Prediction of Movie Box Office Success using YouTube Trailer Comments and Wikipedia Data

by

Tasin Mohammad
20341021
Maliha Binta Islam
20101587
Humayra Binte Jamal
20101591

A Thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2024

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Tasin Mohammad

20341021

Maliha Binta Islam

20101587

Humayra Binte Jamal

20101591

Approval

The thesis titled "Prediction of Movie Box Office Success using YouTube Trailer Comments and Wikipedia Data" submitted by

1. Tasin Mohammad (20341021)
2. Maliha Binta Islam (20101587)
3. Humayra Binte Jamal (20101591)

Of Fall 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 2024.

Examining Committee:

Supervisor:
(Member)

Najeefa Nikhat Choudhury

Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD

Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The movie industry is one of the most highly competitive and profitable businesses in the entertainment world, where predicting a movie's success or failure of a film is a difficult task. Along with the movie trailer, key factors such as - the popularity of actors and actresses, comments, ratings, marketing budgets, critical receptions, likes, and dislikes play a significant role in determining a movie's box office performance. Day by day, the use of social media is increasing. People express their opinions and feelings about anything on social media platforms such as Facebook, Instagram, Twitter, and YouTube. This research paper aims to predict movie box office success using opinions expressed in comments on movie trailers and Wikipedia data.

To predict a movie's box office performance, a dataset containing YouTube trailer comments and Wikipedia data will be collected through Web Scraping using Python's BeautifulSoup library. Machine learning algorithms like Random Forest, Decision Tree, and Logistic Regression will then be implemented to make predictions. This research aims to give film production studios, distributors, and other movie industry collaborators an early insight into a film's box office capabilities.

Keywords: Prediction, Movie box office success, YouTube trailer comments, Wikipedia data, Web scraping, Sentiment analysis, Machine learning, Natural Language Processing

Table of Contents

Chapter 1

Introduction

The movie-going experience for many people has long been the preferred choice when it comes to spending quality time with friends and family. This makes the film industry one of the most rewarding businesses in the entertainment world. The success of a movie can depend on many factors, such as the cast, budget, director, and trailer. This research will use Wikipedia data and YouTube trailer comments to predict a movie's success.

YouTube is among the most popular platforms for movie studios to upload trailers for upcoming films. Moviegoers engage in passionate discussions regarding a particular film and express their opinions and feelings in the comment section.

On the other hand, Wikipedia is one of the most used sites to access information about movies, actors, directors, ratings, and other film industry news.

This research paper aims to use YouTube trailer comments and Wikipedia data to predict the box office success of a movie. Sentiment analysis will be used to categorize YouTube comments and gain insight into the overall reaction of audiences to a film's trailer. This, combined with other factors such as the cast, director, distributor, and production studio, will predict whether a film will succeed at the box office.

Chapter 2

Problem Statement

Predicting a movie's box office success is a complex and challenging task with significant implications for the film industry, such as allowing production companies to allocate resources effectively and make informed decisions. Despite the availability of various data sources, such as YouTube trailer comments and Wikipedia data, there is a lack of comprehensive research exploring their potential to accurately forecast the commercial performance of movies, with most existing approaches primarily relying on traditional marketing strategies and limited data sources, leading to suboptimal predictive models. Therefore, the problem addressed in this thesis is the absence of a robust predictive model that leverages YouTube trailer comments and Wikipedia data to predict the box office success of movies effectively. This study aims to develop a reliable model that can provide valuable insights to filmmakers, distributors, and investors to enable them to make more informed decisions regarding the types of movies they make and the cast and crew they hire for a particular film, backed up by audience approval.

This research aims to address the following challenges:

- **Neglecting the potential of YouTube trailer comments:** YouTube has become a major platform for movie promotion, and user comments on movie trailers reflect audience reactions and expectations. However, most existing prediction models overlook the valuable information embedded in these comments. Therefore, leveraging YouTube trailer comments as a data source and extracting meaningful insights can enhance the accuracy and granularity of box office predictions.

- **Insufficient integration of Wikipedia data with contemporary approaches:** Wikipedia provides much information about movies, including director, release date, and cast details. While previous prediction models utilized Wikipedia data, they often lack integration with newer data sources or advanced machine-learning techniques. This study aims to explore innovative ways to combine Wikipedia data with YouTube trailer comments and employ advanced predictive algorithms to improve the accuracy and reliability of box office predictions.

