Movie Review sentiment classifier

Hari Chandana Kotnani
Noah Hammons

Introduction:

- Our goal was to make a model that can accurately predict the sentiment of a movie review.
- Using our model you could predict quickly and accurately the overall sentiment of reviews.
- This could be used to show movie investors how well received a movie was so they can better decide on what movies to greenlight or not.

Dataset:

- Our dataset is an imdb dataset of 50,000 reviews obtained from kaggle.
- This dataset contains 25,000 highly polarizing reviews and 25,000 random reviews.
- By applying Naive Bayes to the IMDb dataset, we can build a model that can accurately predict the sentiment of new movie reviews.

Data Preparation:

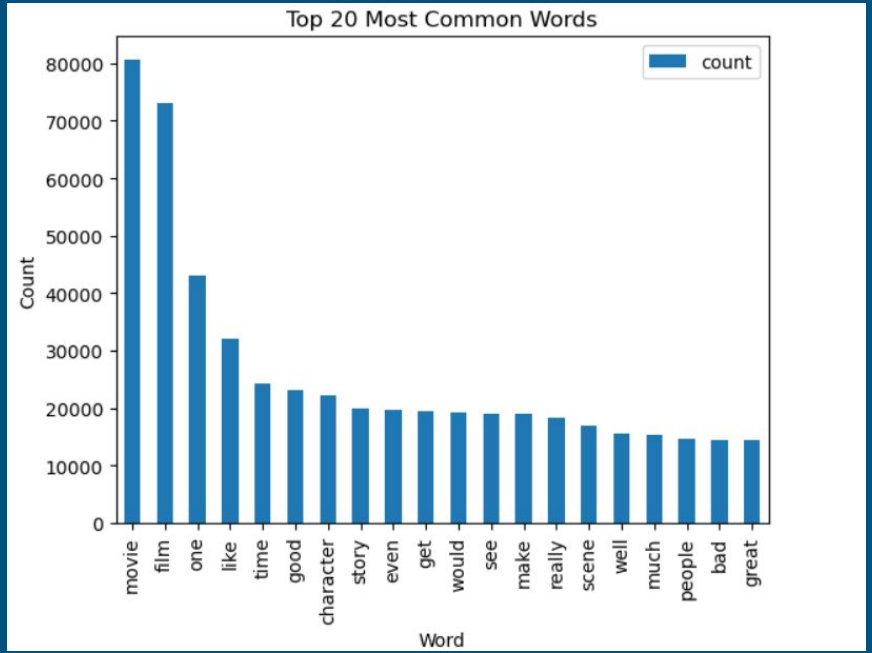
- We performed general cleaning by removing numbers, special characters, punctuation, emojis, and non-ASCII characters.
- We then eliminated extra white spaces, tabs, and newlines.
- To transform our text data into numerical format, we used CountVectorizer to create a matrix of token counts.
- We also generated bigrams using tokenization for each review.

Analysis techniques:

- The data is first split into training and testing sets.
- We used Multinomial Naive Bayes classifier because it is well-suited for text classification tasks.
- **Analysis 1 :**We used a word frequency analysis to identify the most common words in the reviews.
- Bigrams are generated from the text data, and a sentiment score is assigned to each bigram.
- **Analysis 2 :**The top occurring n-grams in positive and negative reviews are then identified by fitting a CountVectorizer, and the results are visualized using a bar chart."

