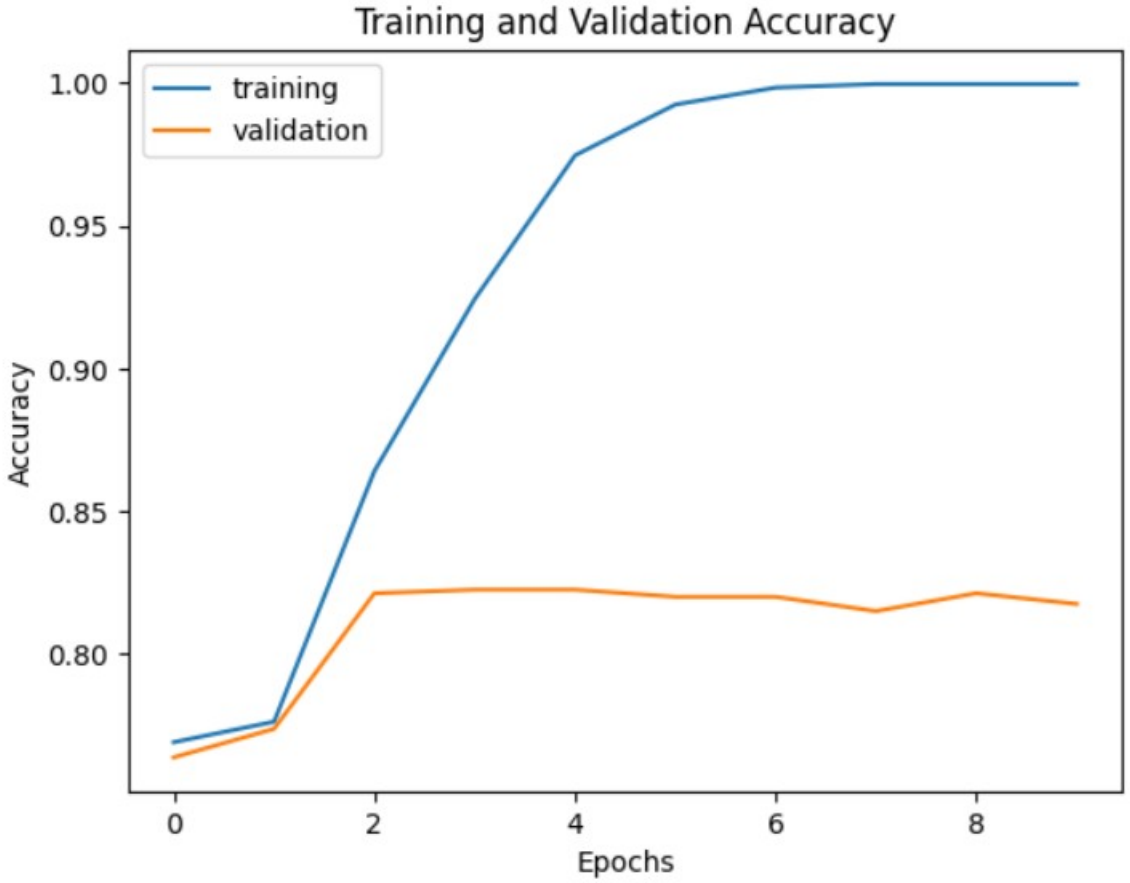
Critical analysis of NLP Sentiment Analysis AI.

The intelligence system created for this report is a Natural Language Processing system that will predict if a review has a negative, positive or neutral sentiment. It was trained and tested on a dataset of 5,000 amazon food reviews with an 80/20 split between training and test data. Data was cleaned and pre-processed, with pre-processing including cleaning, tokenization, negation handling, PoS tagging, lemmatization and stop word removal. It was then vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) before being fit to the model, compiled using adam, in batches of 32 and 10 epochs. Below is a graph showing the learning curve of the model:



(Figure 1: Learning curve of the model)

The blue training line indicates how accurately the model fits to the training data, while the validation is how accurate it is to the testing data over 10 epochs. This graph indicates the model is performing decently well since the validation is almost 0.85 in accuracy, however, this does not show the true issues with the model. Below are some further evaluative scores:

```
32/32 ━━━━━━━━━━━━━━━━━━━━ 1s 8ms/step - accuracy: 0.8320 - loss: 0.5524
Test loss: 0.5524
Test accuracy: 0.8320
F1-score (macro): 0.558
Precision (macro): 0.667
Recall (macro): 0.528
```