Chapter 3

Research Objectives

This research aims to develop a predictive model to predict the box office success of movies using a combination of YouTube trailer comments and Wikipedia data. The main aim is to study the relationship between audience feedback from YouTube comments on movie trailers and the film's subsequent box office performance. This research further aims to analyze the traditional influences on a movie's box office, release date, budget, writer, director and cast members from Wikipedia.

The research objectives are as follows:

1. Explore the potential of YouTube trailer comments as a valuable source of information for predicting movie box office success. By analyzing the sentiment of comments, this study aims to determine if there are significant correlations between user-generated responses and the commercial performance of films.

2. Investigate the role of traditional movie attributes from Wikipedia in conjunction with YouTube trailer comments. By combining these two data sources, the research aims to enhance the prediction accuracy of box office success.

3. Develop a predictive model using machine learning algorithms to predict the box office performance of a movie. The study will evaluate classification techniques, such as Random Forest, Decision Tree, and Logistic Regression, to identify the most suitable approach for predicting movie revenues based on the given features.

4. Evaluate the performance of the proposed algorithms through extensive experimentation and validation. The model will be trained and tested on a diverse dataset of movies spanning different release dates and production budgets. The accuracy, precision, recall, and other relevant metrics will be measured to assess the reliability and effectiveness of the model.

5. Provide insights and recommendations based on the research findings to assist movie studios, distributors, and marketing professionals make informed decisions regarding resource allocation, promotional strategies, and release planning. The aim is to offer actionable recommendations that can improve the chances of achieving box office success for future movie releases.

By accomplishing these objectives, this research aims to contribute to the field of movie industry analysis by leveraging user-generated content on YouTube and traditional movie attributes from Wikipedia to build an accurate and robust predictive model to estimate the box office success of movies.

Chapter 4

Literature Review

In the study by Apala, Krushikanth & Jose, Merin & Motnam, Supreme & Chan, Chien-Chung & Liszka, Kathy & de Gregorio, and Federico [1], data mining is applied to forecast the achievement of the movies. Data is collected from social media like IMDb, Twitter, and Youtube. This paper used a code snippet in PHP to collect data from Youtube, which ran by the Wamp server. From IMDb, they collected the top 50 genres of the movie; from Twitter, they collected the popularity of directors, actors, and actresses; from Youtube, they collected the likes, dislikes, views, and comments. Sentiment analysis is applied to the comments that have been extracted from Youtube. It used a simple min-max method for normalizing the training data, a K-means tool from Weka for clustering, and then created a training set for generating the predictive model. This study showed that uniform and non-uniform weighted approaches produced by Weka (Naive Bayesian classifier and J-48 classifier) could be used to predict a movie's success. The potential gap of this study is that it did not demonstrate which algorithm will give the highest accuracy and relied on a single machine learning algorithm. The study divided the movies into three categories: hit, flop, and neutral but didn't clarify which movie was a hit or flop. Overall, we can conclude that the study only considers the popularity of an actor/actress but does not count views, likes, dislikes, and comments to predict the favorable outcome of a movie.

Vasu Jain [2] worked on sentiment analysis to determine the sentiment expressed in tweets related to movies. In this study, to predict the movie's box office success, sentiment analysis results were used from tweets during a film's release. The sentiments were classified into four categories, which were positive, negative, neutral, and irrelevant. A classifier was trained using the Lingpipe sentiment analyzer, which utilizes an 8-gram language model on character sequences. 48 movies were correctly classified as hits. However, no other factors except Tweet sentiments were considered when making predictions, which makes the study very limited and leaves room for improvements.

