Dataset:

I used the covid tweet dataset: <https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification>. This dataset was already split into training and testing sets, so I did not need to perform splitting. However, the dataset was too large to use in its full size (Google Colab ran out of RAM and crashed), so I used a small subset of both the training and testing data. The input is the text of a tweet, while the output is a sentiment that can take on a range of values from 0 to 4.

0 – Extremely negative

1 – Negative

2 – Neutral

3 – Positive

4 – Extremely Positive

Methods Used:

Sequential:

This produced the best results, at least for the training data. However, there was a large amount of overfitting, where the accuracy on the training data was about 1 while the accuracy of the testing/validation sets hovered around 0.5.

RNN:

This produced poor results for training and testing data. The accuracy for training and testing hovered around 0.3.

LSTM:

This produced poor results for training and testing data. The accuracy for training and testing hovered around 0.2.

I suspect that the reason that my RNN and LSTM networks performed so poorly was due to the fact that I performed no stop word removal, which is generally standard for sentiment analysis tasks. I did not perform stopword splitting because I wanted to see how older neural networks might perform without preprocessing, in comparison with transformer models that I have more experience with.