MILESTONE 1: PROBLEM, DATASET AND EXPLORATORY ANALYSIS

Project by: Albina Cako and Joshua Dalphy

Course: CSML1010

York University

BUSINESS CONTEXT

- Sentiment Analysis is the process of determining whether text is either positive, negative or neutral
- Valuable tool for companies and content creators to:
  - **Gauge public opinion on their products/services**
  - Perform market research
  - Monitor their reputation
- In the entertainment industry sentiment analysis can be used on reviews to gauge the audience's reaction to trailers, movies and tv shows.
  - Provides valuable data to studios
  - Studios can use the data to better understand their customers
  - Studios can use the data to improve or change content to better meet their customer's expectations

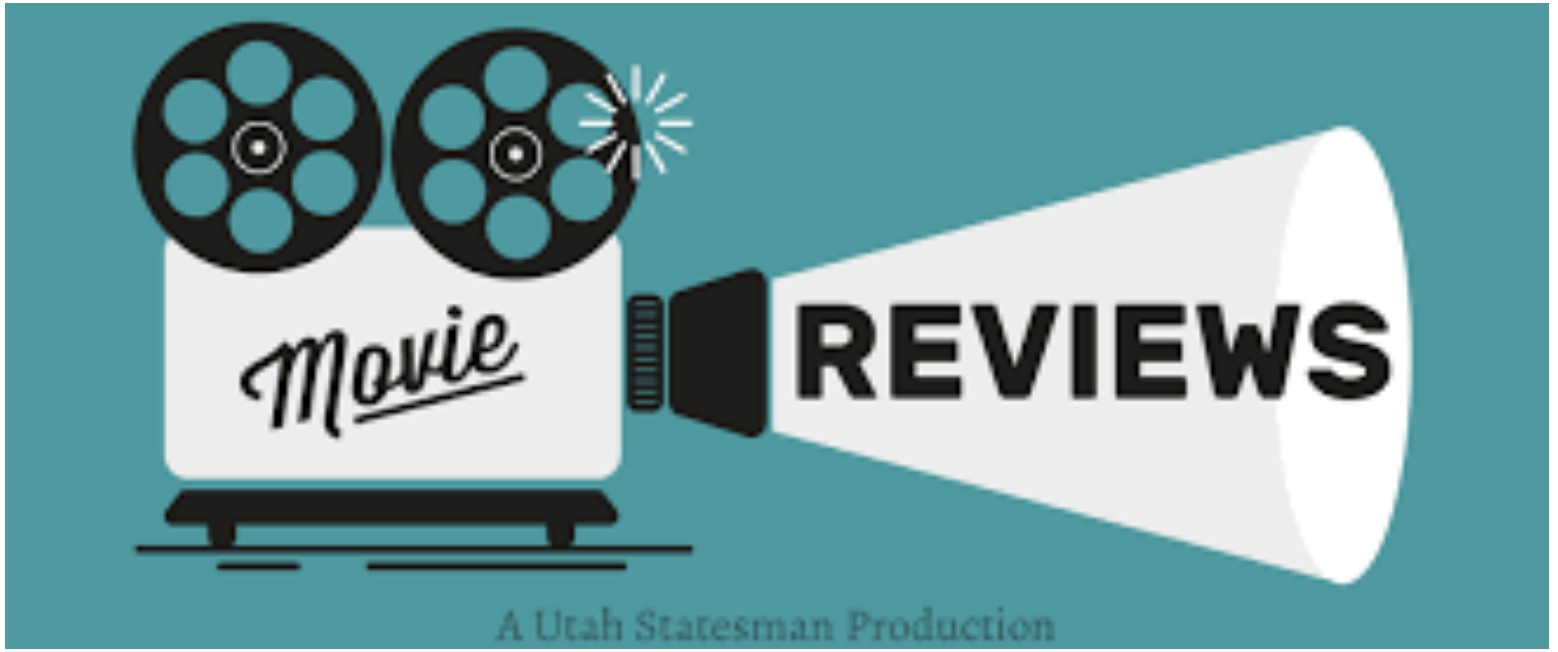
https://www.ksat.com/entertainment/2020/08/07/nothing-new-to-watch-on-netflix-think-again/

