Final Report: Movie Review Sentiment Analysis

Andrew Bryan

Problem Statement

The goal of this project is to enhance the accuracy of movie ratings on Rotten Tomatoes by developing a machine learning model capable of determining the sentiment of movie reviews based solely on the review text. This approach aims to provide better rating results that more accurately represent the general audience's sentiment and mitigate the effects of review bombing.

Dataset

A dataset of 50,000 movie reviews, each labeled with a sentiment (positive or negative), was used for this project. The dataset provides a rich source of textual data that allows for the application of natural language processing (NLP) techniques to discern sentiment directly from the review content.

Project Scope

The project involves the following key activities:

- **Data Collection and Preprocessing:** Cleaning and preparing the dataset for analysis, including removing stop words and tokenization
- **Feature Engineering:** Using techniques like TF-IDF and word embeddings to convert text data into numerical features suitable for machine learning models.
- **Model Development:** Building and training various machine learning models, including random forest and nearest neighbor to predict the sentiment of movie reviews.
- **Model Evaluation and Selection:** Evaluating the models based on accuracy, precision, recall, and F1-score, and selecting the best-performing model.

Constraints

The project faced certain constraints, including:

Data Limitations: The dataset only covers a specific set of 50,000 reviews, which might not capture the full diversity of movie reviews.

Resource Constraints: Limited computational resources for training complex models and conducting extensive hyperparameter tuning.

Data Trustworthiness: The dataset used for this project was sourced from Kaggle. Without additional resources and verification, it is challenging to ascertain the reliability of the dataset. Consequently, we have to rely on the provided sentiment labels and assume their accuracy.

Ambiguity in Neutral Reviews: The current model does not account for neutral reviews, where reviewers may have mixed or moderate opinions about a movie. These reviews, which do not clearly classify a movie as either good or bad, present ambiguity in distinguishing varying levels of neutrality. This could affect the overall accuracy and nuance of the sentiment analysis.

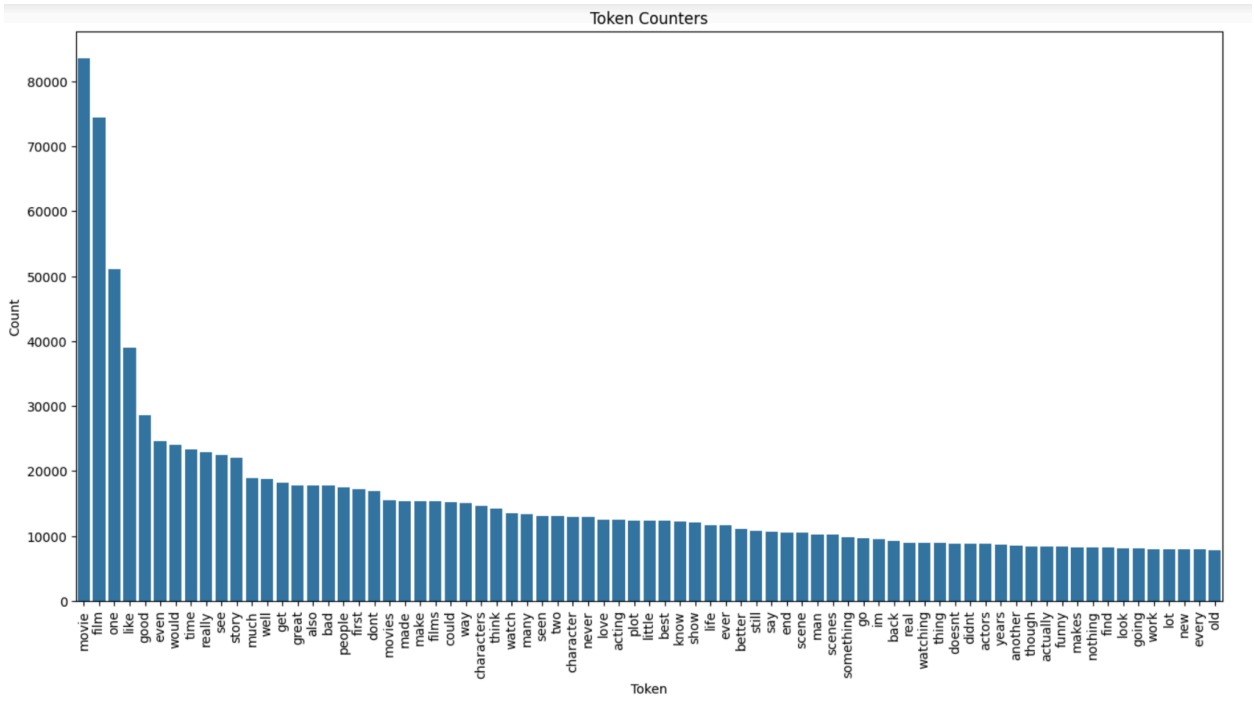
Data Wrangling and Cleaning

The dataset obtained from Kaggle consisted of 50,000 movie reviews along with their overall sentiment labels.

The initial dataset was well-prepared, with no null values and an equal distribution of positive and negative reviews (25,000 each). This cleanliness and balance allowed for a smooth transition into exploratory data analysis (EDA) without the need for extensive data cleaning.

Exploratory Data Analysis

During the exploratory data analysis (EDA), the focus was on identifying word trends across the entire dataset. To achieve this, tokens were extracted from each movie review and analyzed to uncover patterns and insights. Below is a histogram displaying the top 75 most frequently used tokens (words) across all reviews in the dataset.



