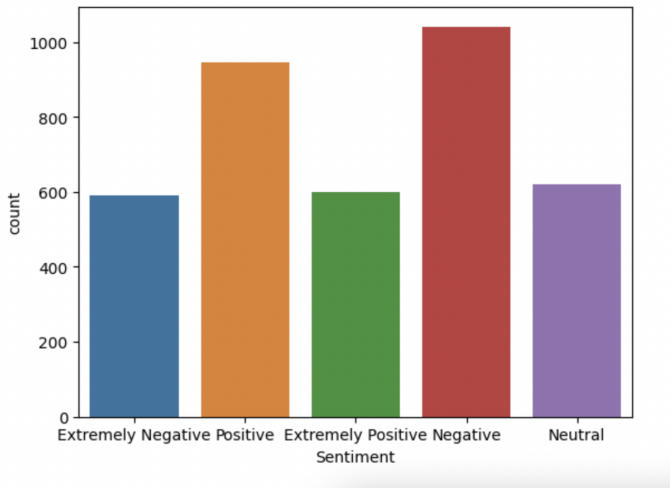
Kyle Chan kxc180021, Ryan Banafshay rcb170002

Text Classification

For this assignment, we used Kaggle datasets to experiment with text classification. The data set we've chosen classifies tweets based on their sentiment related to COVID-19 and the pandemic. Sentiment varies from extremely positive to extremely negative. The distribution of the target classes are as follows: 2464 for extremely negative tweets, 2423 for extremely positive, 4014 for negative, 2813 for neutral, and 4286 for positive. This distribution can be visualized with the following graph.



From this dataset, we created a variety of different models to predict the sentiment scores from the tweets using these multi-label classifications and then test the accuracy of these different approaches with one another. We first vectorized the tweets and set up a sequential model. After setting up this model, we saw it gave us an accuracy score of 52.5%. Seeing as this isn't an amazing accuracy score, we decided to try out different approaches to then test. First we set up a GRU (RNN) Architecture model. We trained a GRU, Gated Recurrent Unit, and used one-hot encoding to vectorize the words of each tweet. This gave us a worse accuracy output of 45.39%. From this, we continued and added an embedding layer to this same GRU model. An embedding is a vector of floating-point values of a predetermined length. Embedding representations are dense compared to sparse one-hot matrix representations. This ended up giving us our best accuracy score of 62.22%.