Rotten Tomatoes - Classifier Analysis

Dataset Link: https://drive.google.com/file/d/1fLagttakoBqDl3E2aaYQNj3ALt-TXOrM/view?usp=share_link

For this analysis, we developed a Classifier that could classify a movie review as Negative or Positive based on the review text, and evaluated its performance. As a first step, we collected mutiple reviews of ~150 movies from Rotten Tomatoes, gathering a total of ~12,575 records. Based on this data, we built a Logistic Regression classifier and a Decision Tree model. Figure 1 below shows details about the review data obtained:

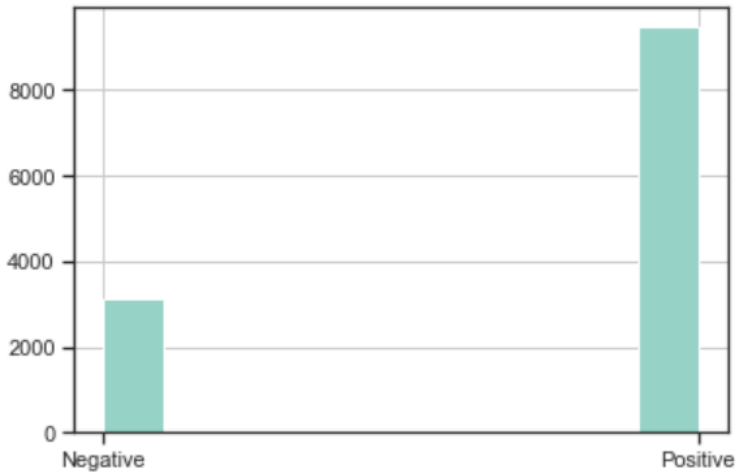

Figure 1: No of Positive and Negative Reviews

From the histogram, we can observe that the data is very skewed. Here, the number of Positive reviews is 9,466 and Negative reviews are 3,109.

Accuracy Evaluation

The Logistic Regression classifier resulted in an accuracy of 76.88%. Meanwhile, the Decision Tree model has an accuracy of 71.6%. We decided to train the Logistic Regression model with an 80/20 split and the Decision Tree model with a 70/30 split, as this helped us achieve the maximum accuracy.

The results obtained suggest that both models have decent accuracy, and overall good predictive performance. In this case, using the Logistic Regression would be the best option, as it proved to have a better performance. However, the proportion of Positive Reviews vs Negative Reviews used in the Training dataset could be impacting the models' performance.

Coincidence Matrix - Precision and Recall Analysis

To further evaluate the effectiveness of models and their predictive power, we conducted a Precision and Recall analysis on the Testing data. Figures 2 and 3 below show the classification results:

```
[[ 332  610]
 [ 476 2358]]
              precision    recall  f1-score   support

    Negative       0.41      0.35      0.38       942
    Positive       0.79      0.83      0.81      2834

    accuracy                           0.71      3776
   macro avg       0.60      0.59      0.60      3776
weighted avg       0.70      0.71      0.70      3776
```

Figure 2: Precision-Recall for **Decision tree**

```
[[ 190  416]
 [ 166 1746]]
              precision    recall  f1-score   support

    Negative       0.53      0.31      0.40       606
    Positive       0.81      0.91      0.86      1912

    accuracy                           0.77      2518
   macro avg       0.67      0.61      0.63      2518
weighted avg       0.74      0.77      0.75      2518
```

Figure 3: Precision-Recall for **Logistic Regression**

We can observe that both models perform better at predicting Positive reviews, since the Precision and Recall values for these are far greater than those for Negative reviews.

Conclusions

After building both classifiers, we can conclude that the Logistic Regression model would be the best option, as it proved to have a better performance. The accuracy of the model appears to be strong at 76%. However, to improve the overall performance of the models, a larger sample of Negative reviews could be included in the training dataset.
In addition, more details about the models' predicted values are presented in Figures 4 and 5 below:

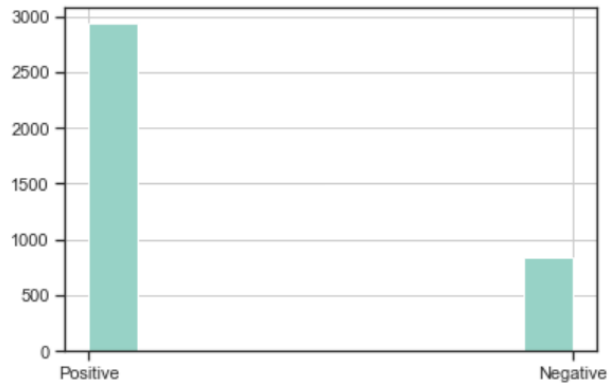

Figure 4: Histogram Chart of **Predicted values for Decision Tree**

The Histogram above shows the number of positive and negative reviews predicted by the model with a 71% accuracy. Out of 3,776 records, 2,936 are positive and 840 are negative.

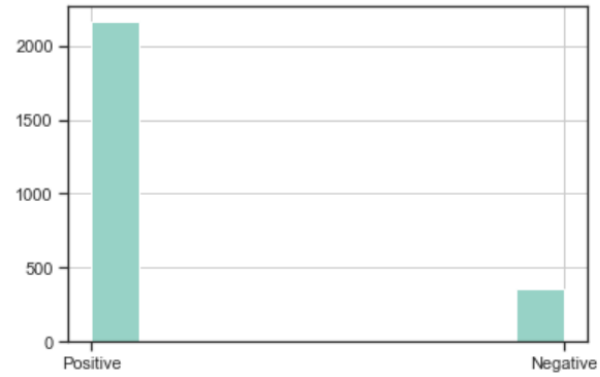

Figure 5: Histogram Chart of **Predicted values for Logistic Regression**

The Histogram above shows the number of positive and negative reviews predicted by the model with a 77% accuracy. Out of 2,518 records, 2,162 are positive and 356 are negative.