Sentiment Analysis on IMDB movie Review

Problem Statement

Performing a sentiment analysis on movie reviews is challenging because of obstacles like sentence negation, sarcasm, terseness, language ambiguity, and etc. This project will be analyzing sentiment of movie reviews from IMDB.

Why does sentiment analysis matter?

The ability to quickly understand consumer attitudes and react is important. The sentiment can be a part of consumer behavior marketing with other factors such as psychological, personal and social. This project will help businesses to automate movie review sentiment analysis and allow quicker understanding of consumer behavior.

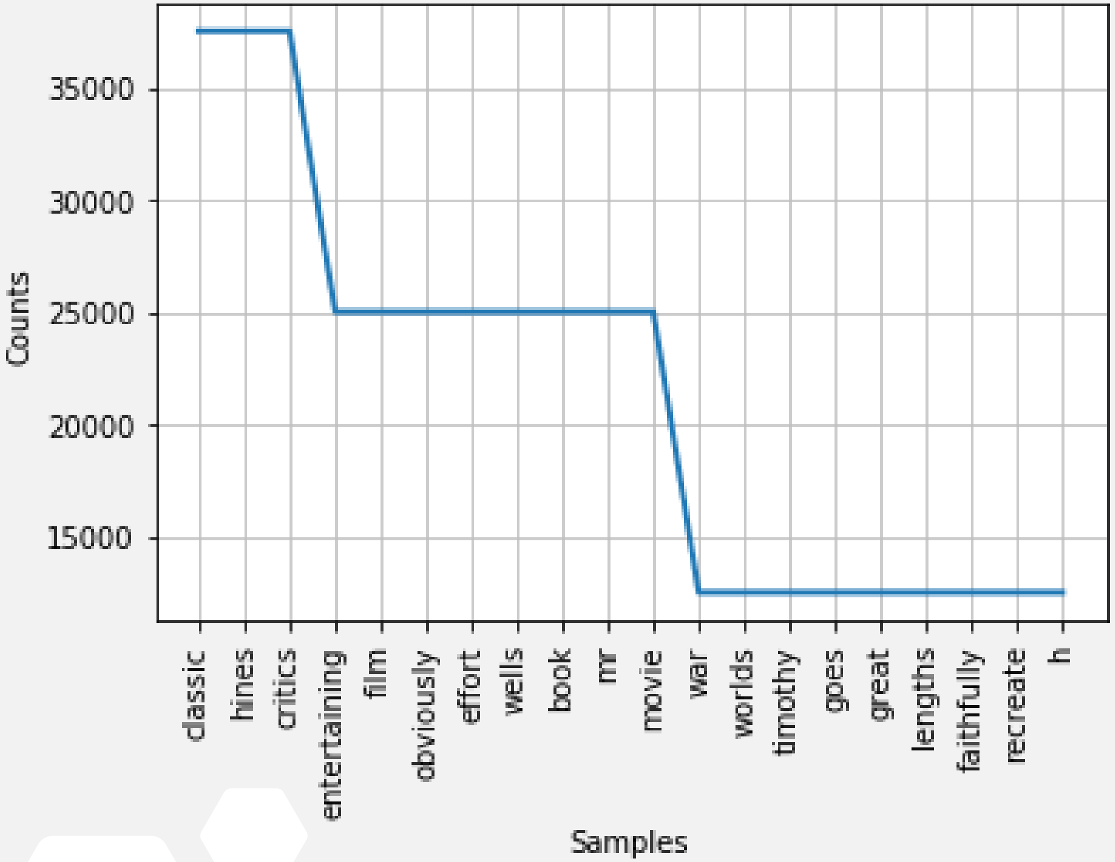
Collection of data and cleaning the data

- 25000 movie reviews in tsv format provided by IMDB including 12500 positive and 12500 negative ones
- The sentiment column indicates whether the review is positive (1) or negative (0)
- Preprocessing
  - Non-ascii, punctuation, numbers, and stop words were removed
  - Letters were converted to lowercase
  - Words were stemmed, lemminized, and tokened.
  - Cleaned words were combined to create a corpus for TF*IDF algorithm

