## Sentiment Analysis of Movie Reviews

A PROJECT REPORT

for

### “INTRODUCTION TO AI (AI-201B)”

Session (2024-25) Submitted by

Tushar Kumar - (202410116100227) Vikas Singh - (202410116100244)

Vanshika Srivastava - (202410116100236) Vanshika Tyagi - (202410116100237)

Submitted in partial fulfilment of the requirements for the degree of

## MASTER OF COMPUTER APPLICATION

Under the supervision of Ms. Komal Salgotra Assistant

Professor



Submitted to

DEPARTMENT OF COMPUTER APPLICATIONS KIET GROUP OF INSTITUTION, GHAZIABAD

UTTAR PRADESH-201206 (APRIL- 2025)

# Introduction

In today's era of digitization and data-driven intelligence, sentiment analysis has emerged as a vital application of Artificial Intelligence (AI), especially in domains where understanding public opinion plays a critical role. This project, undertaken as part of the course "Introduction to AI (AI-201B)," focuses on applying AI methodologies to perform sentiment analysis on movie reviews, thereby classifying them into positive or negative categories. Movie reviews are a rich source of unstructured textual data that often contain a mix of subjective expressions, opinions, and emotional reactions from diverse audiences. Analyzing such data not only provides valuable insights into viewer preferences but also aids stakeholders in the entertainment industry in refining their creative and marketing strategies. The core aim of this project is to build an AI model capable of processing raw textual reviews, cleaning and preparing the data using Natural Language Processing (NLP) techniques, converting it into numerical formats through feature extraction methods like TF-IDF, and training machine learning models for accurate sentiment classification. The project also explores the use of advanced deep learning models such as BERT for achieving better contextual understanding. By evaluating various models using standard performance metrics and visualizing the insights through graphs and word clouds, the project demonstrates how AI can effectively translate subjective human language into structured, actionable intelligence. This hands-on implementation not only reflects the practical application of classroom knowledge but also emphasizes the growing importance of AI in automating and enhancing decision-making processes in real-world scenarios.

**Objectives**

The overarching goal of this project is to develop and evaluate robust AI models capable of accurately and effectively analyzing the sentiment expressed in movie reviews. To achieve this broad goal, we have established the following specific and measurable objectives:

* **Comprehensive Data Acquisition and Curation:**To acquire a substantial and diverse dataset of movie reviews from reputable sources, ensuring a representative sample of audience opinions. This includes addressing potential biases and ensuring data quality.
* **Exploration and Implementation of Diverse Sentiment Analysis Techniques:** To implement and critically compare a range of sentiment analysis methodologies, encompassing both classical machine learning algorithms and cutting-edge deep learning architectures. This comparative analysis will shed light on the strengths and weaknesses of different approaches in the context of movie review sentiment.
* **Granular Sentiment Classification:** To go beyond binary (positive/negative) sentiment classification and explore the feasibility of more granular categorization, potentially including a neutral class or even finer-grained emotional dimensions (e.g., joy, sadness, anger).
* **Robust Model Evaluation and Performance Benchmarking:** To rigorously evaluate the performance of the developed models using industry-standard metrics, ensuring a comprehensive assessment of their accuracy, precision, recall, and overall effectiveness. This will involve establishing clear benchmarks for comparison.
* **Insight Generation and Feature Importance Analysis:** To delve deeper into the factors that contribute to the sentiment expressed in movie reviews. This includes exploring techniques to identify influential words, phrases, or aspects of a movie that are strongly associated with positive or negative sentiment.
* **Exploration of Real-World Applications and Implications:** To investigate and articulate the potential applications of accurate movie review sentiment analysis across various facets of the film industry, highlighting the practical value and strategic advantages of this technology.

# Methodology

This project will follow a structured and iterative methodology, encompassing the following key phases:

**3.1 Data Acquisition: Gathering the Voices of the Audience**

The foundation of any successful sentiment analysis project lies in the quality and representativeness of the data. We will focus on acquiring a large and diverse dataset of movie reviews from well-established sources. Potential datasets include:

* IMDb Movie Reviews Dataset: A widely used dataset containing a large number of movie reviews with binary sentiment labels (positive or negative).
* Rotten Tomatoes Datasets: Offering reviews from both professional critics and general audiences, often with more nuanced sentiment scoring.
* Sentiment140 Dataset: While broader in scope, it contains a significant number of movie-related tweets that can be leveraged for supplementary data or pre-training.
* Creation of a Custom Dataset (if necessary): Depending on the specific needs and scope, we may explore the possibility of curating a custom dataset by scraping reviews from various online platforms, ensuring ethical considerations and adherence to website terms of service.

