Movie Reviews Sentiment Analysis

Phase 4 Project

AVI PATEL

Objective

This project seeks to create a model that classifies Sentiment of a given movie review is Positive or Negative . This information will help Movie Production House to analyze reviews from viewers around various social networking platform.

Business Problem

Movie Production House can get Average user rating of their movies from websites like IMDb, Google Reviews and Yelp. But to get the Overall Sentiment and Word Of Mouth from social networking sites can be quite painful as reviews are available in the form of text and not ratings (numbers). To solve this problem Machine Learning models that analyze the sentiment of a given text review can be helpful.

The Data

The Movie Reviews having ratings (1,2,3,8,9,10) were web scraped from IMDb.

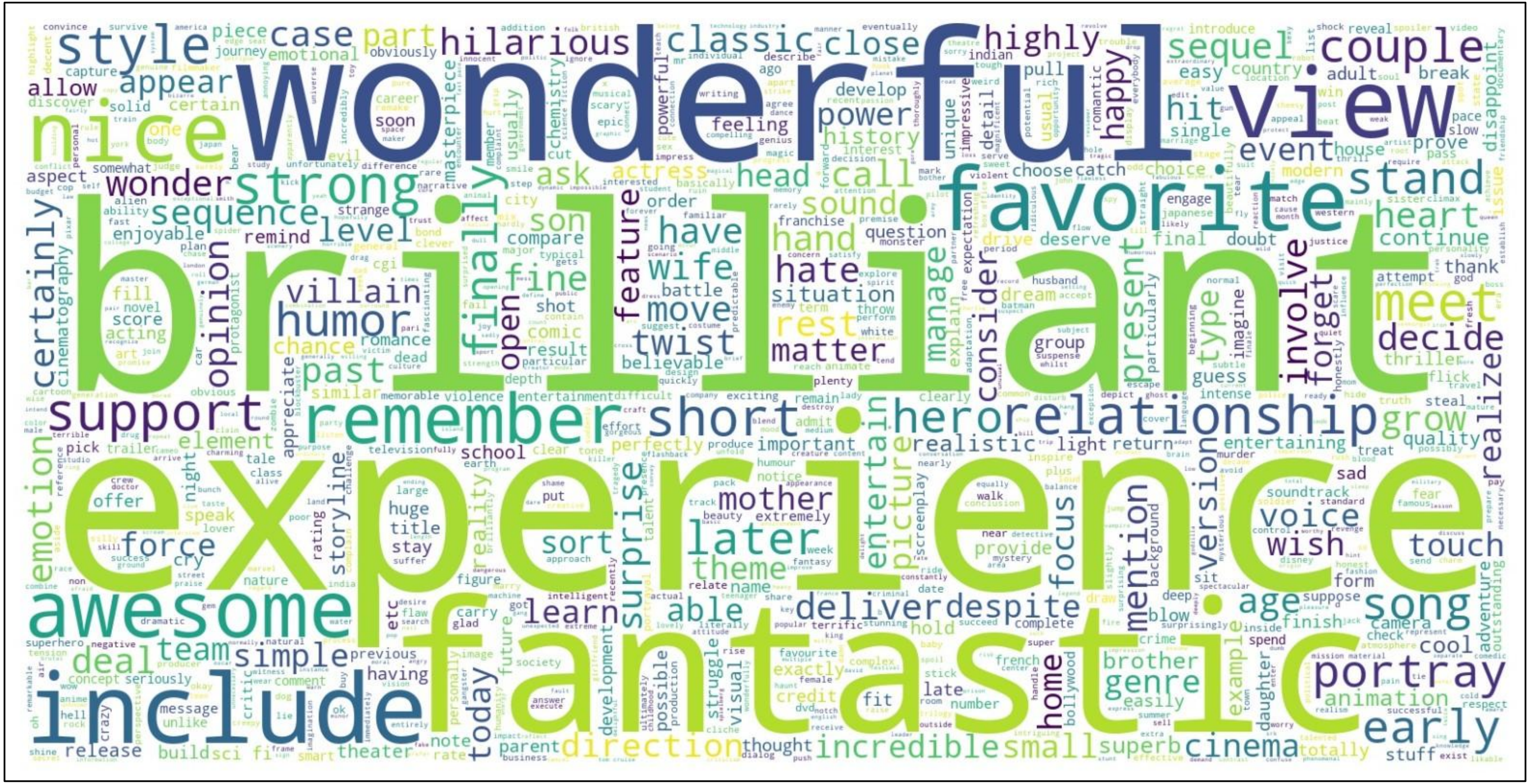
Reviews with rating 1,2,3 were labeled as Negative and reviews with rating 8,9,10 were labeled as Positive.

250K Reviews were used for the analysis.

