Part 1:

Question: Run the code (Word2Vec encoding and LSTM prediction) but include the stop words. How do the F-scores compare? How does the quality of the prediction compare?

Answer: After including the stop words, the F-score increased and the quality of the prediction improved. So I set the flag to use stop words to make the performance better.

Part 2 Extra Credit:

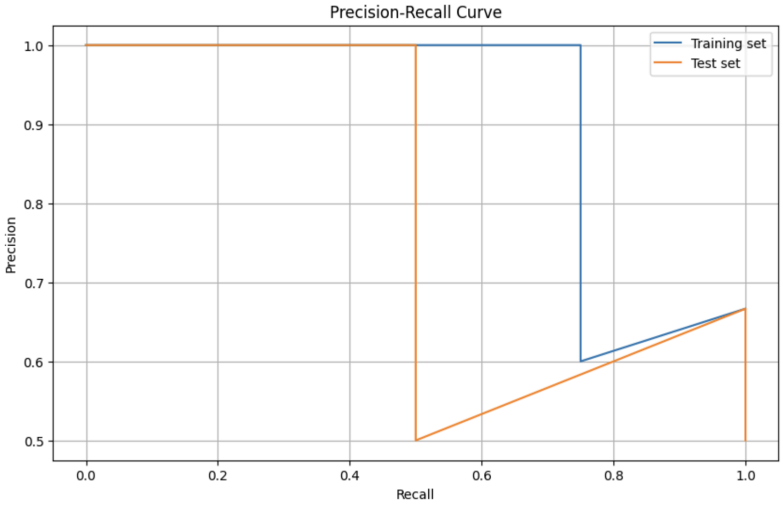
Question: Which one is better?

Answer: Though there are many BERT predictions getting around 50% cosine similarity, the overall performance of GPT2 is better than BERT.

Part 4:

Question: A PDF with a plot of the results and a one-paragraph description of what you found most interesting.

Answer:



The precision-recall curve shows a case of possible overfitting, as indicated by the high precision across all recall levels for the training set (blue line), which sharply contrasts with the performance on the test set (orange line) where precision drops significantly as recall increases. This discrepancy suggests that while the model performs exceptionally well on the data it was trained on, its ability to generalize to new, unseen data is limited. The steep decline in precision on the test set with minimal increases in recall is particularly interesting, implying a model that is quite sensitive to the classification threshold and has not balanced the precision-recall trade-off well on unseen data.