NLP Assignment 4 Report

Sentiment Analysis Model

Classification Report :

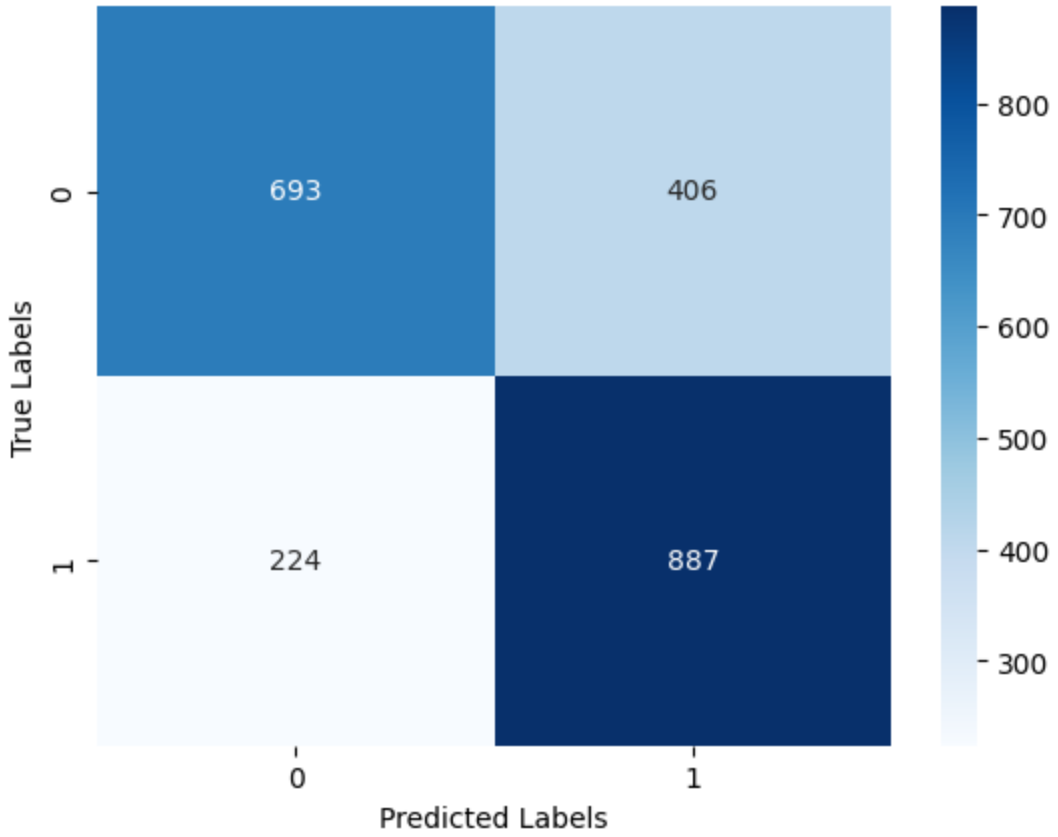
```
                precision    recall  f1-score   support

    Negative       0.76      0.63      0.69      1099
    Positive       0.69      0.80      0.74      1111

    accuracy                           0.71      2210
   macro avg       0.72      0.71      0.71      2210
weighted avg       0.72      0.71      0.71      2210
```

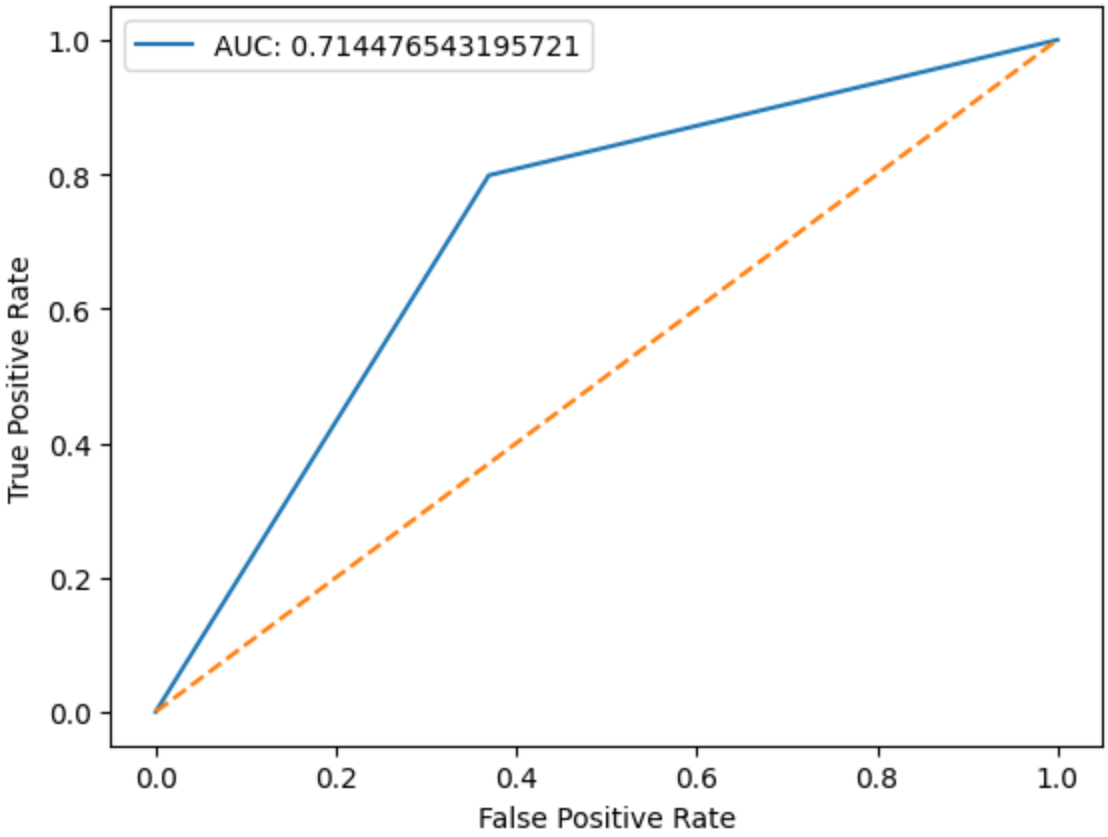
From this classification report, we see that the average accuracy is around 71% . This is fairly satisfactory, given that the training data is not huge. Recall is a measure of how many of the positive cases the classifier correctly predicted, and it is slightly lower for Negative sentences, indicating that the classifier sometimes cannot recognize a negative sentence. F1 score is a measure that combines both recall and precision, and the values obtained here are fairly satisfactory.

