Project Report

🎬 **Movie Success Prediction and Sentiment Study**

Introduction

With the explosion of content in the entertainment industry, especially film, production houses are keen on understanding what makes a movie successful. This project aims to build a system that predicts movie success using textual sentiment derived from movie overviews and structured data such as budget and revenue. It also investigates genre-wise sentiment trends and builds predictive models to support decision-making.

Abstract

This study integrates Natural Language Processing (NLP) with machine learning to predict the commercial success of movies. Using IMDB-style metadata, including budget, revenue, and movie overview, we apply sentiment analysis (via VADER) to extract the emotional tone of descriptions. These sentiment scores are then used alongside numerical features to predict whether a movie is a "Hit" or "Flop". Additionally, a regression model is built to estimate revenue based on key predictors. Genre-wise sentiment trends are also visualized to reveal interesting patterns in audience perception across different film types.

Tools Used

- **Python Libraries:**
  - pandas for data handling
  - nltk (VADER) for sentiment analysis
  - scikit-learn for model building
  - matplotlib and seaborn for visualizations
- **Jupyter Notebook:** For code execution and iterative analysis
- **Excel:** For initial exploration and formatting

Steps Involved in Building the Project

1. **Data Collection and Cleaning:** Used a cleaned dataset containing key movie information such as budget, revenue, overview, genres, etc.
2. **Sentiment Analysis:** Applied VADER sentiment analyzer to the 'overview' text to generate compound sentiment scores for each movie.
3. **Success Classification Logic:**

- If the sentiment is positive → movie is likely a *Hit*
- If sentiment is negative → it's a *Hit* only if revenue ≥ 3× budget
- If sentiment is positive but revenue < budget → classified as *Flop*

4. **Genre-wise Sentiment Visualization:** Extracted individual genres and plotted their average sentiment to understand how audience tone varies across movie types.
5. **Regression Model (Revenue Prediction):** Created a linear regression model to predict revenue using budget, popularity, sentiment score, etc.
6. **Classification Model (Hit/Flop Prediction):** Built a logistic regression model and evaluated it using confusion matrix and ROC curve.

Conclusion

The project highlights how sentiment analysis combined with financial indicators can be a strong tool in forecasting movie outcomes. While sentiment plays a role, budget and revenue are significant drivers. Genre-wise insights provide an added layer of business intelligence. The classification model achieved acceptable accuracy, and ROC curve analysis validated its performance. This approach can assist stakeholders in making informed decisions during production and marketing phases.