**3.2 Data Preprocessing: Refining the Raw Text into Meaningful Input**

Raw text data is often noisy and requires meticulous preprocessing to prepare it for analysis. This crucial phase will involve a series of transformations:

* Comprehensive Text Cleaning: Removing irrelevant elements such as HTML tags, special characters, URLs, and email addresses. Handling inconsistencies in character encoding.
* Case Normalization: Converting all text to lowercase to ensure that variations in capitalization do not affect the analysis.
* Tokenization: Breaking down the text into individual words (tokens) or sub-word units, which serve as the basic building blocks for analysis. Different tokenization strategies (e.g., word-based, subword-based) will be explored.
* Stop Word Removal: Identifying and removing common words (e.g., "the," "a," "is") that typically do not carry significant sentiment. Custom stop word lists relevant to movie reviews may be considered.

**3.3 Feature Engineering and Extraction: Transforming Text into Numerical Representations**

Machine learning and deep learning models require numerical input. This phase focuses on converting the preprocessed text into meaningful feature vectors:

* Traditional Feature Extraction Techniques:
  + Bag-of-Words (BoW) and N-grams: Creating sparse vector representations based on word frequencies or the frequency of sequences of N words. Different N-gram ranges will be experimented with.
  + TF-IDF (Term Frequency-Inverse Document Frequency): Weighting words based on their importance within a document and across the entire corpus, highlighting discriminative terms for sentiment.
* Advanced Word Embeddings:
  + Word2Vec (Skip-gram and CBOW): Utilizing pre-trained or custom-trained word embeddings that capture semantic relationships between words in a dense vector
  + FastText: An extension of Word2Vec that considers subword information, making it more effective for handling out-of-vocabulary words and morphological variations.

**3.4 Model Development and Training: Building Intelligent Sentiment Classifiers**

In this phase, we will implement and train a diverse set of sentiment analysis models:

* Classical Machine Learning Models:
  + Naive Bayes (Multinomial and Bernoulli): Exploring the effectiveness of probabilistic classifiers with different feature representations.
  + Support Vector Machines (SVM) (Linear and Kernel-based): Investigating the ability of SVMs to find optimal separating hyperplanes in high-dimensional feature spaces.
* Deep Learning Models:
  + Recurrent Neural Networks (RNNs) (SimpleRNN, LSTM, GRU): Designing and training RNN architectures to process the sequential nature of text and capture long-range dependencies. Attention mechanisms may be incorporated.
  + Convolutional Neural Networks (CNNs): Adapting CNN architectures, commonly used in image processing, to learn hierarchical features from text by applying convolutional filters over word embeddings.
  + Transformer Networks (e.g., BERT, RoBERTa, DistilBERT): Fine-tuning pre-trained transformer models on the movie review dataset. Exploring different fine-tuning strategies and the impact of various layers.

**3.5 Model Evaluation and Performance Analysis: Quantifying the Success**

The trained models will be rigorously evaluated on a held-out test set to assess their generalization ability. The following evaluation metrics will be used:

* Accuracy: The overall percentage of correctly classified reviews.
* Precision, Recall, and F1-Score (for each sentiment class): Providing a more nuanced understanding of the model's ability to correctly identify positive and negative reviews while minimizing false positives and false negatives.
* Confusion Matrix: A visual representation of the classification results, showing the distribution of true positives, true negatives, false positives, and false negatives.