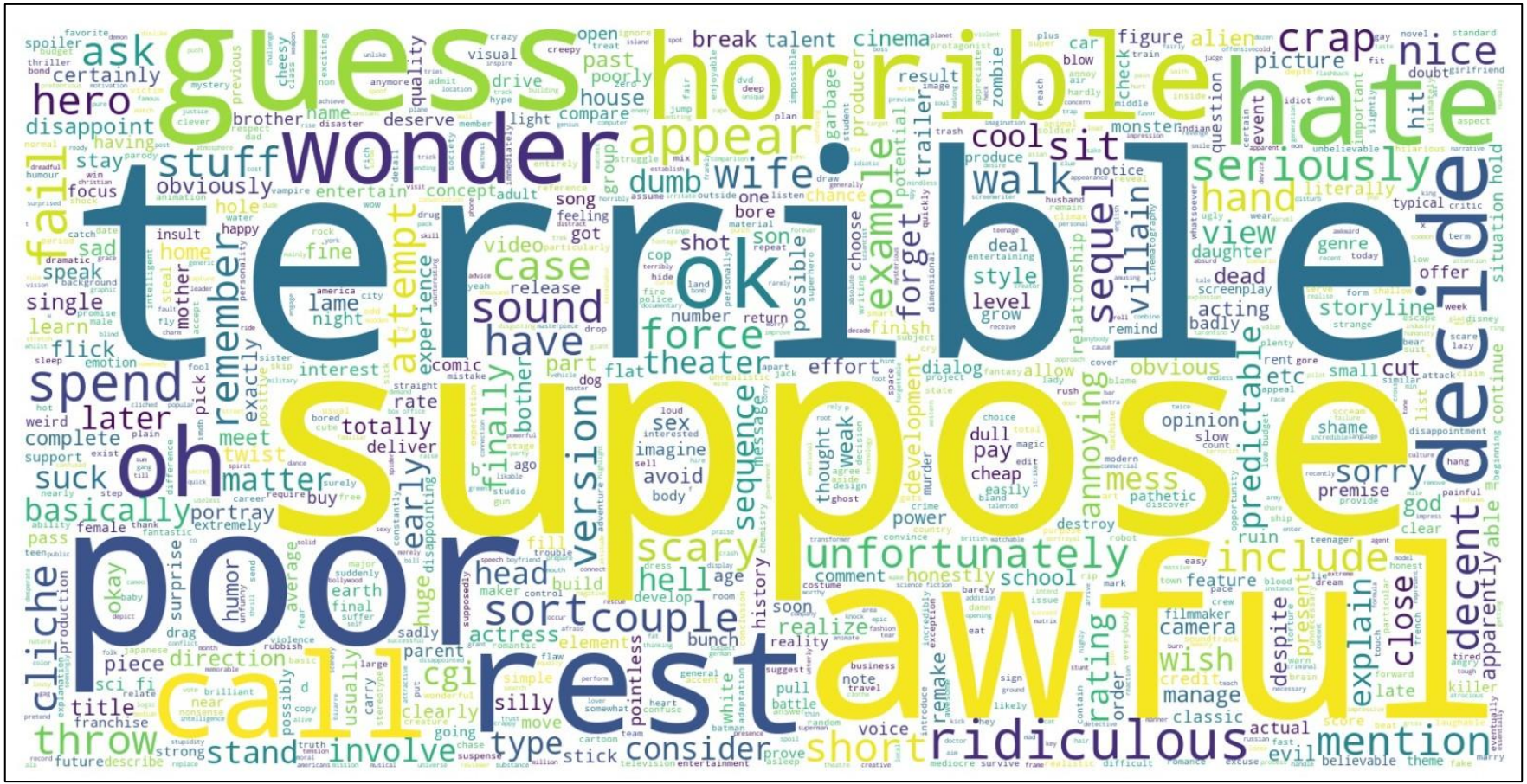
Steps Involved In Cleaning The Text Data

- Removing White Spaces, Digits and HTML Tags
- Normalizing case
- Contraction to Expansion
- Removing Punctuation and Special Characters
- Removing Accented Characters
- Removing Stop Words
- Lemmatizing