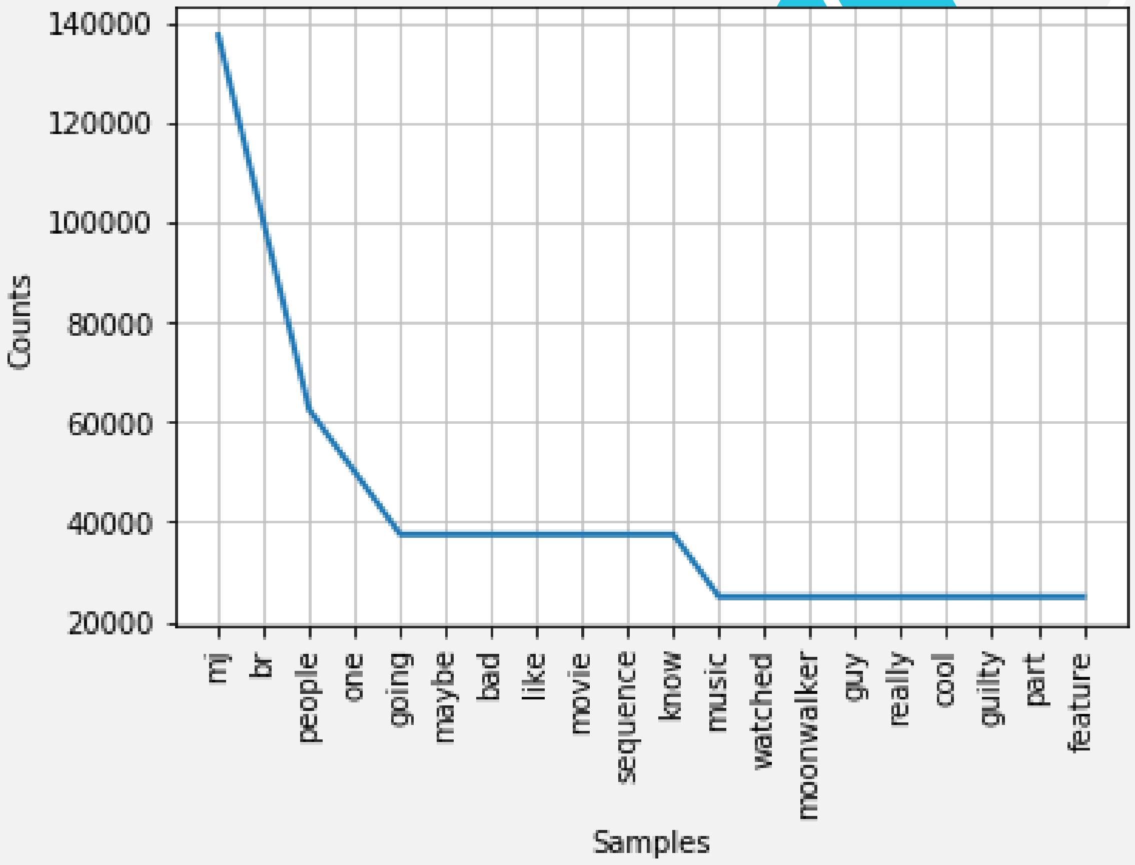
Exploratory Data Analysis

Word Frequency in Movie Reviews

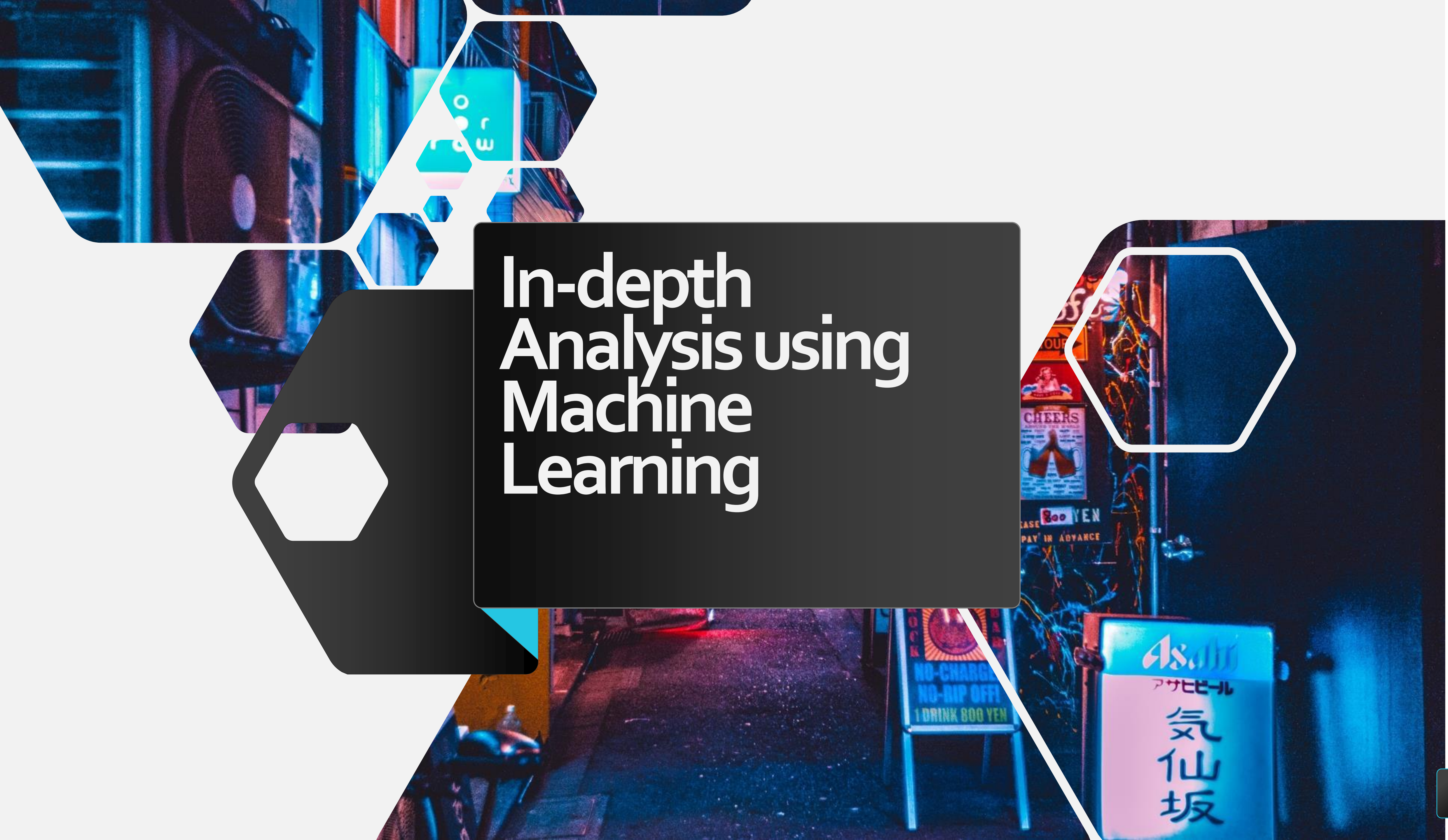
Positive Movie Reviews



Negative Movie Reviews



The graph above shows the top 20 words that were mentioned in the movie reviews. Word ‘entertaining’ contains positive meaning but other words on the charts are neutral.



In-depth Analysis using Machine Learning

TF-IDF Algorithm

TF-IDF is a way to convert textual data to numeric form, and is short for Term Frequency-Inverse Document Frequency.

- The algorithm is used to weigh a keyword in any content and assign the importance to that keyword based on the number of times it appears in the document
- Words that appear in many reviews have a value closer to zero and words that appear in less documents have values closer to 1

- Formula
  - $TF = \frac{number\ of\ times\ term\ appears\ in\ review}{total\ number\ of\ terms\ in\ review}$
  - $IDF = \ln(\frac{number\ of\ reviews}{number\ of\ reviews\ the\ term\ appears\ in})$

Model Performances

Supervised learning - classification

	F-1 score
Logistic classification	88%
Support vector machine classification	88%
Naïve Bayes classification	86%

- 70% to train and 30% to test the model
- Logistic classification is chosen because of higher F-1 score and less expensive computational cost

Thank You

👤 Jaesuk Kim

✉ jaesukkim4402@gmail.com

🔗 www.linkedin.com/in/jaesuk-kim

Please see Jupyter Notebook for further analysis