Analysis of the Impact of Movie Titles and Genres on Market Success and Film Awards

Team: Laslo Uri E2 163/2023, Denis Dautović E2 166/2023, Nikola Janković E2 130/2025

1. Project Goal Definition

The goal of the project is to develop a system for **predicting the market success of movies** based on the analysis of movie title content and genre characteristics. The system uses natural language processing (NLP) and machine learning techniques for binary classification of movies, as well as for predicting the probability of nominations for film awards.

Specific tasks:

1. Classification of movies by market success (successful/unsuccessful) based on ROI
2. Prediction of the probability of Oscar nomination based on genre and title characteristics
3. Identification of which title characteristics (sentiment, lexical diversity, genre) most influence movie success

2. Problem Motivation

The film industry generates over 100 billion dollars annually, but investments in production carry high risk. A system for predicting movie success has practical value for:

- **Producers:** More informed investment decisions in projects
- **Streaming platforms:** Improved recommendations and prediction of content popularity
- **Researchers:** Quantitative analysis of the relationship between narrative content and audience reception

The combination of textual content analysis (titles) with genre and financial data represents an approach that is not sufficiently explored in existing literature.

3. Relevant Literature

3.1 Detecting Emotional Scenes Using Semantic Analysis on Subtitles [1]

Research goal: Development of a system for automatic detection of emotional scenes from movies using NLP techniques applied to subtitles.

Methodology: Two classifiers were implemented: (a) naive counting classifier based on word frequency and semantic constructs; (b) MaxEnt classifier with unigram analysis and emotional words.

Dataset: 8 movies of different genres (Gladiator, Love Actually, Remember the Titans, X-men trilogy, Transformers, Troy) with manually annotated ~8,000 sentences, grouped into segments of 20 sentences each.

Evaluation: Testing on each movie individually with manually annotated gold standard labels. Metric: classification accuracy.

Results: Naive classifier achieved average accuracy of 62% (max. 80%), while MaxEnt classifier reached 78% on Remember the Titans.

Relevance: The study demonstrates the efficiency of NLP techniques in analyzing movie subtitles. In our project we expand this approach by combining it with genre and financial characteristics, using TF-IDF instead of simple word counting.

3.2 Screenplay Quality Assessment: Can We Predict Who Gets Nominated? [2]

Research goal: Automatic evaluation of screenplay quality with focus on predicting Oscar nominations.

Methodology: TF-IDF text transformation, SVM classification, combination of linguistic, emotional and structural characteristics. Comparison with deep learning models (BERT).

Dataset: ScriptBase corpus (897 screenplays) and Movie Screenplay Corpus (868 screenplays). Binary labels: nominated/not nominated.

Evaluation: Split 80/10/10 (training/validation/test). Metrics: F1-score, precision, recall.

Results: Feature-based approach (TF-IDF + SVM) outperformed deep learning methods (BERT) with F1-score of 62.35% (ScriptBase) and 64.79% (MSC).

Relevance: Directly confirms the possibility of predicting Oscar nominations based on text. We will use similar feature engineering techniques, but instead of screenplays we analyze subtitles and add financial data. We will use ensemble methods instead of just SVM.

3.3 Success in Books: A Big Data Approach to Bestsellers [3]

Research goal: Analysis of sales patterns of books from the New York Times bestseller list and development of a model for predicting total sales based on early sales data.

Methodology: Statistical model based on three mechanisms: (1) book fitness (quality), (2) preferential attachment (popularity attracts more buyers), (3) aging (decline of interest over time). Combination of these factors into a single formula for sales prediction.

Dataset: 2468 fiction and 2025 nonfiction bestsellers from NYT list (2008-2016), with weekly sales data from NPD BookScan.

Evaluation: Model fitting on first 25-50 weeks of sales, prediction of total sales. Metric: R^2 for fit quality.

Results: Model accurately predicts total book sales based on first 6 months ($R^2 > 0.9$). Fiction sells better than nonfiction. Initial position on the list predicts duration of stay.

Relevance: Although it deals with books, the principles of analyzing success of cultural products are applicable to movies. We will use a similar approach for analyzing the relationship between genre and success, but focused on prediction BEFORE movie release (not post-hoc sales analysis).