PROJECT DEFINITION

- Develop a machine learning model to perform sentiment analysis on movie reviews.
- The model will take written reviews as inputs and classify them as either positive or negative.
- Use the machine learning life cycle principles learned throughout this course and apply them to our project.

https://www.marketmotive.com/blog/discipline-specific/social-media/sentiment-analysis-article

DATA SELECTION



https://ususstatesman.com/movie-review-it/

- The data used for the project is the Large Movie Review Dataset obtained from the ai.standford.edu website.
- The data contains a total of 50,000 reviews divided evenly between testing and training folders.
  - The data is balanced and has 25,000 positive and negative reviews.
- For the purpose of this project 7000 movie reviews were selected.
- The dataset selected is balanced, containing 3500 positive and 3500 negative reviews.

DATA PREPARATION AND CLEANING

THE DATASET WAS NORMALIZED USING TEXT_NORMALIZER.PY

THE CLEANED DATASET WAS SAVED INTO A CSV FILE AS MOVIE_REVIEWS_CLEAN.CSV

DATA EXPLORATION

Sample review (positive sentiment):

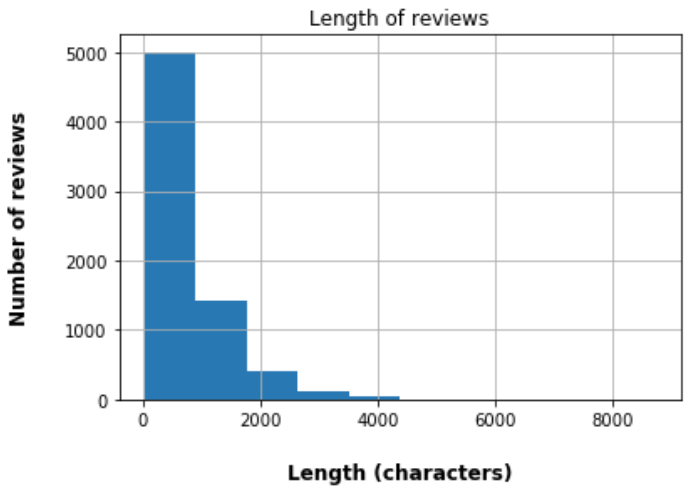
production quality cast premise authentic new england waterbury ct locale lush john williams score result 3 4 star collector item unfortunately get passable 2 star decent flick mostly memorable try bring art house style film mainstream small town locale story ordinary people genre well satisfy grownup jane fonda unable hide braininess enough make character believable wonder not post doctorate yale instead work dead end factory job waterbury robert diniros character bit contrived illiterate nice guy loser turn actually little help janes character 1990 version henry ford thomas edison genre successfully handle nobodys fool mid 90 year 2003 schmidt wish main stream studio would try stuff post adolescent reserve couple screen multi cinema complex effort give effort

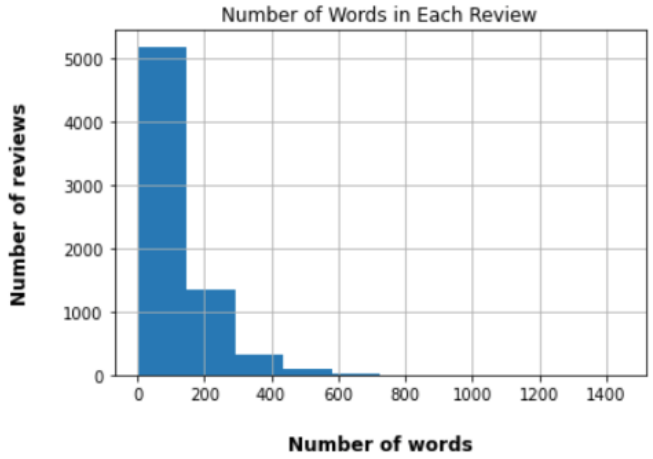
https://www.datapine.com/blog/data-exploration-vs-data-presentation/

