Project Report: Movie Success Prediction & Sentiment Study

1. Introduction

This project aims to analyze movie reviews and metadata to predict box office success and understand public sentiment. Using machine learning and natural language processing (NLP), we explore how user sentiment influences movie popularity and financial performance.

2. Abstract

We utilized IMDB user review data and TMDB movie metadata. Sentiment analysis was performed using VADER, and a regression model (Random Forest) was trained to predict movie revenue. Word clouds and visualizations were created to interpret sentiment distribution. Results show that review sentiment and factors like budget, popularity, and vote count strongly affect revenue.

3. Tools Used

- Python (pandas, matplotlib, seaborn, sklearn, NLTK, VADER)
- Google Colab (development environment)
- WordCloud (visual sentiment keywords)

4. Steps Involved

1. Data Collection: IMDB reviews + TMDB movie metadata.
2. Preprocessing:
   - Cleaned text and structured numeric data
   - Removed missing/null values
3. Sentiment Analysis:
   - Applied VADER to extract sentiment from reviews
   - Classified as positive, neutral, or negative
4. Genre-Based Analysis:
   - Grouped reviews by randomly mapped genres
   - Created sentiment distribution graphs
5. Revenue Prediction:
   - Trained a Random Forest Regressor
   - Features: budget, popularity, votes, runtime
   - Achieved RMSE: ~[your value] | R^2 Score: ~[your value]
6. Visualizations:
   - Word clouds for positive/negative reviews
   - Feature importance for prediction

5. Conclusion

The model provides good accuracy in predicting movie revenue based on key metadata and review sentiment. VADER effectively captures emotional tone in reviews, aiding insight into audience perception. This project demonstrates how NLP and ML can be used for success forecasting in the entertainment industry.