Dataset:

https://drive.google.com/file/d/1mijMGQQu5Cb26QqtRIxb2wYDzAv1f6CU/view?usp=sharing

The above-mentioned dataset has more than 3.2 Lakh IMDB ratings and reviews. The code implements four different algorithms, namely

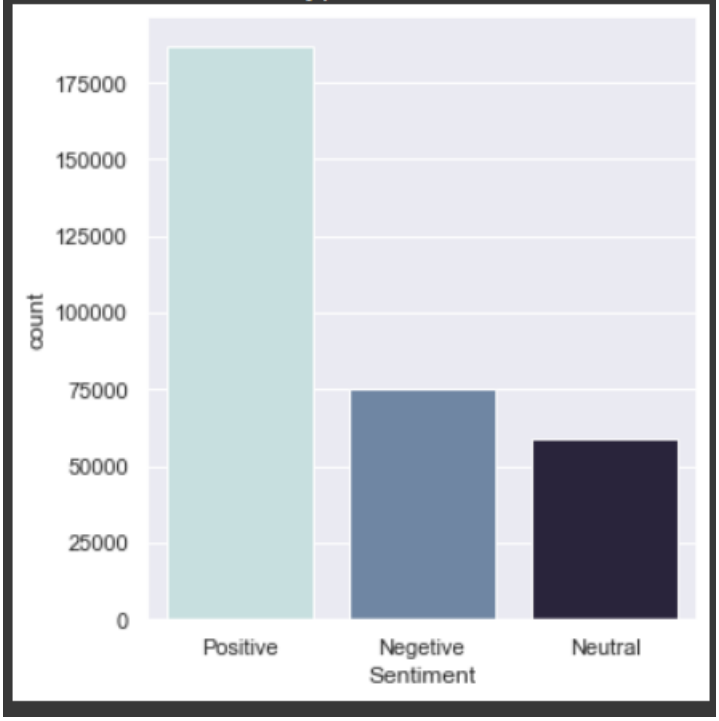
- Naïve Bayes Classifier
- Stochastic Gradient Descent – SGD Classifier
- Random Forest Classifier
- Logistic Regression

The goal is to compare the test accuracies achieved by each of the above-mentioned algorithms for our given dataset. Below are the results achieved:

	Model	Test accuracy
3	Stochastic Gradient Decent	0.739205
0	Logistic Regression	0.739034
2	Naive Bayes	0.691863
1	Random Forest	0.682993

The dataset consists of 3.2 Lakh comments, with the following data distribution:

```
Positive    186757
Negetive     75030
Neutral      58960
Name: Sentiment, dtype: int64
```



The reviews data is pre-processed using the following steps, making it fit for the algorithms to run on:

- Tokenization
- Stemming
- Feature Extraction
- Removing Stop Words

The dataset is split into a 8:2 Training-Testing ratio.

Output:
Apart from the categorization, the model also provides with a word cloud and a graph count of the positive and negative words used in descending order of their frequencies.

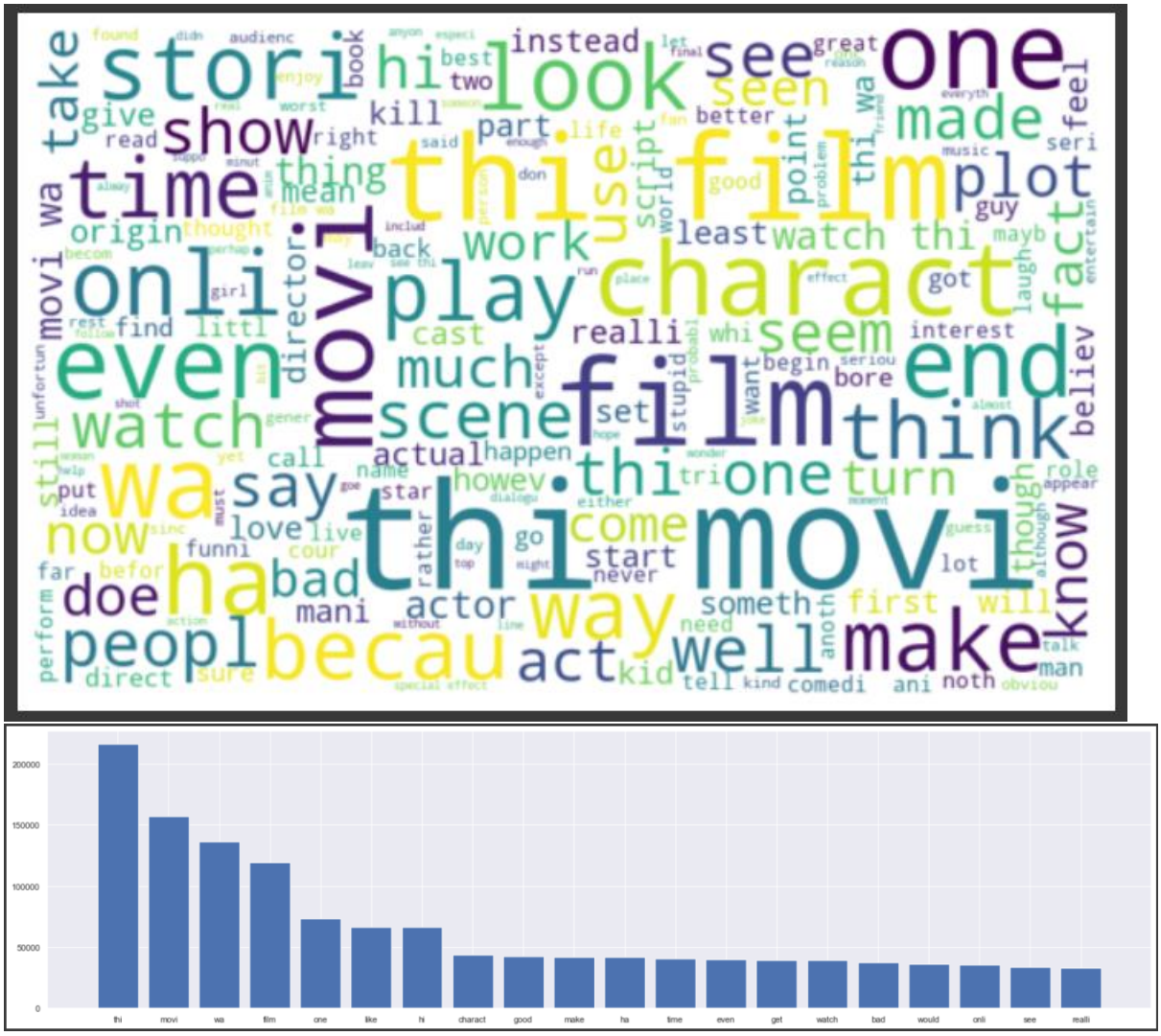


Figure 1: Negative Words Preview

Figure 2: Positive Words Preview