# 🎬 Movie Success Prediction and Sentiment Study

## Introduction

The entertainment industry relies heavily on audience feedback and revenue data to evaluate a film's performance. In an era where data analytics shape business decisions, predicting a movie's commercial success using measurable variables like budget, sentiment, and critic scores can offer valuable insights. This project explores how Natural Language Processing (NLP) and machine learning can be applied to predict a film's box office performance and understand the sentiment patterns associated with movie genres.

## Abstract

This project leverages data from IMDB/Kaggle to predict the box office success (gross) of a movie using a regression-based machine learning model. The workflow integrates sentiment analysis using VADER on plot summaries (via plot\_keywords) to understand viewer orientation toward a film's content. We use these sentiment scores along with numerical features like budget and imdb\_score to predict a movie's commercial potential. The outcome includes statistical predictions, visualizations, genre-wise sentiment analysis, and a comprehensive Excel dashboard for exploratory insights.

## Tools Used

- Python Libraries:  
 - NLTK, VADER: Sentiment analysis  
 - Pandas, NumPy: Data handling and cleaning  
 - Matplotlib, Seaborn: Data visualization  
 - Scikit-learn: Regression modeling  
- Excel:  
 - Exported multi-sheet summaries including pivot tables, descriptive stats, and predictive performance metrics

## Steps Involved in Building the Project

1. Data Loading  
 Loaded and preprocessed movie metadata from IMDB/Kaggle including essential fields like budget, gross, imdb\_score, genres, plot\_keywords, and cast information.  
  
2. Data Cleaning  
 Filled missing values using appropriate techniques: mode for categorical, median for numeric, and zero/default where applicable.  
  
3. Sentiment Analysis  
 Applied VADER sentiment analysis on the plot\_keywords column to derive compound sentiment scores. This helped reflect the thematic tone of a movie in numeric form.  
  
4. Regression Modeling  
 Built a Linear Regression model using budget, imdb\_score, and Sentiment\_Score as predictors for gross. Evaluated the model using R² Score and RMSE, which were then exported to Excel.  
  
5. Visualization and Pivot Tables  
 Generated insightful plots such as:  
 - Sentiment vs Genre (bar plot)  
 - IMDB Score vs Average Gross (line plot)  
 - Budget vs Gross Revenue (scatter plot)  
 Created Excel pivot tables for:  
 - Genre-wise sentiment  
 - Revenue by content rating, director, and country  
 - Correlation matrix of numerical features  
  
6. Excel Reporting  
 Exported all results into an Excel workbook (Movie\_Success\_Predictions.xlsx) with sheets for:  
 - Raw predictions  
 - Genre sentiment  
 - Revenue summaries  
 - Model performance  
 - Descriptive statistics

## Conclusion

This study demonstrates how predictive analytics combined with sentiment analysis can provide meaningful insights into movie performance. While box office outcomes are influenced by many unquantifiable factors, the model shows that integrating emotional tone (sentiment) with traditional numeric features enhances predictive understanding. The project delivers both analytical depth and practical reporting formats, enabling future business or production decisions to be guided by data-driven evidence.  
  
Note: All analysis is limited to publicly available metadata and sentiment proxies from textual descriptions. Further accuracy can be achieved by incorporating real-time reviews and social media engagement metrics.