M. S. Neethu and R. Rajasree [3] focused on sentiment analysis, particularly on analyzing sentiments expressed in Twitter posts about electronic products using a machine learning approach. The study identified some of the challenges with analyzing the sentiments of Tweets, which include the presence of slang words, misspellings, and the constraint of a 140-character limit per Tweet. The study compares the performance of three types of basic classifiers (SVM, Naive Bayes, and Maximum Entropy) and an ensemble classifier for sentiment classification. SVM and Naive Bayes classifiers were implemented using built-in functions in Matlab, while the Maximum Entropy classifier was implemented using MaxEnt software. The Naive Bayes classifier exhibited better precision than the other three classifiers but slightly lower accuracy and recall.On the other hand, SVM, Maximum Entropy Classifier, and ensemble classifiers demonstrated similar accuracy, precision, and recall. These classifiers achieved an accuracy of 90%, while Naive Bayes achieved an accuracy of 89.5%. The study concludes that the chosen feature vector for the product domain plays a crucial role in sentiment analysis. Despite the variation in classifier performance, the quality of the feature vector enables better sentiment analysis results. They emphasize that the feature vector's effectiveness does not depend on the specific classifier.

M. S. Usha and M. Indra Devi (2013) [4] propose a Combined Sentiment Topic (C.S.T.) model, which offers a novel approach to sentiment analysis by simultaneously detecting sentiments and topics in the text. Unlike existing supervised and semi-supervised learning methods that rely on labeled documents for classification, the C.S.T. model is purely based on unsupervised learning techniques and utilizes unlabeled documents for classification. This unsupervised nature of the C.S.T. model enhances its portability to different domains. The paper reports that the C.S.T. model outperforms existing semi-supervised approaches, demonstrating its effectiveness in sentiment analysis tasks. One limitation of the C.S.T. model is its inability to detect neutral opinions, as it currently focuses only on classifying positive and negative sentiments.

Vr, Nithin & Pranav, M & Babu, PB & Lijiya, A. [5] focus on forecasting the movie's success depending on IMDb data. This study collected the dataset from IMDb, Wikipedia, and Rotten Tomatoes. This study only prioritizes the movie which was released between 2000-2012 and in the English language. As items can be duplicated, they use the mean and median methods as central tendencies while the data preprocessing time. This study only worked with a numerical value. That's why the central tendency is used to convert nominal attributes to numerical ones in data integration and transformation time. Different machine learning algorithms, such as - Linear Regression, Logistic Regression, and Support Vector Machine Regression (SVN) models have been run to predict how the movie is going to be. It shows that the linear regression model has the highest accuracy in predicting movie success among other machine learning algorithms. One potential gap in the study is that this study only concentrates on IMDb and Wikipedia data. Other data, such as the youtube trailer, comments under that trailer, Twitter, and social media comments, were ignored. These factors play an important role in identifying whether the movie is going to be successful or not. Briefly, this study gives an idea of predicting the success of movies by collecting the dataset from IMDb, Wikipedia, and Rotten Tomatoes and solving the dataset using machine learning algorithms. Also, it can be seen that among the 20 features in the dataset, actor1, writer, actor2, director, budget, and reviews played the most significant features. Briefly, it gives an idea of how to forecast whether a movie is going to be successful or not by collecting the dataset from Rotten Tomatoes, IMDb, and Wikipedia and solving the dataset using machine learning algorithms.

A. Bhave, H. Kulkarni, V. Biramane, and P. Kosamkar [6] highlighted the importance of considering both classical and social factors while predicting a movie's achievement at the box office. Factors such as cast, producer, and director are considered classical, while online responses to a film and a film's engagement on social media platforms are considered social factors. The study argues that integrating classical and social factors can lead to more accurate results when predictions are made regarding the success of a film. The study suggests considering YouTube view count and comments, sentiment analysis of Tweets regarding a film, and Wikipedia view and edit counts to improve the prediction success rate.

R. Dhir and A. Raj [7] analyzed the IMDb dataset and provided a prediction of IMDb scores for movies. The dataset, which includes information on 5,043 movies spanning 100 years and 66 nations, is analyzed using machine-learning methods. Then the outcomes from the methods are compared to decide the most successful one. Various machine learning techniques like K-Nearest Neighbors, Support Vector Machine, Gradient Boost, Ada Boost, and Random Forest are used here to analyze the dataset. Being one of the most informative resources, IMDb covers many variables like movie titles, duration, genre, release time, director names, star popularity, and movie rating from critics. Additional factors like the number of critics for reviews, the number of user votes, Facebook reactions, movie runtime, budget, and gross collection all significantly impact IMDb scores, making it a great source to consider datasets from. According to the findings and predictions received from all the algorithms, Random Forest had the highest rate of prediction accuracy among the evaluated. Compared to earlier studies, the suggested model predicts movie success with greater accuracy, proving it successful. A possible limitation in this study includes limited exploration of social media data like youtube, Facebook, and Twitter comments. Fully relying on IMDb scores may result in ignoring other significant factors like box-office revenue, awards, and critical acclaim. Moreover, the limitation over sample size is compromising the proposed model to be a global one. Still, the researchers also expressed their desire to work on it using other learning models.

