Assignment: Movie Review

Sentiment Analysis on Movie Review

By:- Amrendra Singh Rathore

Problem Inspection

The dataset is a set of movie reviews which is similar to what data analysts at IMDB handle. The data belongs to/taken from IMDB, having 2,000 reviews and a sentiment corresponding to each review.

The main aim of this project is to identify the underlying sentiment of a movie review on the basis of its textual information. In this project, we try to classify whether a person liked the movie or not based on the review they give for the movie

Data Exploration

Data Exploration is to understand the dataset by checking the general info about the data(i.e. data types, columns, missing values, number of values)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   Review     2000 non-null   object
 1   Sentiment  2000 non-null   object
dtypes: object(2)
memory usage: 31.4+ KB
```

Pre-Processing

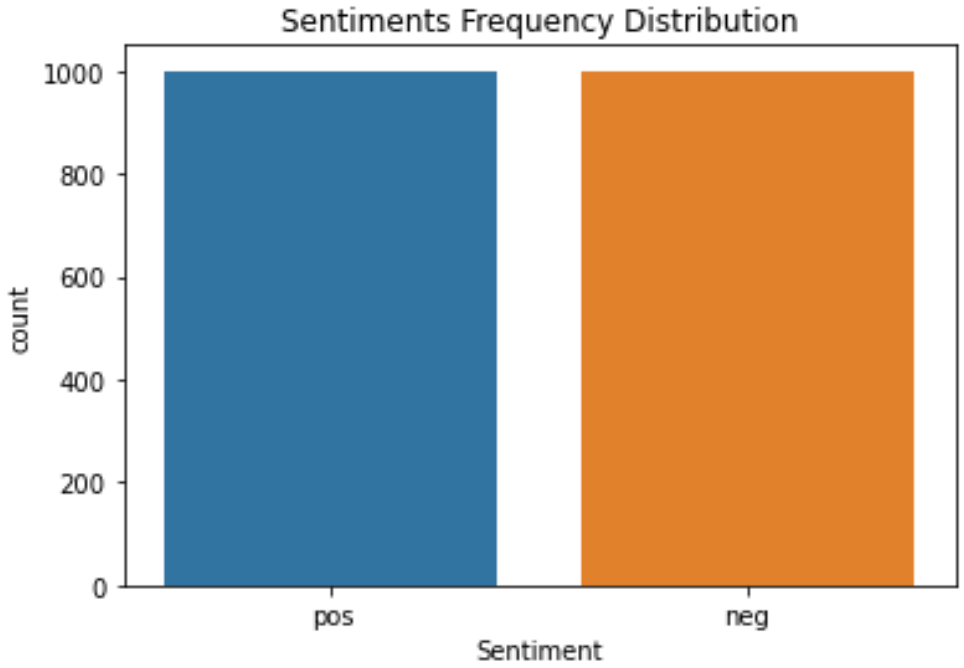
For preprocessing the data I have done following things:

- Removed noise from the data
- Stemming
- Removed stop words
- TF-IDF Vectorization

by creating the functions.

Data Visualization

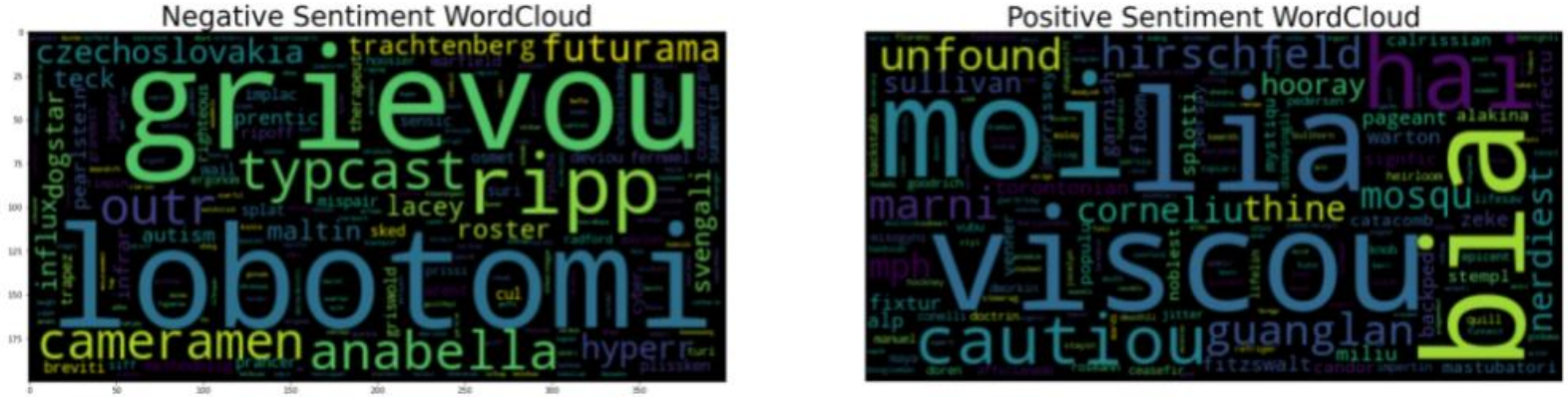
- At first I have checked for the frequency of labels, which is found that there is equal frequency of both positive and negative labels.



- After which using Wordcloud I have checked for high frequency words before and after removing the common words.

Before

After



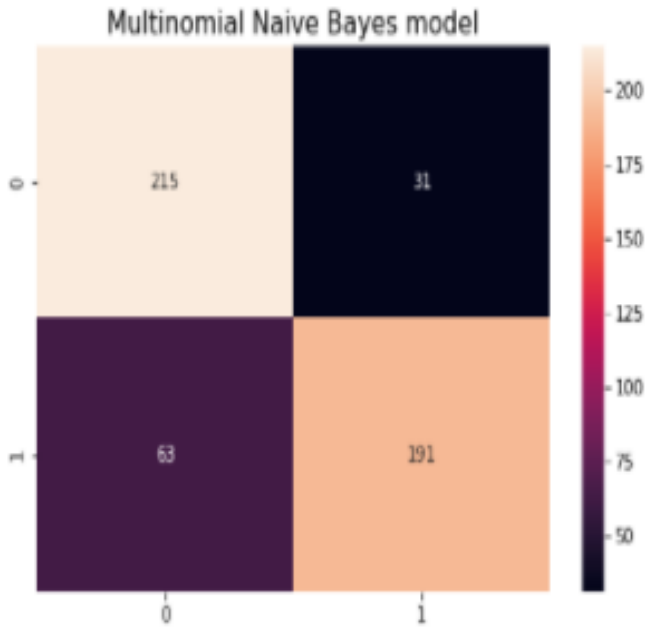
Model

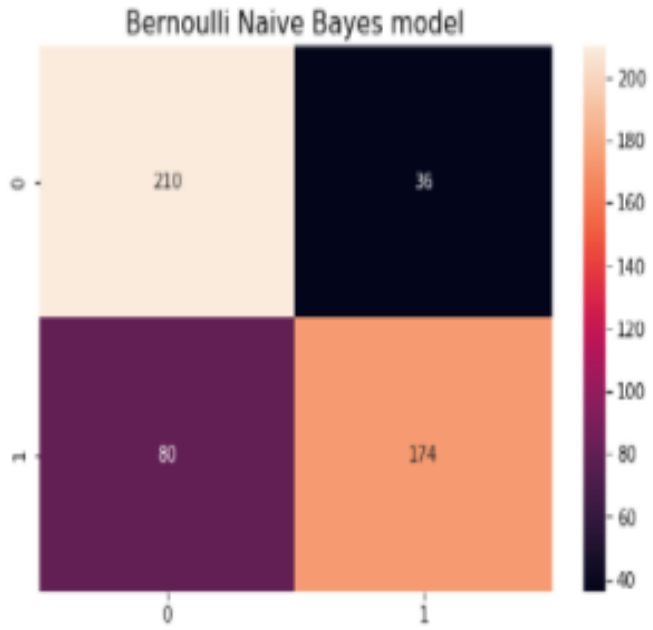
For model selection I have used 3 classification algorithm, they are

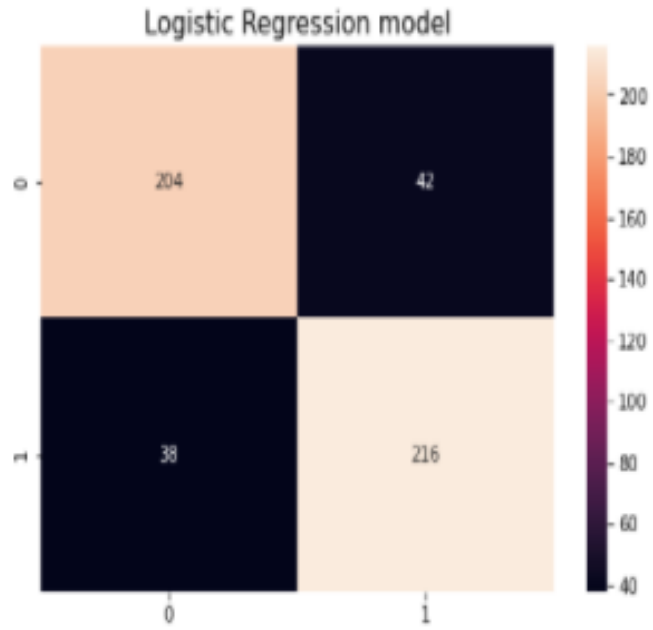
- Multinomial Naive Bayes
- Bernoulli Naive Bayes
- Logistic Regression