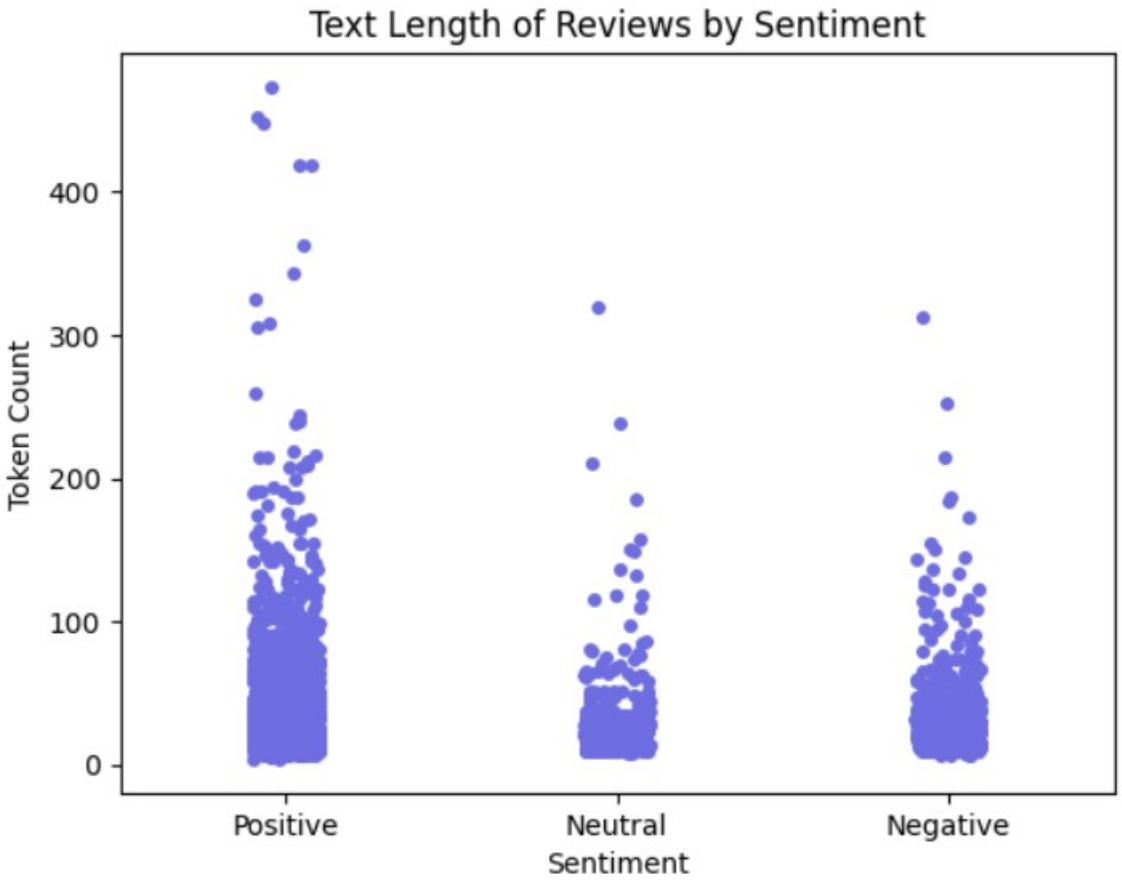
(Figure 2: Evaluation scores of the model)

Accuracy is the percentage of correct predictions the model makes, whereas loss shows the difference between predicted values and the actual values. Although the model is getting a decent accuracy score, it is getting significant loss, indicating it is not actually performing very well. The precision score shows how many of the models positive classifications (both true and false positives) are truly positive, and recall is the proportion of true positives that were correctly classified as positives. These two scores are both used to calculate the F1 score (Google, 2024). All

three of these have been calculated with a macro average to calculate the average F1 score for all the classes, disregarding class imbalance. The F1 score for this model shows the model is not getting true positive results as often as the accuracy may suggest.