Confusion Matrix :



This is the confusion matrix for the same. We can see that the correct predictions are dominant. Out of incorrect classifications, negative being classified as positive is common, as we saw with the low recall value of negative.

ROC Curve :



The above is the ROC curve for this model. The orange line indicates a bad random classifier. The blue line is the ROC curve, and it is fairly good. A perfect ROC curve is extremely bent.

NLI Classifier Model

The NLI dataset is huge, and due to lack of compute, I was only able to train for 10 epochs.

```
Epoch: 1, Train Loss: 1.060, Train Accuracy: 0.406
Epoch: 2, Train Loss: 0.980, Train Accuracy: 0.486
Epoch: 3, Train Loss: 0.943, Train Accuracy: 0.515
Epoch: 4, Train Loss: 0.914, Train Accuracy: 0.533
Epoch: 5, Train Loss: 0.886, Train Accuracy: 0.551
Epoch: 6, Train Loss: 0.857, Train Accuracy: 0.567
Epoch: 7, Train Loss: 0.827, Train Accuracy: 0.584
Epoch: 8, Train Loss: 0.803, Train Accuracy: 0.597
Epoch: 9, Train Loss: 0.776, Train Accuracy: 0.612
Epoch: 10, Train Loss: 0.749, Train Accuracy: 0.624
```

```
Test Loss: 0.903, Test Accuracy: 0.596
```

The test accuracy for 10 epoch training stands at 0.596. This will most definitely improve with more epochs, which I could not achieve due to lack of compute.

Classification Report :

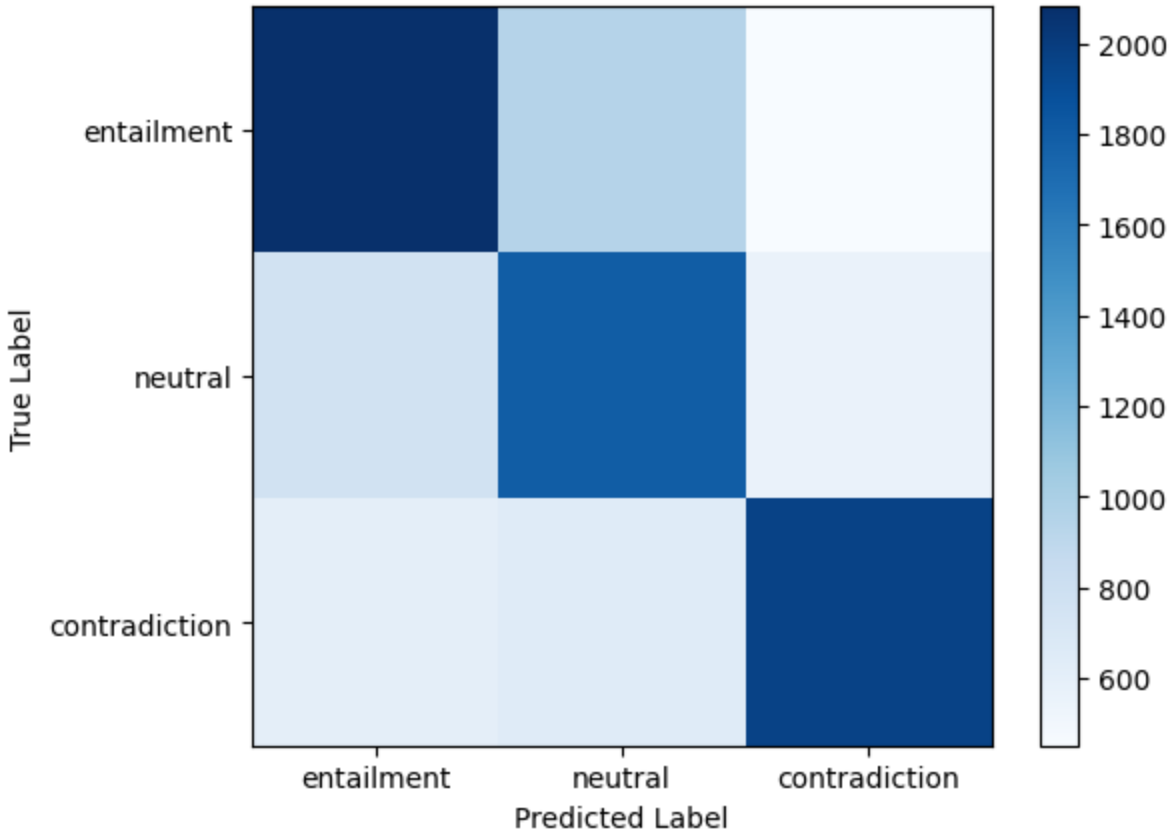
```
                precision    recall  f1-score   support

    entailment       0.60      0.60      0.60      3479
       neutral       0.53      0.58      0.55      3123
 contradiction       0.66      0.61      0.64      3213

      accuracy                           0.60      9815
     macro avg       0.60      0.60      0.60      9815
  weighted avg       0.60      0.60      0.60      9815
```