W. Lu [8] explores how to collect and evaluate online comments to predict the movie box office. For data acquisition, the author gathered comments on individual movies using the Baidu search engine and pre-processed the data to extract subjective criticism and make it more uniform. For emotional orientation analysis, two kinds of approaches: machine learning(using a supervised training method) and semantic orientation(using an unsupervised approach), are introduced here. Finally, the study proposes a K.N.N. (K-Nearest Neighbors) model-based prediction method for box office, which classifies comments as positive or negative, and calculating their correlation coefficient proves the model's great accuracy. The potential limitations in the study include using a specific search engine and a specific platform for collecting data, primarily focusing on the K.N.N. model regardless of other machine learning techniques being available, and excluding multiple additional factors like marketing campaigns, the popularity of casts, release timing, movie genre, music ratings, etc.

Verma, Garima and Verma, Hemraj. (2019) [9] chose a logistic regression model because of its versatility and capacity to handle binary outcomes. This L.R. model can predict with an accuracy of up to 80% whether a Bollywood film will be a ”Hit” or a ”Flop” before their debut. For data processing and analysis, it uses IBM SPSS 21.0. The authors chose the number of screens, IMDb rating, and MusicRating as predictors, with the movie verdict to be hit or flop as the outcome variable. The data was acquired from online sources like IMDb, bollymoviereviewz.com, planetbollywood.com, boxofficeindia.com, and bollywoodhungama.com through Web Scraping, after extracting data for more than 2000 movies, resulting in the final dataset including 116 movies. The model successfully predicted hit or flop movies with 78.1% accuracy for training data and 80% for test data. The possible shortcomings of this study were that it used limited predictors and was confined to Bollywood movies, which may not give accurate results for movies outside of the Bollywood industry because of Bollywood’s distinctive characteristics, and not taking into consideration multiple additional factors like audience preference as well as opinion, marketing strategies, star power, etc.

M. D. Athira and K. S. Lakshmi [10] emphasize the importance of predicting movie success based on a strategy that improves prediction accuracy by combining movie metadata and reviews data. For review data collection, IMDb is used for the dataset labeling positive, negative, and neutral reviews. In contrast, meta-data is collected from Github that includes information such as the number of critical reviews, duration, number of Facebook reactions, release time, IMDb score, aspect ratio, film budget, gross revenue, etc. An ensemble classifier technique is used in the study for predicting the success rate, with algorithms that involve Random Forest, Naive Bayes, Logistic Regression and Max Voting. The max voting approach will identify the movie as a success, failure, or neutral depending on similar outcomes from two or more methods among these classifiers. Finally, the researchers concluded that combining metadata and review data with the results obtained using the Max Voting technique provides us with an almost 90% accuracy rate, which is better than using any features independently. There are some limitations in this study also, though it can provide us with a much higher accuracy rate. The possible gaps are mainly a lack of versatility in the datasets, difficulty in ensuring quality review data, missing usage of some additional evaluation metrics like recall, F1 score, R.O.C. curve, etc., along with some external factors like marketing strategies, audience preference, film budget, critical acclaim and some more.

N. Darapaneni et al. [11] aimed to forecast the movie box office performance using various Machine Learning algorithms, including K-Nearest Neighbours, Random Forest, XGBoost Classifier, Decision Tree, and Deep Neural Network. The authors implemented these algorithms on an IMDB dataset which focused on features such as movie crew, plot, audience, and critics’ reviews/ratings, to predict their success rate. The study found that the XGBoost Classifier was the most accurate model, achieving an accuracy rate of nearly 90%. It identified user reviews, cast, and budget as key features for predicting the success of movies. A potential gap in the study is that it does not consider other key features, such as the production studio, the film’s director, and the film’s budget, which have historically been known to impact the success of a movie significantly. Examples include movies produced by Disney and movies directed by Christopher Nolan.

