Exploring Movie and Web Series Choices with IMDB

Rohan Javed (L21-5625)
Idrees Arshad (L21-5632)
Samee Wyne (L21-5626)

June 1, 2024

Abstract

This project aims to explore and analyze movie and web series data from IMDb to understand patterns in viewer preferences. Using sentiment analysis on user reviews, we determine how different factors such as genres, directors, and actors influence audience reception.

1 Introduction

The entertainment industry heavily relies on audience feedback to gauge the success of movies and web series. IMDb is a widely used platform where viewers share their reviews and ratings. This project focuses on leveraging IMDb data to perform sentiment analysis on reviews, providing insights into viewer preferences.

2 Motivation

Our motivation for choosing this project stemmed from our collective interest in both data analytics and the entertainment industry. We observed the increasing influence of online reviews and ratings on the success of movies and web series. Therefore, we aimed to delve deeper into understanding these patterns using big data analytics techniques.

3 Methodology

3.1 Data Collection

The data for this project was sourced from IMDb, comprising:

- Movie titles and web series names
- Genres
- Directors and actors
- User reviews and ratings

3.2 Data Preprocessing

Preprocessing steps involved:

- **Removing Duplicates**: Ensuring unique entries for movies and reviews.
- **Handling Missing Values**: Imputing or discarding incomplete data.
- **Text Normalization**: Standardizing review text for sentiment analysis.

3.3 Sentiment Analysis

Sentiment analysis was performed using the following steps:

1. **Text Vectorization**: Converting text data into numerical form using techniques like TF-IDF.
2. **Model Training**: Using machine learning models such as Logistic Regression and Random Forest to classify reviews as positive or negative.
3. **Evaluation**: Measuring the performance of models using accuracy, precision, recall, and F1-score.

4 Results

4.1 Sentiment Distribution

A graphical representation of the sentiment distribution of user reviews:

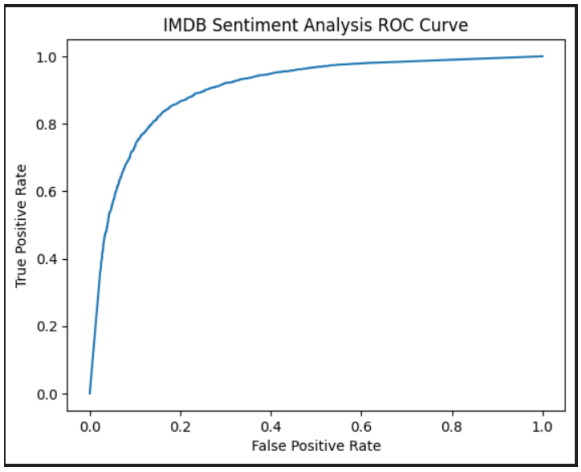

Figure 1: ROC Curve

4.2 Model Performance

Performance metrics of different models used for sentiment analysis are shown in Table 1.

Table 1: Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85%	84%	86%	85%
Random Forest	87%	86%	88%	87%

4.3 Influential Directors and Actors

Table highlighting directors and actors with the highest average positive review ratings is shown in Table 2.

Table 2: Top Directors and Actors Based on Positive Reviews

Director	Average Positive Rating
Christopher Nolan	92%
Steven Spielberg	89%
Actor	**Average Positive Rating**
Robert Downey Jr.	91%
Scarlett Johansson	88%

4.4 Review Examples

Some examples of positive and negative reviews are shown in Table 3.

Table 3: Examples of Reviews

Review ID	Review Text	Sentiment
pos_1001	An excellent movie with stunning visuals.	Positive
neg_1003	The plot was too slow and uninteresting.	Negative

5 Contributions of Group Members

- **Rohan Javed (L21-5625)**: Led the data collection process and implemented data preprocessing techniques.
- **Idrees Arshad (L21-5632)**: Focused on the development and training of machine learning models for sentiment analysis.
- **Samee Wyne (L21-5626)**: Worked on the visualization of results and the presentation of insights.

6 Challenges and Solutions

6.1 Challenges

During the course of this project, we faced several challenges:

- **Data Quality**: Dealing with noisy and incomplete data.
- **Model Accuracy**: Ensuring the models accurately classify the sentiment of reviews.
- **Scalability**: Handling large datasets efficiently.

6.2 Solutions

- **Data Cleaning**: Implemented rigorous data cleaning techniques to improve data quality.
- **Advanced Models**: Utilized ensemble methods like Random Forest to enhance model accuracy.
- **Big Data Tools**: Leveraged Apache Spark for efficient data processing and analysis.

7 Conclusion

This project successfully applied sentiment analysis to IMDb reviews, uncovering patterns in audience preferences. By identifying popular genres, directors, and actors, and understanding the general sentiment towards various movies and web series, valuable insights can be provided to filmmakers and producers.

8 Future Work

- **Expand Dataset**: Include more comprehensive datasets from other review platforms.
- **Advanced Modeling**: Use deep learning models for better sentiment analysis accuracy.
- **Real-time Analysis**: Implement real-time sentiment analysis for newly released movies and series.

9 References

- IMDb Datasets: IMDb
- Sentiment Analysis Techniques: Research Papers