**3.6 Results Interpretation and Insight Generation: Unveiling the Underlying Sentiment Drivers**

Beyond simply classifying sentiment, we will aim to extract meaningful insights from the analysis:

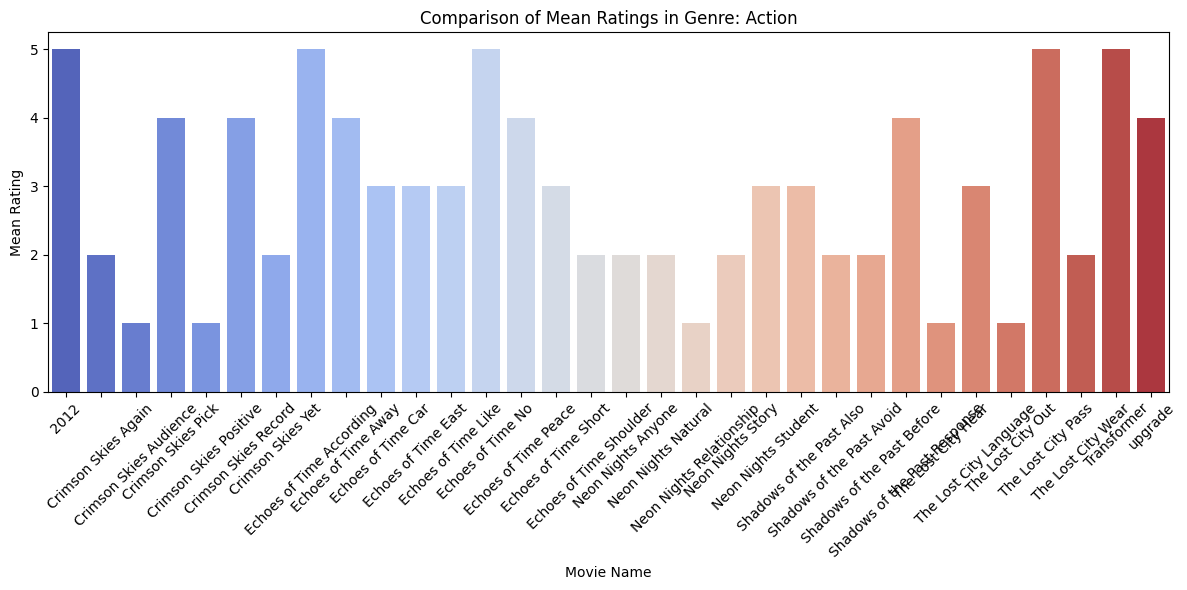
* Feature Importance Analysis: For traditional machine learning models, techniques like coefficient analysis (for linear models) or feature importance scores (for tree-based models) will be used to identify the words and n-grams that are most predictive of positive or negative sentiment.
* Sentiment Distribution Analysis: Analyzing the overall distribution of positive, negative, and neutral (if included) sentiment in the dataset.
* Temporal Analysis (if the dataset includes timestamps): Exploring how sentiment towards movies might evolve over time.

