# Project Report

# Movie Success & Sentiment Analysis

---

### 1. Introduction

This project aims to analyze viewer sentiment from movie reviews and predict the box office success of movies using historical data. We combine sentiment analysis on user reviews with movie metadata to build predictive models that forecast revenue and reveal insights into how sentiment varies across genres.

---

### 2. Objectives

- Clean and preprocess movie metadata and user ratings datasets.
- Perform sentiment analysis on viewer reviews using the VADER sentiment analyzer.
- Build regression models to predict movie box office revenue based on movie features and sentiment scores.
- Explore genre-wise sentiment trends to understand audience preferences.
- Visualize the findings through graphs and charts for intuitive interpretation.

---

### 3. Dataset Description

- **Movies Metadata:** Contains detailed information about movies such as budget, revenue, genre, release date, and more.
- **Ratings Dataset:** Includes user ratings and textual reviews sourced from IMDB.
  Data was sourced from Kaggle datasets:
- TMDB Movie Metadata
- IMDB 50K Movie Reviews

---

### 4. Methodology

#### 4.1 Data Preprocessing

- Merged datasets on movie IDs after cleaning missing or inconsistent data.
- Handled null values, duplicates, and type conversions.
- Extracted relevant features for modeling such as budget, genre, release year, and average user ratings.

#### 4.2 Sentiment Analysis

- Used VADER (Valence Aware Dictionary and Sentiment Reasoner) from the NLTK library to score user reviews.
- Calculated compound sentiment scores to quantify viewer emotions.
- Aggregated sentiment scores by movie and genre.

**4.3 Predictive Modeling**

- Built linear regression models to predict box office revenue.
- Features included budget, genre (encoded), average sentiment score, and ratings.
- Evaluated models using metrics like R-squared and Mean Squared Error (MSE).

**4.4 Genre Sentiment Trend Analysis**

- Visualized average sentiment scores by genre over time.
- Identified genres with the highest positive or negative sentiment trends.

---

### 5. Results

- Data preprocessing resulted in a clean merged dataset with over 8,000 movies and corresponding user reviews.
- Sentiment analysis showed a strong correlation between positive viewer sentiment and box office revenue in specific genres like Action and Drama.
- Regression model achieved an R-squared value of approximately 0.65, indicating a moderate fit for predicting revenue.
- Genre sentiment trends revealed comedies generally had more positive sentiment compared to horror or thriller genres.

---

### 6. Visualizations

- Sentiment score distribution across all movies.
- Revenue vs. Sentiment scatter plots.
- Genre-wise sentiment trends over years.
- Sample visualization:

---

### 7. Challenges

- Handling missing data and inconsistent metadata was time-consuming.
- Sentiment analysis accuracy limited by VADER's lexicon approach; some sarcasm or slang was missed.
- Revenue prediction influenced by many external factors not included in datasets.

## 8. Future Work

- Incorporate deep learning-based sentiment models (e.g., BERT) for improved accuracy.
- Integrate additional data sources such as critic reviews, social media sentiment, and marketing budgets.
- Develop an interactive dashboard using Streamlit or Dash for live data exploration.
- Expand prediction models to classify movies into success tiers rather than predicting exact revenue.

## 9. Conclusion

This project demonstrated that combining sentiment analysis of user reviews with movie metadata provides meaningful insights into movie success patterns. While the regression model's predictive power was moderate, it established a foundation for more advanced models and richer datasets to improve accuracy. The genre sentiment trends highlight how audience emotions differ across movie types, useful for marketing and production decisions.

## 10. References

- Kaggle TMDB Movie Metadata: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata
- Kaggle IMDB 50K Movie Reviews: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
- VADER Sentiment Analyzer: https://github.com/cjhutto/vaderSentiment
- Scikit-learn Documentation: https://scikit-learn.org/stable/

## 11. Appendix

- List of notebooks and their purpose:
  - 1_data_preprocessing.ipynb: Data cleaning and merging
  - 2_sentiment_analysis.ipynb: Sentiment scoring with VADER
  - 3_box_office_prediction.ipynb: Regression modeling for revenue prediction
  - 4_genre_sentiment_trends.ipynb: Visualization of sentiment trends by genre