Sivakumar, Pirunthavi; Abishankar, Kamalanathan; Rajeswaren, Vithusia Puvaneswaren; Mehendran, Yanusha; Ekanayake, E.M.U.W.J.B. [12] showed how to assess the success and rating of a movie using Random Forest and Naive Bayes models before its release. Using data collected from IMDb, Box Office Mojo, and YouTube comments, this paper highlights various factors that influence the success of a movie. Using YouTube API for scraping 1000 comments per video from YouTube and using a dataset from Kaggle for 200 movies as a substitute for IMDb scraping, they covered the part of data collection. A Random Forest Algorithm is trained using IMDb attributes to predict movie success, while a Naive Bayes model is trained using YouTube user reviews to predict movie ratings. The models perform well, with the Success Prediction model having an overall accuracy of 70%. Still, a problem arises due to YouTube API's tendency to prioritize negative comments over positive ones. Therefore, the study proposes the developed models will work successfully on the data collected from the internet rather than the comments collected from YouTube. The potential gaps in this study comprise the limitation of YouTube's API, which only allows access to a limited number of user comments per video, failing to consider a larger and more diverse dataset required to improve the models' generalizability and being unaware of other factors such as critical acclaim, cultural impact, and long-term profitability that contribute to a film's overall success.

Dewan Muhammad Qaseem, Nashit Ali, Waseem Akram, Aman Ullah, and Kemal Polat (2022) [13] focus on forecasting the success rates of movies by applying a technique to spectators' tweets on movie trailers. The study collected tweets from different movies using the hashtag method. Different machine learning algorithms have been used, such as - Linear S.V.C., K.N.N., Naive Bayes, and decision trees. Also, a lexical-based approach algorithm has been used. The outcomes of this paper showed that Linear S.V.C. has the maximum validity among other machine learning algorithms in predicting movie success by using tweets. Also, the lexical-based approach has used three different dictionaries with three different word counts. Among them, ratings_Warriner dictionary gives a more accurate result. Then the result was compared with other sites, such as IMDb. The potential gap in this study only focuses on tweets. Still, it does not pay attention to the comment section under the youtube trailers, social media comments, and IMDb ratings and comments. Also, it did not consider important key features such as the popularity of actors, actresses, cast, movie budget, quality, promotion, sound effects, etc. In short, this study only used Twitter to predict a movie's success using different machine learning algorithms ignoring some significant features and a lexical-based approach from which Linear S.V.C. and ratings_Warriner dictionary give the highest accuracy.

Chapter 5

Data Description

5.1 Overview

The primary dataset used in this research is an amalgamation of Wikipedia data and YouTube trailer comments that have been collected firsthand by the authors. The dataset initially comprised of 23 features and 2020 rows of data when it was originally collected from Wikipedia, but was left with 18 features and 1780 rows once the data was cleaned. The featues of the final dataset are described in Table ??.

5.2 Wikipedia Data

The raw movie data utilized in this research was web scraped from Wikipedia using Python's beautiful soup library. The scope was limited to major American film productions released theatrically in the United States of America between 2010 and 2022 inclusive. In total, data on 2020 films was collected.

The web scraping methodology produced a dataset that required substantial cleaning and preprocessing before analysis could proceed. Three main data issues were addressed: embedded HTML reference tags, heterogeneous data types for key variables, and non-standardized date variables.

First, the raw Wikipedia data retained various HTML tags used for internal Wikimedia formatting and references. As these provided no analytical value, a find-and-replace script removed all HTML tags across the entire dataset. Second, the running time information for films contained heterogeneous values - some records displayed hours and minutes while others contained only minutes. A simple calculation was scripted to standardize all running times to integer minutes.Similarly, budget and worldwide box office gross values required standardization to a integer data type and domestic currency. Records contained a variety string formats: written numbers, numerals, and currency symbols. Values were parsed to remove non-numeric characters, then converted to integer type, and scaled to hundreds of millions of USD. Finally, movie release dates in the raw data took several forms, such as "June 23, 2017" and "2017-06-23" across records. The Python datetime library was leveraged to convert all release dates into standardized Python datetime objects.