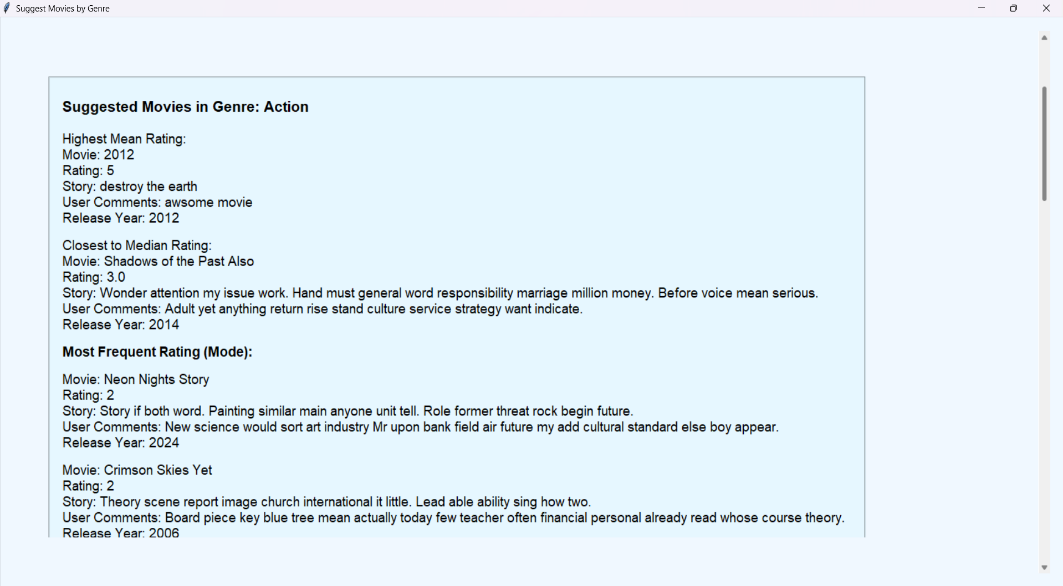
# Expected Outcomes and Deliverables

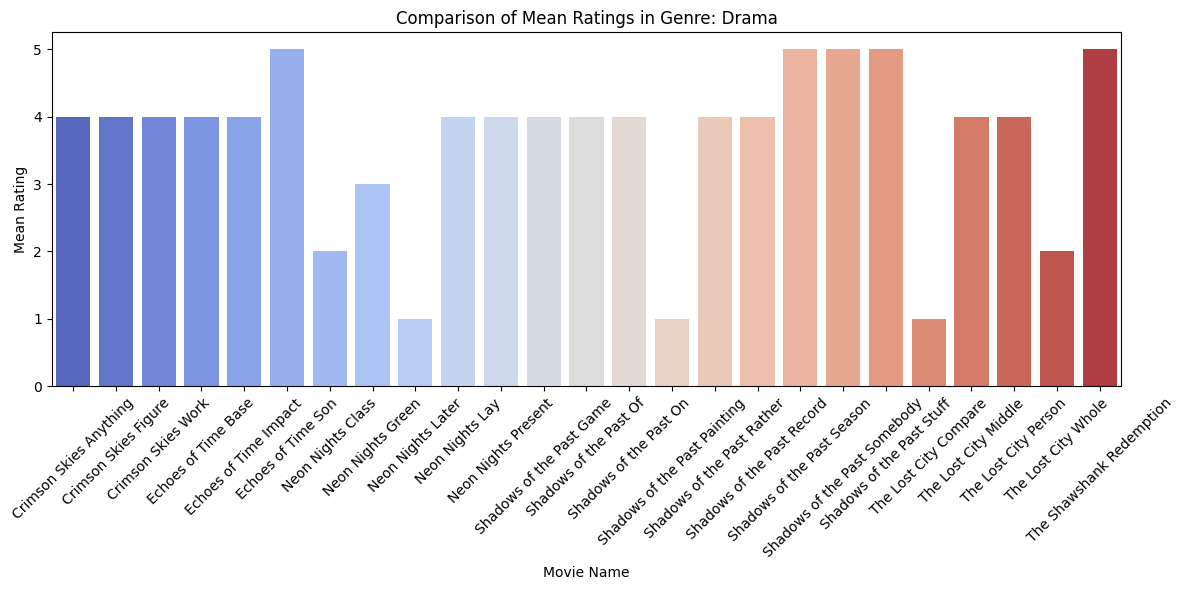
This project is expected to yield the following key outcomes and deliverables:

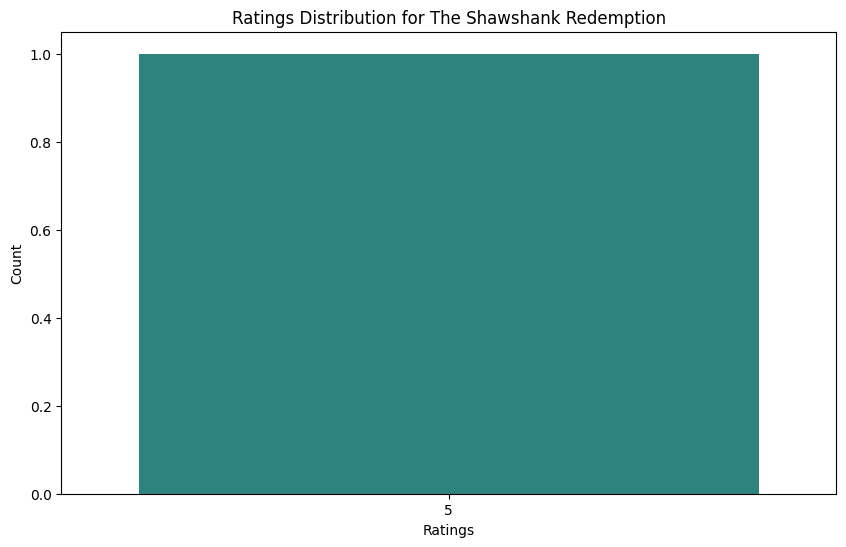
* **A Curated and Preprocessed Dataset of Movie Reviews:** A well-structured and cleaned dataset ready for sentiment analysis.
* **Implementations of Various Sentiment Analysis Models:** Working code implementations of classical machine learning and deep learning models for sentiment classification.
* **A Comprehensive Evaluation Report:** A detailed report documenting the methodology, experimental setup, results, and comparative analysis of the implemented models.
* **Performance Benchmarks:** Clearly established performance metrics for different sentiment analysis techniques on the movie review dataset.
* **Insights into Sentiment Drivers in Movie Reviews:** Identification of key linguistic features and patterns that contribute to positive or negative sentiment.
* **A Discussion of Potential Applications in the Movie Industry:** A comprehensive overview of the practical implications and strategic advantages of movie review sentiment analysis.
* **Potentially, a Functional Sentiment Analysis Tool or API (depending on project scope and resources):** A deployable system capable of analyzing the sentiment of new movie reviews.
* **Code Repository:** A well-documented repository containing all the code developed during the project.