---

4. Dataset

We will compile the dataset ourselves by combining multiple sources:

Data sources:

- **TMDB API** (themoviedb.org) — title, year, genres, budget, box office revenue
- **OpenSubtitles API** (opensubtitles.stoplight.io) — movie subtitle texts
- **Academy Awards Database** (awardsdatabase.oscars.org) — historical Oscar nomination data

Prediction attributes:

- Movie genres (action, drama, comedy, horror...)
- NLP subtitle characteristics (TF-IDF, sentiment, lexical diversity)
- Movie budget (log-transformed)

Target labels:

- **Market success:** Class 1 if ROI ≥ 2.0 (revenue ≥ 2× budget), otherwise class 0. Labels calculated automatically from budget and revenue.
- **Oscar nomination:** Class 1 if movie is nominated, otherwise class 0. Labels taken from Academy Awards Database.

Expected size: ~5,000-10,000 movies (2000-2024) with complete data (budget, revenue and subtitles). We expect imbalanced classes — most movies are neither successful nor nominated.

---

5. Methodology

After collecting movie data, we will combine data from TMDB API (title, genre, budget, revenue) with subtitles from OpenSubtitles using title and year as keys. Movies without complete data (missing budget, revenue or subtitles) will be excluded from analysis.

Feature extraction from subtitles: For each movie we will extract from subtitle text: (1) TF-IDF vectors of most frequent words and phrases appearing in dialogues, (2) sentiment score measuring whether dialogues are predominantly positive or negative, (3) lexical diversity showing vocabulary richness in the movie. These characteristics will be combined with movie genre (action, drama, comedy...) which we will encode as binary attributes.

Modeling: Input to classifiers will be: movie genres, NLP subtitle characteristics and budget ratio. Output is binary prediction: successful (ROI ≥ 2) or unsuccessful movie. We will compare Random Forest, XGBoost and SVM as these models have proven effective for textual characteristics in relevant literature [2]. Given that we expect more unsuccessful than successful movies, we will apply SMOTE for class balancing before training.

Results analysis: We will analyze which subtitle characteristics most contribute to success prediction — e.g., whether movies with more positive sentiment have higher ROI, or whether certain genres (action vs drama) show different success patterns.

---

6. Evaluation Method

6.1 Evaluation Procedure

Data will be split into training (70%), validation (15%) and test set (15%) with stratified sampling to maintain class proportions. For more robust performance assessment we will use 5-fold stratified cross-validation on the training set, while final evaluation will be performed on a separate test set that the model has never seen.

6.2 Evaluation Metrics

Given that we expect imbalanced classes (more unsuccessful movies than successful, more non-nominated than nominated), **accuracy is not an adequate metric**. As the primary metric we will use **F1-score** as it balances precision and recall, which is crucial when the minority class (successful movies / nominated movies) is of greater interest. Additionally, we will monitor **AUC-ROC** for assessing the model's ability to rank movies by success probability.

For comparing different models (Random Forest vs XGBoost vs SVM) we will use paired t-test on cross-validation results. We will analyze the confusion matrix to identify error types — e.g., whether the model more often misclassifies low-budget hits or high-budget flops.

7. Software (Optional)

- **Programming language:** Python 3.10+ with Jupyter Notebook environment
- **Data wrangling:** Pandas, NumPy
- **ML:** Scikit-learn (Random Forest, SVM), XGBoost, Imbalanced-learn
- **NLP:** SpaCy, NLTK, Gensim (Word2Vec)
- **Visualization:** Matplotlib, Seaborn, Plotly

8. Work Plan

Milestones:

- **First milestone:** Complete dataset, initial EDA, baseline model
- **Second milestone:** Optimized models, evaluation, error analysis
- **Final submission:** Complete system with documentation and report

Potential risks: Missing data for some movies, computational complexity of NLP processing, imbalanced classes. Mitigation: data filtering, batch processing, SMOTE/undersampling.

References

[1] Detecting Emotional Scenes Using Semantic Analysis on Subtitles, Stanford NLP Course Project, 2009. https://nlp.stanford.edu/courses/cs224n/2009/fp/6.pdf

[2] M.-C. Chiu, T. Feng, X. Ren, S. Narayanan, "Screenplay Quality Assessment: Can We Predict Who Gets Nominated?", arXiv:2005.06123, 2020. https://arxiv.org/abs/2005.06123

[3] Y. Yucesoy et al., "Success in books: a big data approach to bestsellers", EPJ Data Science, vol. 7, 2018. https://link.springer.com/article/10.1140/epjds/s13688-018-0135-y