Project Report

🎬 Movie Success Prediction and Sentiment Study

Introduction

With the rapid growth of the entertainment industry, especially in film, you may find it increasingly important to understand what factors contribute to a movie's success. In this project, you'll build a system to predict movie success by analyzing both structured data (like budget and revenue) and unstructured text (like movie overviews). You'll also explore sentiment trends across genres and use machine learning to help make better data-driven decisions.

Abstract

In this study, you'll combine Natural Language Processing (NLP) with machine learning techniques to predict how successful a movie will be. By working with IMDB-style metadata—such as budget, revenue, and movie overviews—you'll apply the VADER sentiment analyzer to understand the emotional tone of each movie description. These sentiment scores will then be integrated with numerical features to classify movies as "Hit" or "Flop." You'll also build a regression model to estimate revenue and visualize sentiment trends by genre to gain deeper insights into audience perception.

Tools Used

You'll utilize the following tools and libraries:

- **Python Libraries**:
    - pandas for managing and manipulating data
    - nltk (with VADER) for performing sentiment analysis
    - scikit-learn for training and evaluating machine learning models
    - matplotlib and seaborn for creating visualizations
- **Jupyter Notebook**: For running your code interactively
- **Excel**: For preliminary data exploration and formatting

Steps You'll Follow in This Project

1. **Data Collection and Cleaning**:
   Start with a pre-cleaned dataset that contains essential information like budget, revenue, overview, genres, etc.
2. **Sentiment Analysis**:
   Use the VADER sentiment analyzer on the overview column to generate compound sentiment scores for each movie.
3. **Success Classification Logic**:
   - If the sentiment score is positive → classify the movie as a *Hit*
   - If the sentiment is negative → classify as a *Hit* only if revenue is at least 3× the budget
   - If the sentiment is positive but revenue is less than the budget → classify as a *Flop*
4. **Genre-wise Sentiment Visualization**:
   Break down multi-genre entries and plot the average sentiment per genre. This helps you see how different types of movies are generally perceived.
5. **Regression Model (Revenue Prediction)**:
   Build a linear regression model to predict movie revenue using features like budget, popularity, and sentiment score.
6. **Classification Model (Hit/Flop Prediction)**:
   Use logistic regression to classify movies and evaluate your model using tools like the confusion matrix and ROC curve.

Conclusion

Through this project, you'll see how combining sentiment analysis with financial data can help predict movie performance. Although sentiment has a measurable impact, budget and revenue remain critical factors. By analyzing genre-specific sentiment, you'll uncover deeper audience trends. The models you build—both regression and classification—offer valuable tools for filmmakers and marketers to make smarter, more informed decisions.