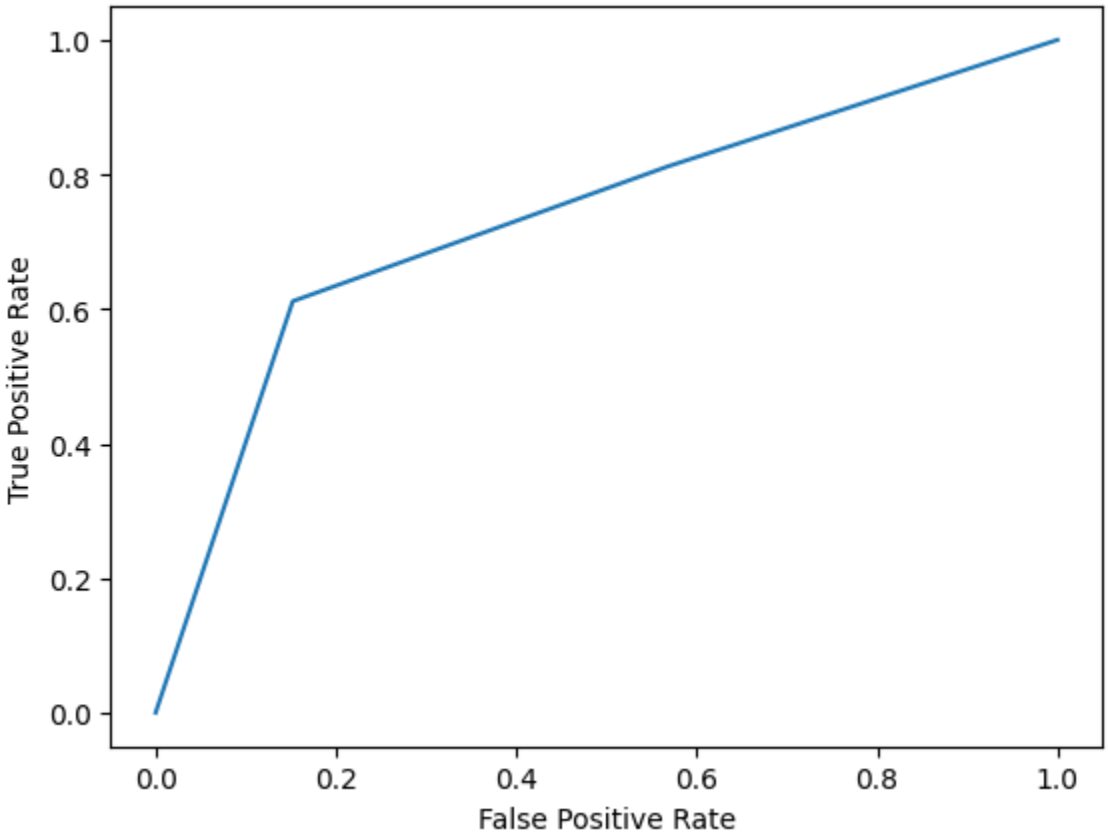
As we can see here, neutral classification has the overall worst scores in all departments. This can be attributed to the nature of the sentences, as humans also might be varied in their opinions as to whether some sentence is neutral or not. Identifying words for entailment and contradiction are present, but not necessarily for neutral. This therefore explains these scores.

Confusion Matrix :



As observed above in the classification report, wrong classification between contradiction to entailment and vice versa is low. This is clearly shown here.

ROC Curve :



As discussed before, a square ROC curve is good, and this curve is fairly satisfactory for the 10 epochs the model was trained for.