For this assignment, 2 transformer models were implemented and trained.

- Original model
    - 1 attention head
    - Dropout rate of 0.1
    - Using learned positional encoding
- Modified model
    - 4 attention heads
    - Dropout rate of 0.2 (Multiple training runs with different dropout rate parameters concluded that the ideal dropout rate is 0.2)
    - Using custom positional encoding layer (Sinusoidal position encoding)

The dataset used in training was augmented using synonym replacement strategy

```
776/776 ━━━━━━━━━━━━━━━━━━━━ 0s 378us/step
776/776 ━━━━━━━━━━━━━━━━━━━━ 2s 943us/step
                Accuracy  Precision  Recall  F1-Score  AUC-ROC
Original Model    0.8405     0.8682  0.8040    0.8349   0.9192
Modified Model    0.8343     0.8499  0.8132    0.8312   0.9170
```

Figure 1: Final evaluation metric after training both models

From figure 1, we can observe the following:

Sinusoidal Advantage: The Modified Model achieved higher **recall**, proving that mathematical positional encoding helps the model generalize across different sentence structures better than learned embeddings
Regularization Trade-off: The slightly lower overall accuracy in the Modified Model is a result of the **increased dropout** layers. While this lowers the "peak" score, it prevents overfitting, making the Modified Model more reliable for future, unseen data

Data augmentation benefits for transformer models

```
776/776 ━━━━━━━━━━━━━━━━━━━━ 3s 767us/step
776/776 ━━━━━━━━━━━━━━━━━━━━ 0s 367us/step
                Accuracy  Precision  Recall  F1-Score  AUC-ROC
Original Model    0.8455     0.8654  0.8195    0.8418   0.9246
Modified Model    0.8305     0.8458  0.8096    0.8273   0.9136
```

Figure 2: Evaluation metrics for dataset which did not undergo data augmentation

Referring to figure 2, we observe that without data augmentation, the original model was surpassing the modified model across all metrics. With data augmentation, the size of the dataset increased from 25000 to 50000. This resulted in a better performance across the metrics for the modified model. The experiment proved that transformer architectures (both original and modified) require significant data volume (augmented to 50,000 samples) to compete with simpler architectures or to realize the benefits of complex encoding