Movie Success Prediction & Sentiment Study

Introduction

The film industry generates vast amounts of data, including box office collections, viewer ratings, and user reviews. Leveraging this data enables us to identify key factors driving movie success and to capture audience sentiment. This project combines predictive modeling and natural language processing (NLP) to analyze how sentiment from reviews influences movie performance.

Abstract

This project aims to predict the box office success of movies while simultaneously analyzing audience sentiment from user reviews. IMDb/Kaggle-style datasets are used to create predictive models with regression techniques. Sentiment analysis is carried out using VADER, which classifies reviews into positive, negative, and neutral sentiments. The outcome provides insights into genre-specific sentiment trends and identifies the most impactful factors influencing box office revenue.

Tools Used

• Python
• Libraries: pandas, numpy, matplotlib, scikit-learn, nltk (VADER)
• Excel/CSV for dataset storage and handling
• Jupyter Notebook for development and documentation

Steps Involved in Building the Project

1. Data Collection: Import IMDb/Kaggle datasets containing movie metadata and user reviews.
2. Data Cleaning: Handle missing values, encode genres, and prepare the dataset for modeling.
3. Sentiment Analysis: Apply VADER to user reviews to compute compound, positive, negative, and neutral scores.
4. Aggregation: Generate sentiment summaries by movie and by genre.
5. Predictive Modeling: Build regression models (Linear Regression, Random Forest) to predict box office revenue.
6. Evaluation: Assess performance using metrics such as R^2 and Mean Absolute Error (MAE).
7. Visualization: Produce charts showing genre sentiment distribution, sentiment vs. ratings, and feature importance.

Conclusion

The project demonstrates how sentiment analysis and predictive modeling can be integrated to evaluate movie success. By analyzing user reviews alongside traditional metadata, we gain deeper insights into audience perceptions and their correlation with box office performance. This framework provides a foundation for further exploration using large-scale real-world datasets.