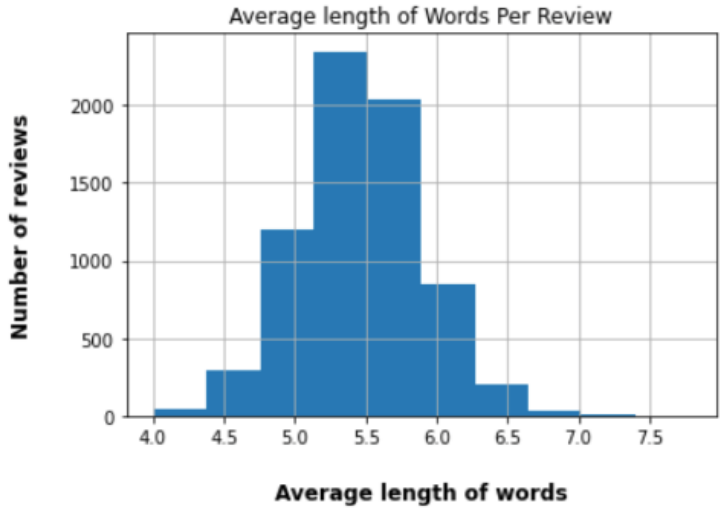
- The dataset contained 2 columns: reviews and sentiments
- The dataset contained 7000 rows, with 6988 unique reviews.
- The sentiment column was either ”Positive” or “Negative” with a total of 3501 positive ones and 3499 negative ones.

LENGTH ANALYSIS OF REVIEWS

- Approx. 70 % of reviews were less then 1000 characters
- The remaining 30% were between 1000 – 4000 characters.
- Most reviews had 200 words or less, while the longest reviews were over 700 words.
- The average length of word in each review was around 5 letters.