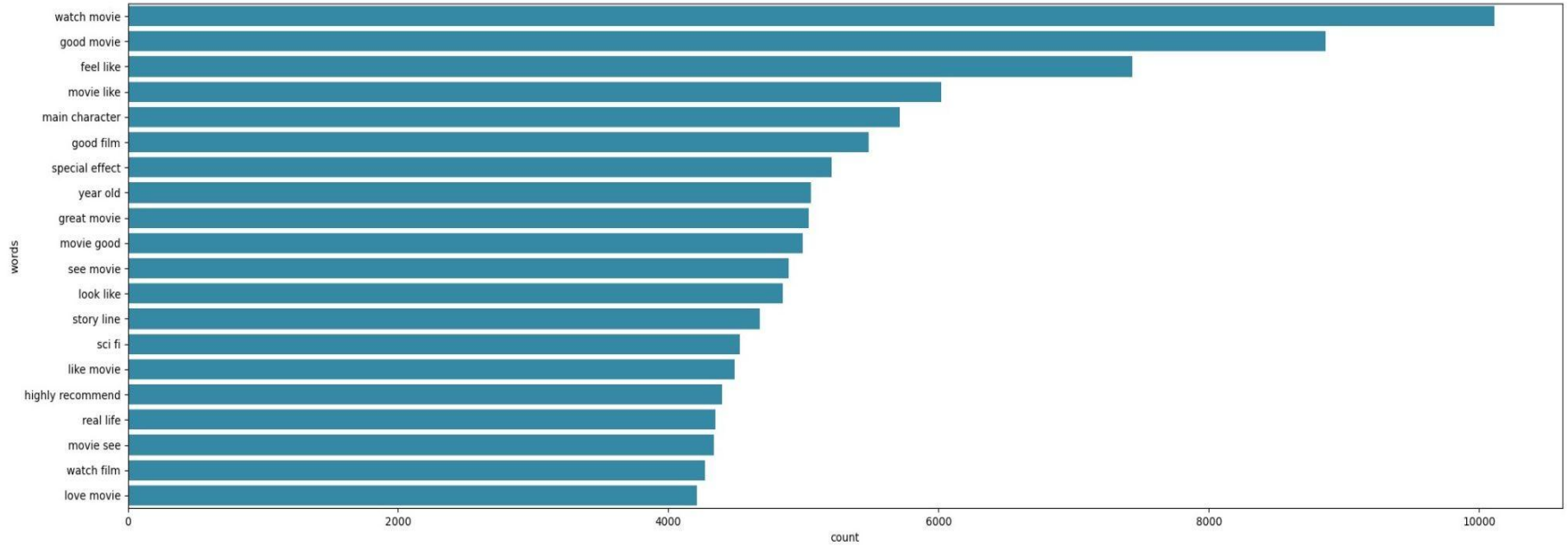
Word Cloud Of Positive Reviews (Top 1000)



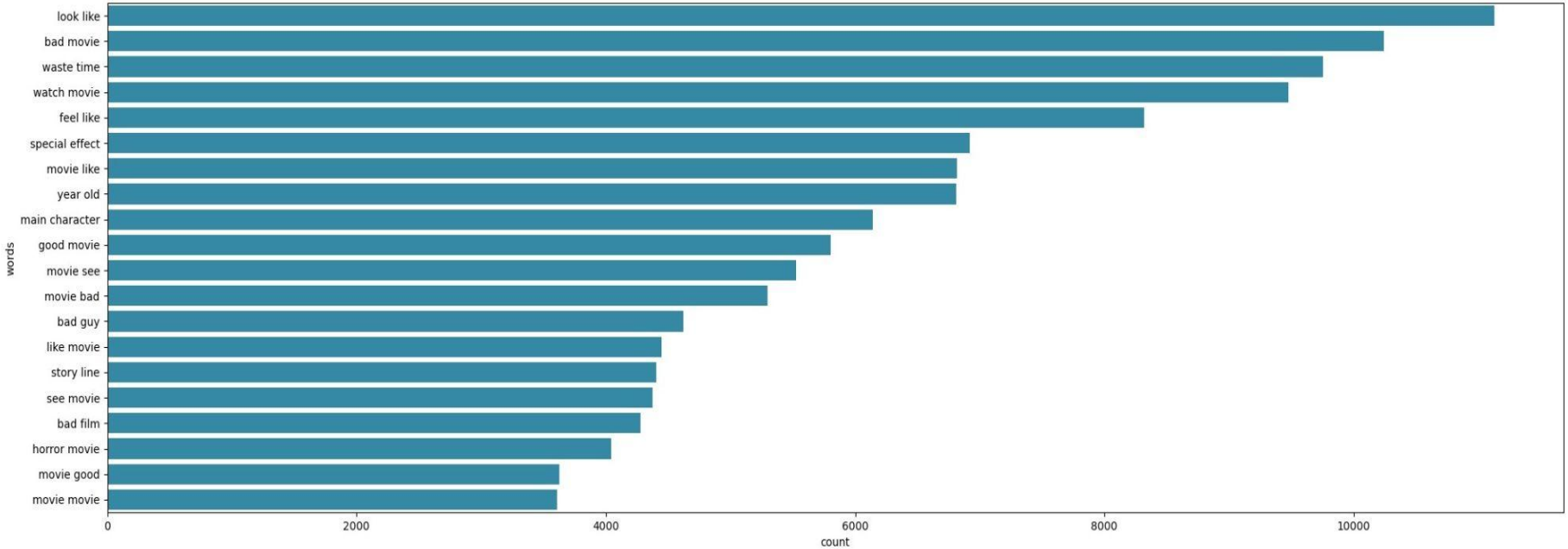
Word Cloud Of Negative Reviews (Top 1000)



Bigrams Of Positive Reviews (Top 20)



Bigrams Of Negative Reviews (Top 20)



Results

2 classification models were performed to determine best fit:
Logistic Regression and Multinomial Naive Bayes.

The Logistic Regression Model reported to be the best model for predicting Sentiment from given movie review using TFIDF Technique.

ACCURACY : 93% (*TRAIN*) and 91% (*TEST*)

Next Steps

- **Location, Time, Name and Numerals Stopwords** can be use to reduced the the document length.
- Removing **Low Occurring** features.
- Using **Recurrent Neural Network** to enhance the performance.

THANK YOU

https://github.com/avithekkc