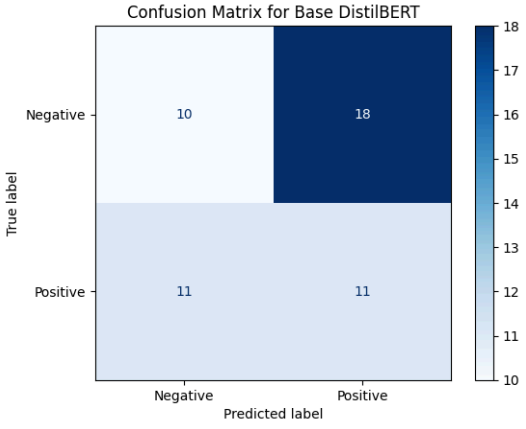
Results:

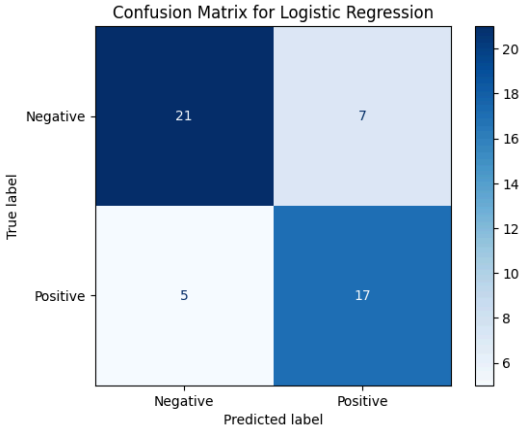
Epoch	Training Loss	Validation Loss	Accuracy
1	0.591900	0.480761	0.820000
2	0.179000	0.544055	0.780000
3	0.048000	0.467841	0.820000

```
Base DistilBERT Accuracy: 0.5600
Fine-tuned DistilBERT Accuracy: 0.8200
Fine-tuned DistilBERT performs better than the base model.
Improvement: 0.2600
```

```
Logistic Regression Accuracy: 0.7600
Base DistilBERT Accuracy: 0.4200
Fine-tuned DistilBERT Accuracy: 0.8200
The fine-tuned DistilBERT model performs the best.
```








```
Logistic Regression Metrics:
```

```
{'Negative': {'precision': 0.8076923076923077, 'recall': 0.75, 'f1-score':
0.7777777777777778, 'support': 28.0}, 'Positive': {'precision': 0.7083333333333334,
'recall': 0.7727272727272727, 'f1-score': 0.7391304347826086, 'support': 22.0},
'accuracy': 0.76, 'macro avg': {'precision': 0.7580128205128205, 'recall':
0.7613636363636364, 'f1-score': 0.7584541062801933, 'support': 50.0}, 'weighted avg':
{'precision': 0.763974358974359, 'recall': 0.76, 'f1-score': 0.7607729468599034,
'support': 50.0}}
```

```
Base DistilBERT Metrics:
```

```
{'Negative': {'precision': 0.4761904761904762, 'recall': 0.35714285714285715,
'f1-score': 0.40816326530612246, 'support': 28.0}, 'Positive': {'precision':
0.3793103448275862, 'recall': 0.5, 'f1-score': 0.43137254901960786, 'support': 22.0},
'accuracy': 0.42, 'macro avg': {'precision': 0.4277504105090312, 'recall':
0.4285714285714286, 'f1-score': 0.41976790716286516, 'support': 50.0}, 'weighted avg':
```

{'precision': 0.4335632183908046, 'recall': 0.42, 'f1-score': 0.418375350140056, 'support': 50.0}}

Fine-tuned DistilBERT Metrics:

{'Negative': {'precision': 0.9523809523809523, 'recall': 0.7142857142857143, 'f1-score': 0.8163265306122449, 'support': 28.0}, 'Positive': {'precision': 0.7241379310344828, 'recall': 0.9545454545454546, 'f1-score': 0.8235294117647058, 'support': 22.0}, 'accuracy': 0.82, 'macro avg': {'precision': 0.8382594417077176, 'recall': 0.8344155844155845, 'f1-score': 0.8199279711884754, 'support': 50.0}, 'weighted avg': {'precision': 0.8519540229885056, 'recall': 0.82, 'f1-score': 0.8194957983193276, 'support': 50.0}}

Model Comparison:

Model	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Avg)
Logistic Regression	0.76	0.7083	0.7727	0.7391
Base DistilBERT	0.42	0.3793	0.5	0.4314
Fine-tuned DistilBERT	0.82	0.7241	0.9545	0.8235