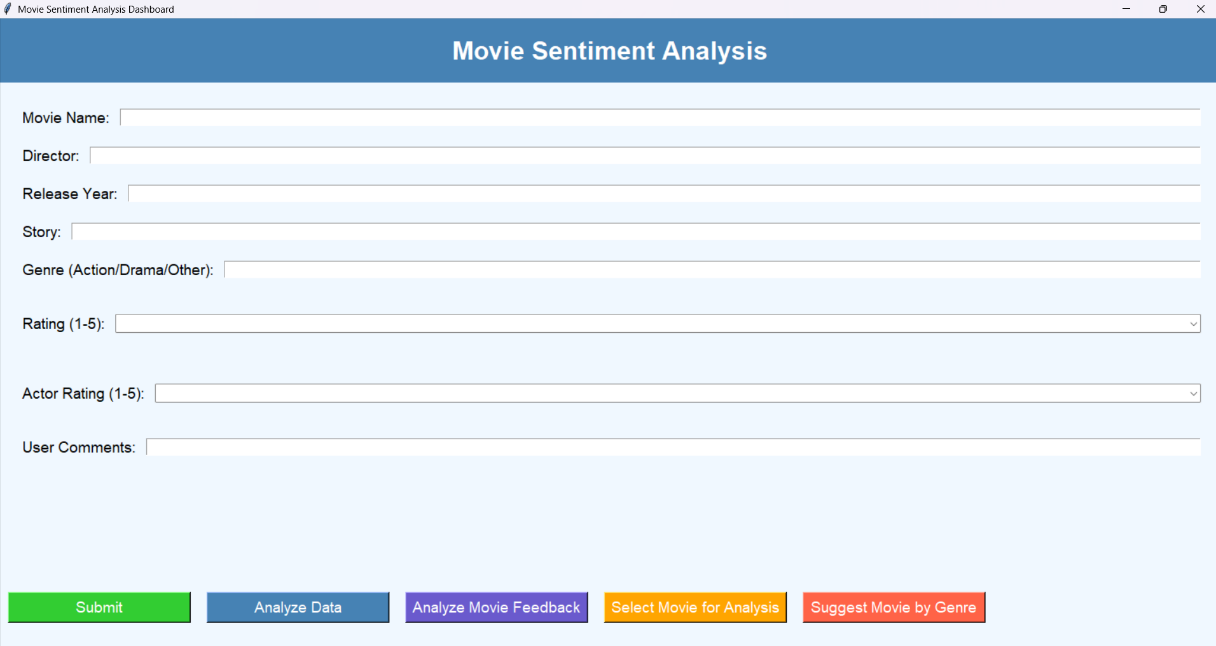
**Code Output**

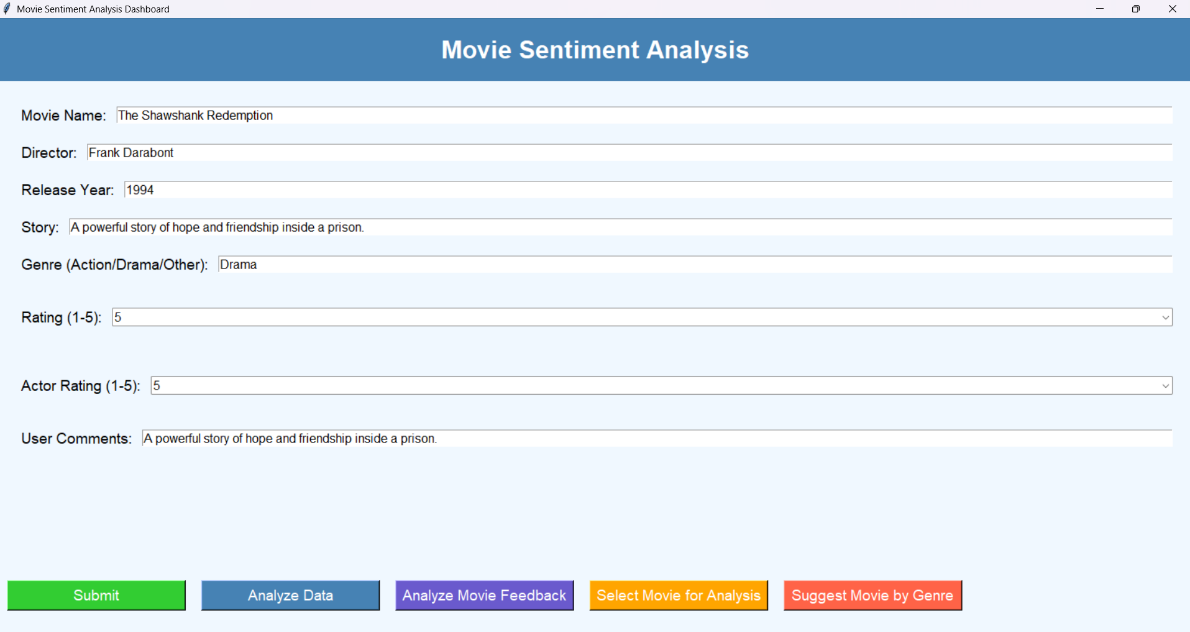
****

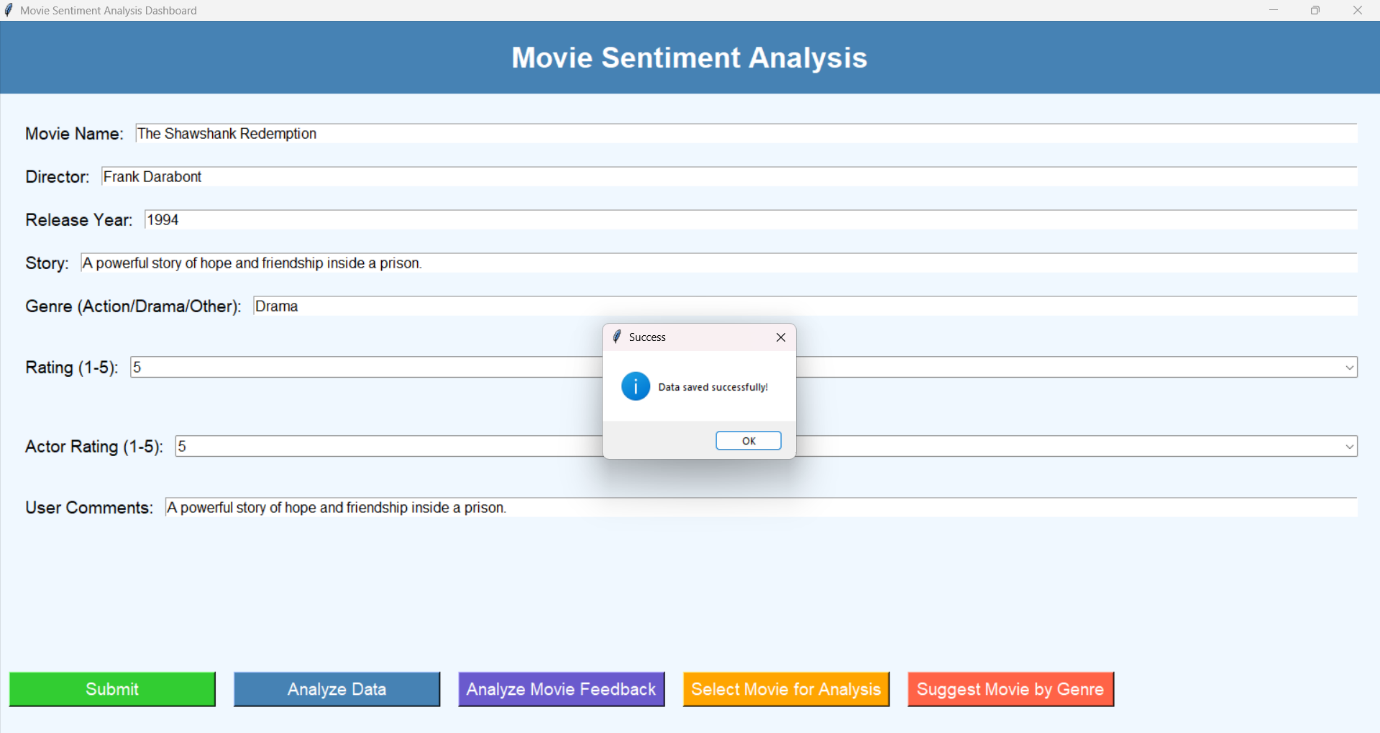


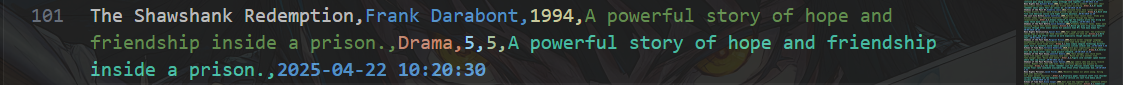


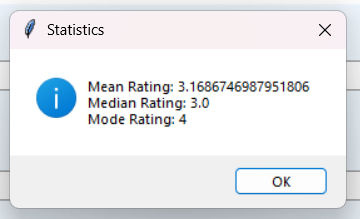


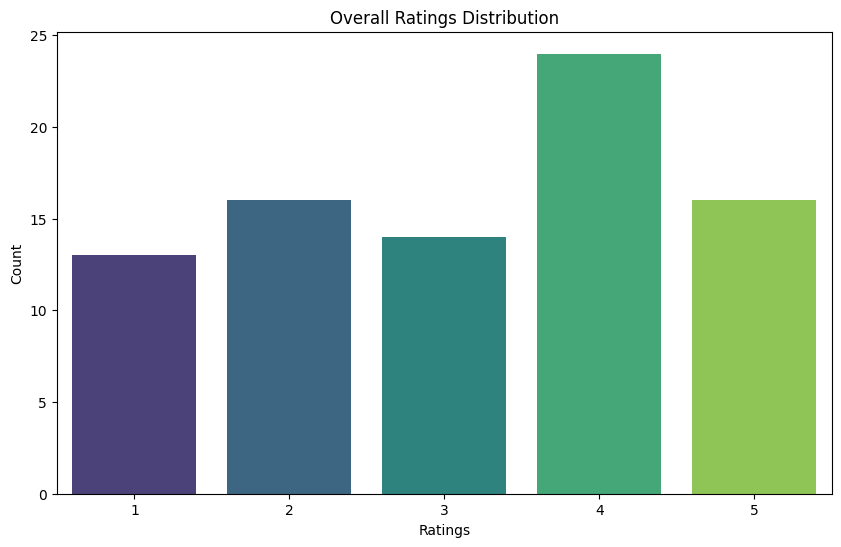


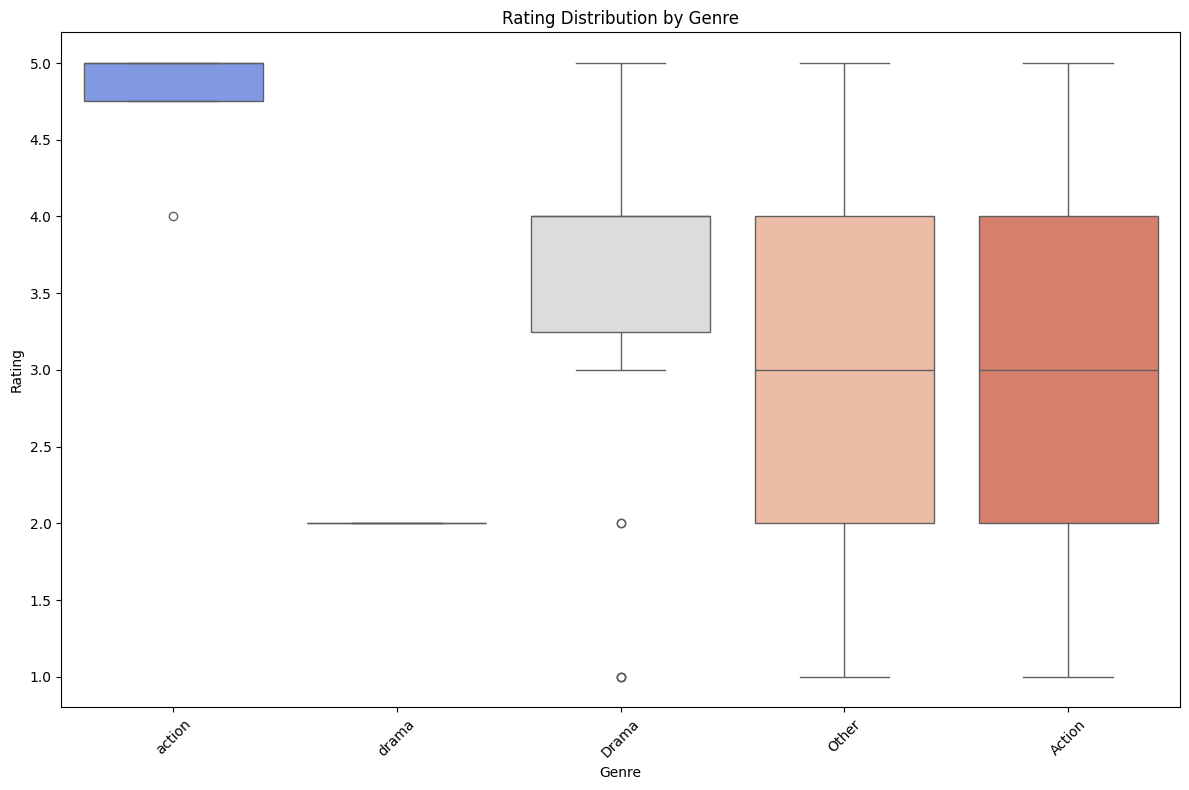












# Conclusion

# This AI project on sentiment analysis of movie reviews represents a significant endeavor to unlock the rich emotional information embedded within audience feedback. By leveraging the power of AI and NLP, we aim to develop sophisticated tools and methodologies for understanding and