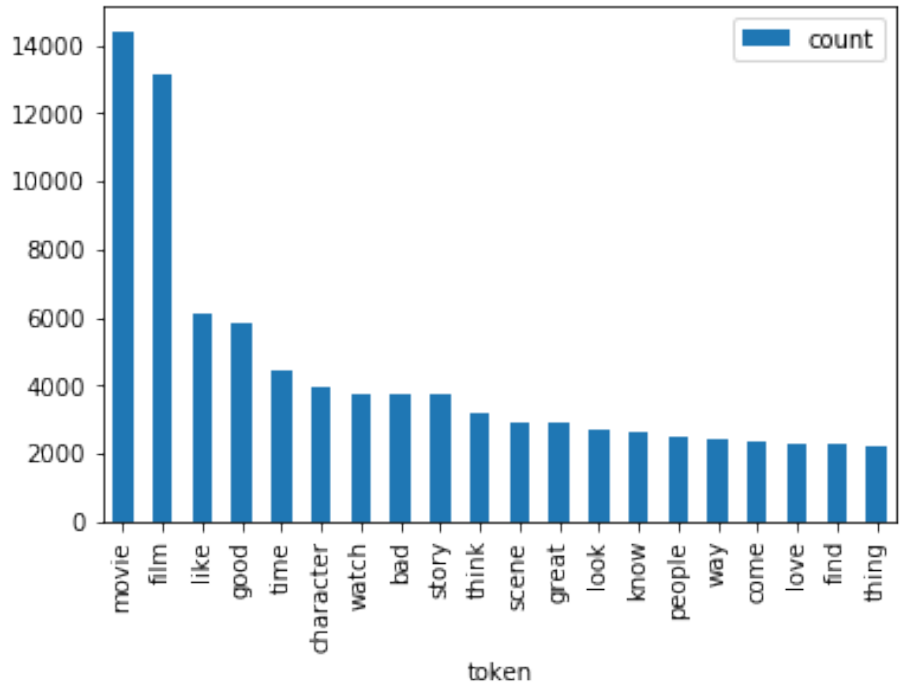






EXPLORING THE DATASET AS A WHOLE

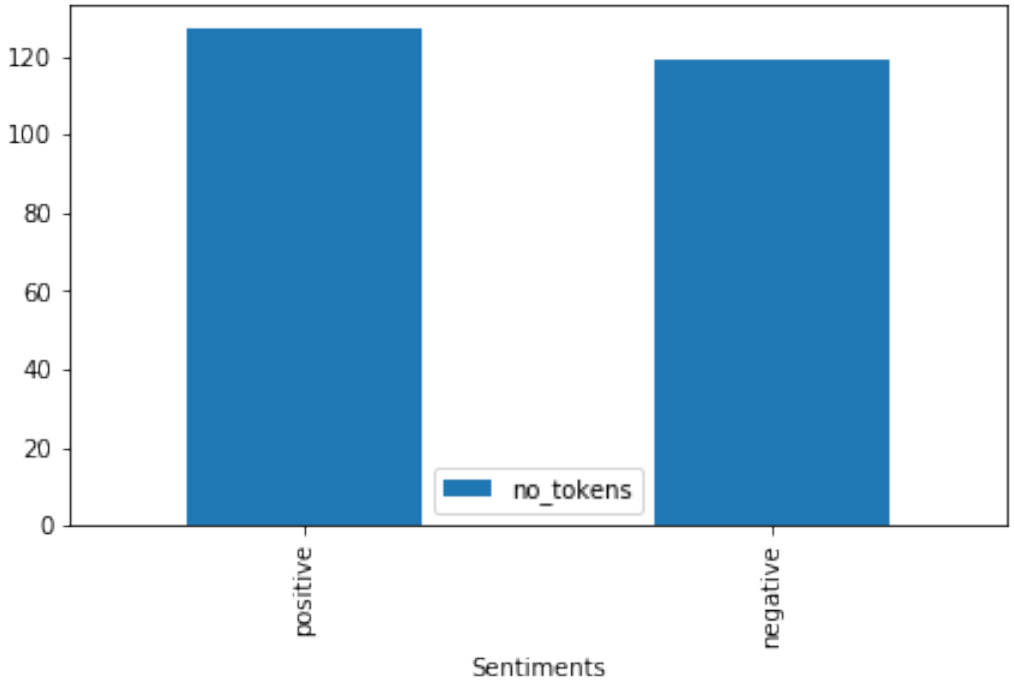
The 20 most common tokens in the dataset