```
Accuracy of Multinomial Naive Bayes Model:
0.812
Accuracy of Bernoulli Naive Bayes Model:
0.768
Accuracy of Logistic Regression Model:
0.84
Mutinomial Naive Bayes model:
              precision    recall  f1-score   support

         neg       0.77      0.87      0.82       246
         pos       0.86      0.75      0.80       254

    accuracy                           0.81       500
   macro avg       0.82      0.81      0.81       500
weighted avg       0.82      0.81      0.81       500

Bernoulli Naive Bayes model:
              precision    recall  f1-score   support

         neg       0.72      0.85      0.78       246
         pos       0.83      0.69      0.75       254

    accuracy                           0.77       500
   macro avg       0.78      0.77      0.77       500
weighted avg       0.78      0.77      0.77       500

Logistic Regression model:
              precision    recall  f1-score   support

         neg       0.84      0.83      0.84       246
         pos       0.84      0.85      0.84       254

    accuracy                           0.84       500
   macro avg       0.84      0.84      0.84       500
weighted avg       0.84      0.84      0.84       500
```

Confusion Matrix







Conclusion

We can observed that all three Logistic Regression, Bernoulli Naive Bayes and Multinomial Naive Bayes models are performing well but in comparison Logistic Regression model works better among them.

Still we can improve the accuracy of the models by preprocessing data and by using lexicon models like Textblob.

THANK YOU