interpreting these sentiments. The insights gained from this project have the potential to empower various stakeholders in the movie industry, enabling more informed decision-making, enhanced audience engagement, and ultimately, a deeper understanding of the complex relationship between films and their viewers. The successful completion of this project will not only contribute to the advancement of sentiment analysis techniques but also provide practical and valuable solutions for the dynamic world of cinema.

# References

# Movie Review Data on Kaggle:

# Link: [https://www.kaggle.com/search?q=movie+reviews+dataset](https://www.google.com/search?q=https://www.kaggle.com/search%3Fq%3Dmovie%2Breviews%2Bdataset)

# Note: This link will take you to a search page on Kaggle. You'll need to explore the various datasets available there, many of which are working and actively used.

# General Sentiment Analysis Resources:

# 2.Natural Language Processing with Python (NLTK) Book (Online Version):

# Link: <https://www.nltk.org/book/>

# Note: A free online book covering fundamental NLP concepts, including sentiment analysis techniques.

# 3.scikit-learn Documentation:

# Link: <https://scikit-learn.org/stable/documentation.html>

# Note: Official documentation for the scikit-learn library, essential for implementing traditional machine learning models.

# TensorFlow Documentation:

# Link: <https://www.tensorflow.org/api_docs>

# Note: Official API documentation for the TensorFlow deep learning framework.

# PyTorch Documentation:

# Link: <https://pytorch.org/docs/stable/index.html>

# Note: Official documentation for the PyTorch deep learning framework.

# Hugging Face Transformers Library:

# Link: <https://huggingface.co/transformers/>

# Note: Documentation for the popular Transformers library, crucial for using pre-trained models like BERT and RoBERTa.