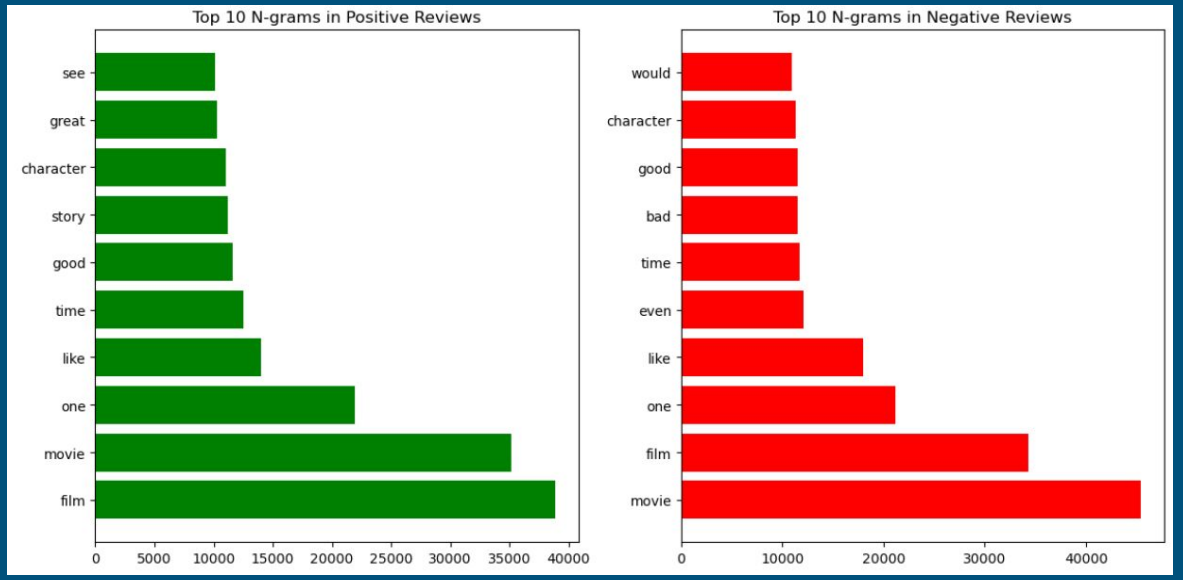
Analysis 1 :Most commonly used words

We found the most common words which led us to wonder if it was different for both positive and negative reviews.



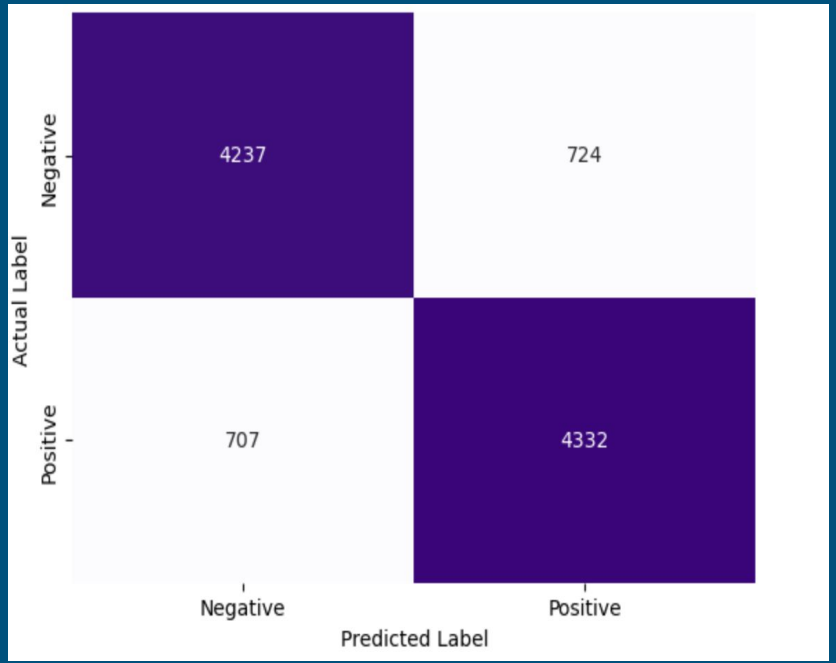
Analysis 2 : Does sentiment change wording.

We found that many of the words used in both positive and negative review have many similar words.



Results:

- After training our model, we constructed a confusion matrix to visualize the performance of the sentiment classification model.
- Our model correctly predicted 86% of the test samples overall.
- The precision, recall, and f1-scores were also 85-86% for both classes, indicating a balanced performance of the model.



THANK YOU