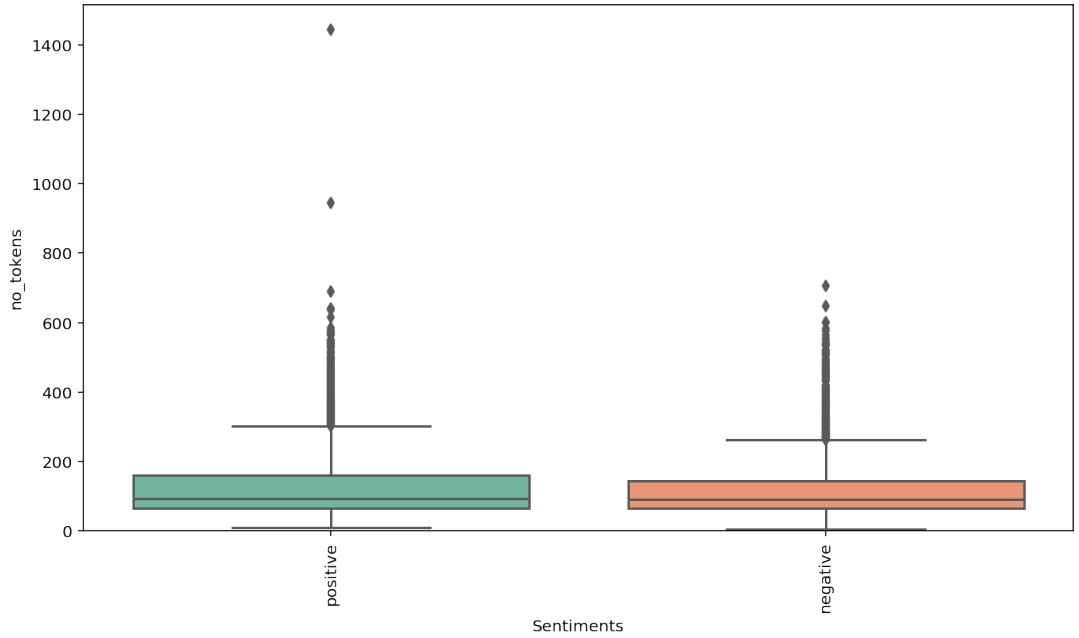


Wordcloud representation of most frequent tokens

EXPLORING THE DATASET USING SENTIMENTS

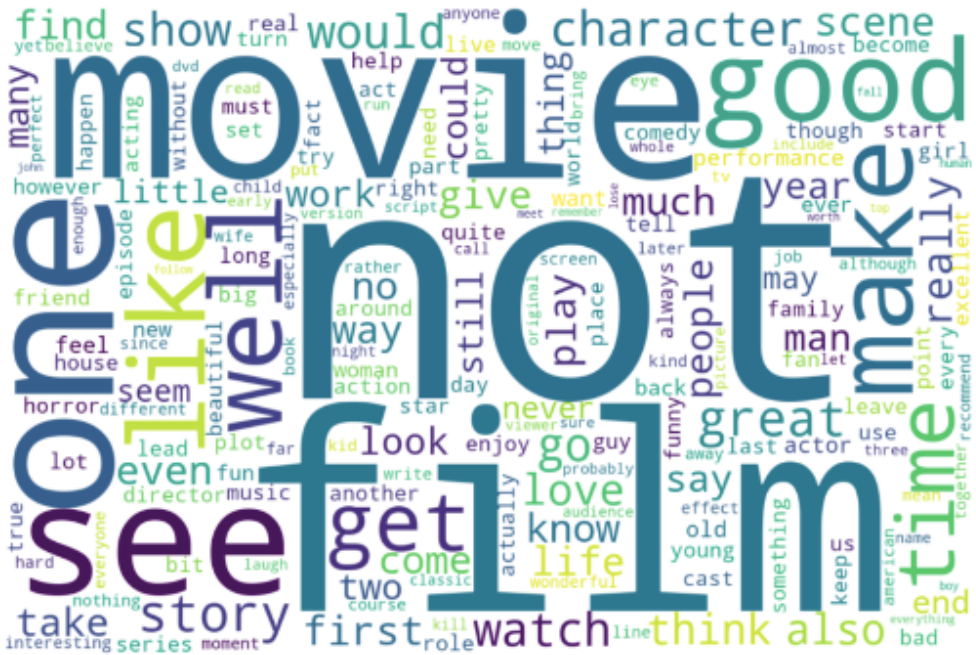
- The positive reviews are slightly longer than the negative reviews
- Outliers identified in both positive and negative reviews



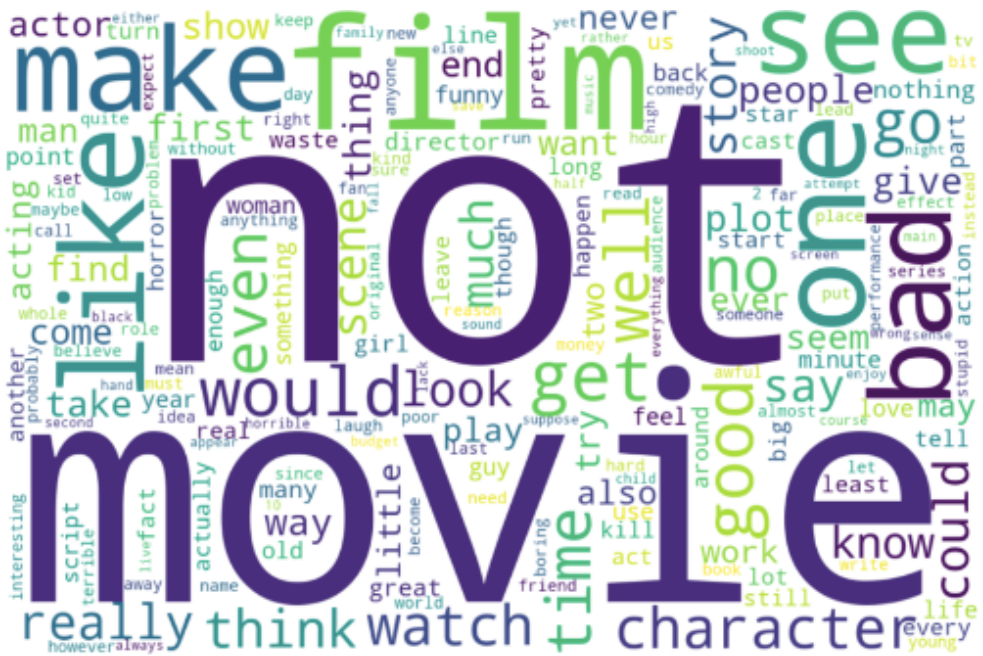


EXPLORING THE DATASET USING SENTIMENTS: COMMON WORDS

Positive reviews common words



Negative reviews common words



- Difficult to distinguish sentiment
- Approximately, 1.73 % of total tokens were common between positive and negative reviews
- It might be beneficial to consider using bag of n-grams to better capture the sentiment of the review

REFERENCES

- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).*