The reason for these scores is the huge class imbalance in the dataset used, as explored in the EDA:

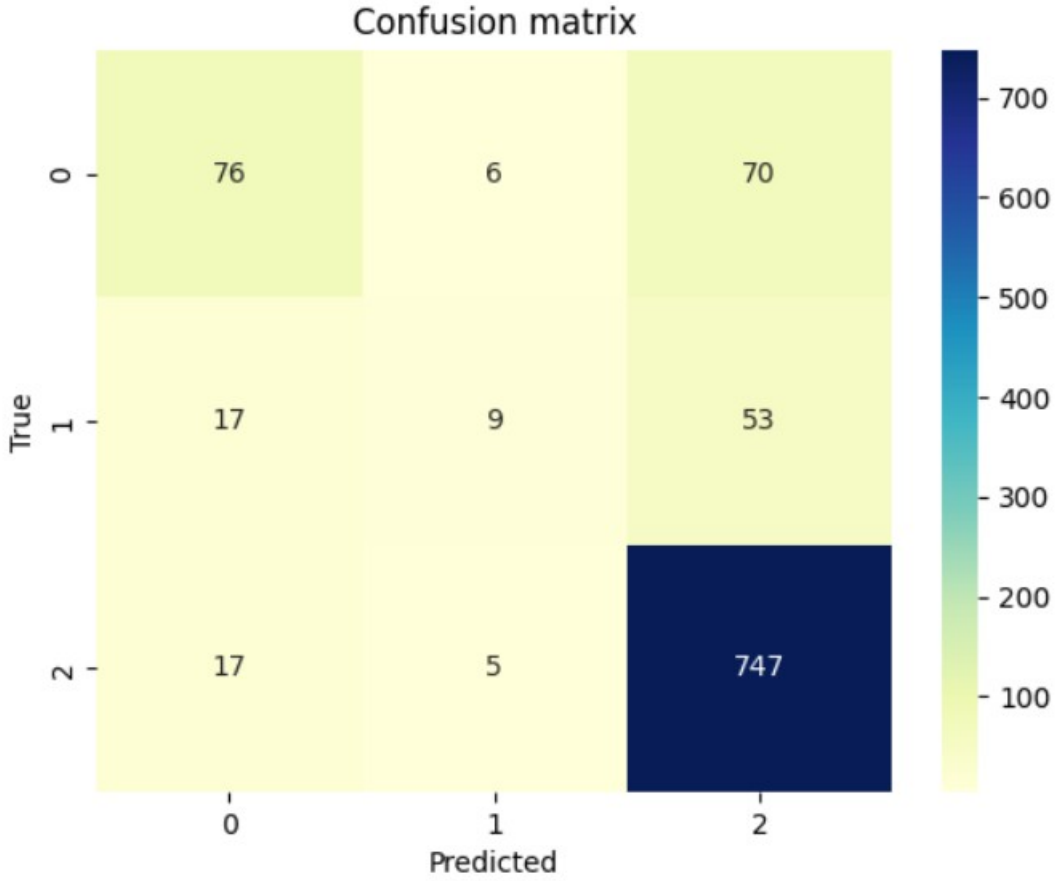
(Figure 3: Bar chart of the number of reviews in each rating in the dataset)



(Figure 4: Scatter plot of the text length of reviews in each sentiment class)

These graphs show that the positive sentiment class makes up a massive portion of the dataset, and also that positive reviews generally have the longest reviews. making it very imbalanced and biased. This means that the model does not have enough data from the neutral and

negative classes to accurately predict them, but it excels in predicting positive reviews. This huge class imbalance is the reason for the high accuracy, as the positive sentiments skew the accuracy as they are mostly predicted correctly, but the loss and F1 score show the model struggling with neutral and negative sentiments. This is further evidenced by the confusion matrix:



(Figure 5: Confusion matrix of the model)

It clearly displays that the model correctly predicts a major amount of the largest sentiment class and struggles with negative and neutral sentiments, with neutral (with the smallest amount of reviews) being the worst of them all. That is the reason that in the final demonstration in testing the model on three random reviews, the model struggled with the neutral review:

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 113ms/step
Predicted sentiment: Positive
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 62ms/step
Predicted sentiment: Positive
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 59ms/step
Predicted sentiment: Negative
```

(Figure 6: Output of final demonstration. Middle review was actually neutral, but predicted positive.)

To improve this model, a balanced dataset with an equal amount of reviews from each class should be used so that the model has ample data to draw its conclusions.