Movie Sentiment Intelligence Platform

CAP5771 Spring 2025 Project Report

Sai Nilasha Varma Indukuri

April 24, 2025

Objective

The aim of this project is to develop an end-to-end machine learning pipeline that applies natural language processing to analyze sentiment from movie descriptions. The ultimate goal is to assist content strategists and media professionals in making data-driven decisions about film narratives by assessing emotional tone.

Methodology: CRISP-DM

This project follows the CRISP-DM (Cross Industry Standard Process for Data Mining) framework to ensure reproducibility and scalability.

Project Milestones

Milestone 1: Data Collection, Preprocessing and Exploratory Analysis

- Merged datasets from IMDb, Netflix, Disney+, Hulu, and Amazon Prime
- Cleaned data: removed missing values, standardized formats
- Performed exploratory data analysis (EDA) with visualizations and correlations

Milestone 2: Feature Engineering and Modeling

- Engineered features: popularity score, ROI, is_franchise
- Applied TF-IDF vectorization on movie overviews
- Trained models: Logistic Regression, SVM, Random Forest, MLP
- Performed hyperparameter tuning using GridSearchCV

Milestone 3: Evaluation and Deployment

- Evaluated RandomForestRegressor using MAE, RMSE, and R^2
- Built an interactive dashboard using Streamlit
- Delivered a demo video, PDF report, and presentation slides

Technology Stack

Component	**Tools Used**
Language	Python 3.10
Data Handling	pandas, NumPy
Modeling	scikit-learn, joblib
NLP	TfidfVectorizer
Visualization	matplotlib, seaborn
Interface	Streamlit

Data Sources

- IMDb movie metadata and overviews
- Netflix, Disney+, Hulu, Amazon Prime metadata
- Top Movies revenue and rating dataset (Kaggle)

Model Evaluation

Metric	Value
MAE	≈ 0.087
RMSE	≈ 0.165
R^2 Score	≈ 0.84
Model	RandomForestRegressor (100 trees)

Key Insights

- Sentiment scores vary significantly by genre
- TF-IDF embeddings captured strong narrative patterns
- Overview length and tone correlate with sentiment

Tool Features

- Real-time sentiment prediction based on input plot summaries
- Dashboard sections: Home, Model Evaluation, Live Prediction, About
- UF Gator-branded design with clean UI and metrics

Deliverables

- Trained model and vectorizer (`.pkl`)
- Interactive `app.py` Streamlit dashboard
- Milestone reports (`Milestone1.pdf`, `Milestone2.pdf`, `Milestone3.pdf`)
- GitHub repository with all assets
- Demo video and presentation slides