AJ Broderick
UIN - 01244170

CS 733 Module 2 Assignment: Multi-class Sentiment Analysis (SA)

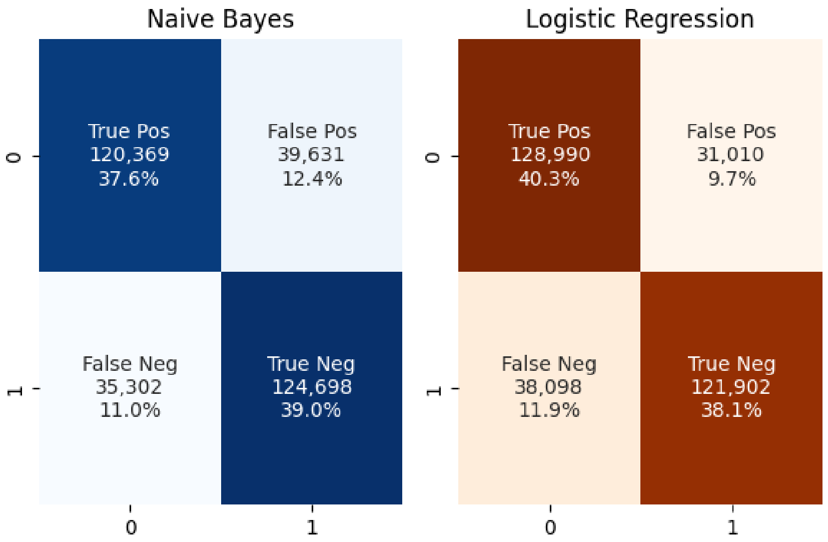
This analysis was performed to attempt to train a Naïve Bayes and Logistic Regression model against a set of 1.6million tweets to determine it's overall sentiment. To train the models TF-DIF was chosen as the feature extraction method for the sentiment analysis. This was done over a bag of words approach since it's a little more nuanced in the way that each of the words is given a weight as opposed to simply counting the words. In a model of this size, the words that have the higher frequency will have a larger impact to the end output.

Using the TF-DIF vector, the training data was fed into the model and then applied to the test set. Afterwards classification reports of the two models is created and displayed below.

```
Naive Bayes Classification Report:

                    precision    recall  f1-score   support

                 0       0.76      0.78      0.77    160000
                 4       0.77      0.75      0.76    160000

          accuracy                           0.77    320000
         macro avg       0.77      0.77      0.77    320000
      weighted avg       0.77      0.77      0.77    320000

Logistic Regression Classification Report:

                    precision    recall  f1-score   support

                 0       0.80      0.76      0.78    160000
                 4       0.77      0.81      0.79    160000

          accuracy                           0.78    320000
         macro avg       0.78      0.78      0.78    320000
      weighted avg       0.78      0.78      0.78    320000
```

From the modeling that was done, there wasn't too much of a difference between the two models. The output from the Classification Report showed that they performed pretty much the same, with only a 1% increase in performance in the Logistic Regression model.

One thing to point out though is the difference in the true positives that were displayed in the Confusion Matrices. Whilst the two models looked to have the same performance, the Logistic Regression model did a better job at predicting the true positives. This would lend itself to being the preferred model for trying to find tweets that have an overall positive feel to them.



If the models were to be refined, the feature engineering of the dataset would need to be reworked. Given that there were so many tweets to process, the vectors that were created might have been too broad to predict the sentiment. Even with trying to remove the bottom 1/3 of values in the vector did not seem to increase the models performances. One way that it could be enhanced would be to use the TD-DIF approach combined with a dictionary of positive and negative words. This would create more impactful weights on the words that would be fed into the model.