Initially, the dataset appeared well-prepared, but the investigation revealed that some minor cleaning was necessary. A significant number of tokens had a usage count of only 1, which, upon further inspection, were identified as typos. To enhance data quality and improve model accuracy, it was decided to remove these low-frequency tokens. After additional investigation, a threshold was set: any tokens with 3 or fewer usages were dropped. This step helped clean the data and provided a more reliable foundation for subsequent analysis. Below is a histogram displaying the top 75 most frequently used tokens (words) across all reviews in the dataset, after cleaning.

	token	count
102399	vanquishbr	1
102397	ladyboyfriends	1
102396	kellysclara	1
102395	insidenow	1
102394	axethe	1
...	...	...
102331	objectprop	1
102329	myselfyeah	1
102328	introtutorials	1
102322	330ambr	1
181542	yosemitebr	1

100 rows × 2 columns

Model Preparation

Term Frequency-Inverse Document Frequency (TF-IDF) Approach
For the modeling technique, Term Frequency-Inverse Document Frequency (TF-IDF) was selected. This approach values words based on their frequency of use, giving higher weight to words that are more indicative of positive or negative sentiment.

Word Embeddings Approach using Word2Vec
The Word2Vec package was installed and used to implement a word embeddings approach. This method captures not only word usage but also meaning, which can improve the accuracy of sentiment analysis. After creating word vectors for each review, these vectors were used as features to train machine learning models.

Modeling

Traditional Machine Learning with TF-IDF

Two traditional machine learning algorithms, Random Forest and Nearest Neighbors, were applied to the TF-IDF approach:

- **Random Forest**: This ensemble learning method was chosen as the best model for the TF-IDF approach, achieving an accuracy of 85% on the test set. Random Forest is well-suited for handling high-dimensional data like text data and tends to generalize well to new data.
- **Nearest Neighbors**: While not as successful as Random Forest, Nearest Neighbors achieved 77% accuracy with the optimal number of neighbors. This model calculates the similarity between samples and assigns a label based on the labels of the nearest neighbors.

Grid search cross-validation was conducted on the training set to optimize hyperparameters, confirming that the original settings yielded the best accuracy scores.

Word Embeddings Approach with Random Forest

For the word embeddings approach:

- **Random Forest**: Chosen as the model due to its superior performance with TF-IDF. After creating vectors for each review using Word2Vec, the Random Forest model achieved 84% accuracy on the test set. This approach captures

semantic meaning and contextual relationships between words, allowing for a deeper understanding of the text data.

OpenAI API Approach

An additional approach involved using an API with a chatbot (OpenAI) to analyze movie reviews and determine sentiment:

This model achieved a high accuracy of 94%, indicating superior performance compared to traditional machine learning and NLP approaches. The OpenAI API uses advanced natural language processing techniques to interpret the sentiment of movie reviews accurately.

Summary

The findings suggest that the OpenAI API provides the most accurate and efficient method for sentiment analysis, offering potential applications in improving movie rating systems like Rotten Tomatoes.

Applications

Rotten Tomatoes can begin using this model today and it would provide the following applications:

Automated Review Aggregation: Sentiment analysis can be used to automatically classify the sentiment of user-generated reviews on Rotten Tomatoes, providing a quicker and more efficient method to aggregate sentiments and update movie scores. This can help in maintaining up-to-date movie ratings that reflect the latest viewer opinions, especially during the initial release period when large volumes of reviews are posted.

Enhanced User Experience: By utilizing sentiment analysis, Rotten Tomatoes can offer more nuanced insights into movie reviews. For example, the site could provide breakdowns of positive, negative, and neutral sentiments, along with intensity scores. This detailed feedback could help users make more informed decisions about which movies to watch based on more than just a simple star rating.

Content Moderation and Quality Control: Sentiment analysis could also be applied to monitor and manage the quality of reviews posted on Rotten Tomatoes. By identifying extreme sentiment, either overly negative or positive, the system could flag potential

review bombs or unusually biased reviews. This could help maintain the integrity and reliability of the review scores, ensuring they represent a balanced view of the audience's opinions.

Future Research

This project initiates a preliminary exploration into sentiment analysis for movie reviews. Each research idea below has the potential to further advance the field of sentiment analysis in the context of movie reviews, offering new insights and methods for understanding audience opinions and preferences.

Fine-Grained Sentiment Analysis: Explore methods to perform fine-grained sentiment analysis on movie reviews. This could involve detecting and analyzing specific aspects of movies such as plot, acting, cinematography, and soundtracks. By understanding the sentiment towards these aspects separately, moviegoers can get more detailed insights into what aspects of a movie are liked or disliked.

Temporal Analysis of Sentiment: Investigate how sentiment towards movies changes over time. This could involve analyzing sentiment trends across different phases of a movie's lifecycle, from pre-release hype to post-release reactions. Understanding how sentiment evolves over time can provide valuable insights into the dynamics of audience opinions and preferences.

Multimodal Sentiment Analysis: Combine textual reviews with other forms of data, such as audio and video reviews, to perform multimodal sentiment analysis. This could involve analyzing sentiment expressed in YouTube video reviews, podcasts, or social media posts related to movies. By integrating multiple modalities, a more comprehensive and accurate understanding of audience sentiment can be achieved.