Table 5.1: Final Dataset Before Preprocessing

Feature	**Data Type**	**Description**
Title	string	The titles of each movie in string format.
Directed by	list	A list of strings containing the names of directors of the film.
Produced by	list	A list of strings containing the names of producers of the movie.
Cinematography	list	A list of strings containing the names of cinematographers of the movie.
Runtime	int	The movie's total duration in minutes.
Distributed by	list	A list of strings containing the names of distributors of the movie.
Language	string	The primary language spoken in the movie.
Written by	list	A list of strings containing the names of writers of the movie.
Cast	list	A list of strings containing the names of actors in the movie.
Edited by	list	A list of strings containing the names of editors of the movie.
Production companies	list	A list of strings containing the names of production companies who produced the movie.
Release Date	datetime	The release date of movie in %Y-%m-%d %H:%M:%S format.
MPAA Rating	string	The MPAA rating that the movie received. The values are G, PG, PG-13, R, and NC-17.
Sentiment Score	float	The sentiment score calculated from the YouTube trailer comments of each movie.
Budget	int	The movie's production budget.
Box Office	int	The total box office gross of the movie.
Box Office Status	string	An estimation of whether or not the movie was profitable in the box office.

With the dataset cleaned, an additional column was engineered called "Box Office Status", which is a binary indicator of whether a given film achieved financial success. The informal Hollywood rule-of-thumb states that for a film to break even, it must gross double its production budget. An even stricter criterion was adapted whereby $\frac{2}{3}$rds of a movie's box office gross amounted to greater than its budget multiplied by 1.5 were labeled box office successes, while films failing to meet that benchmark were labeled as failures. This criterion was formulated by film journalist John Campea where the total budget of a movie (production budget + marketing budget) is assumed to 1.5 times its production budget, and the take home box office revenue for production companies is assumed to be $\frac{2}{3}$rds of the total box office revenue. The primary assumptions are that the marketing budget of movies is 50% of the production budget, and that production companies pay $\frac{1}{3}$rd of total box office revenue to theaters, which leaves them with $\frac{2}{3}$rds of the total revenue. These assumptions were necessary because information on marketing budget and take home box office numbers are not available publicly.

$$\text{Box Office Success} = (\frac{2}{3} \times \text{Box Office Revenue}) > (1.5 \times \text{Budget}) \tag{5.1}$$

5.3 YouTube Trailer Comments

Building on the movie dataset from Wikipedia, additional data was collected from YouTube utilizing the YouTube Data API. The objective was to gather viewer commentary and discussion around trailers for the movies contained in the Wikipedia dataset.

The YouTube Data API allows programmatic access to metadata on videos uploaded to the platform. This capability was leveraged to search for and identify the official trailers for each of the 2,020 American films released between 2010-2022 in the dataset. The trailer's video ID was collected using the API's search method to search for the movie trailers using the movie's title. The trailer video IDs were then used to request user comments made on that trailer video. For each movie trailer video, the oldest 100 comments were collected. This was done to try and ensure that the comments were older than the movie's release date in order to avoid biased data. In cases where fewer than 100 comments existed for a given trailer video, all available comments were gathered. Any comments consisting solely of non-alphanumeric characters were filtered out. In total, 160,138 YouTube comments were collected across 2,020 trailers. The discrepancy in numbers arises from the fact that some obscure indie films had very little identifiable trailer comments on YouTube whatsoever. Moreover, some movie trailers had their comments turned off, which prevented users from commenting on those trailers. For films with trailer comments, additional features appended to the dataset include: Trailer ID, Trailer Title and Comment Text.

Sentiment analysis was then performed on these trailer comments and an average sentiment score was derived for each movie. This additional information was then integrated with the core Wikipedia movie features by matching records on movie title. This combined multi-modal dataset of both structured and unstructured data formed the foundation for investigating what signal, if any, YouTube trailer commentary provides in predicting financial performance of movies.

Chapter 6

Data Analysis

6.1 Correlation

Correlation refers to the relationship between two variables. It is a statistical measure of how a change in one variable impacts another variable. In this study we have utilized the correlation coefficient to establish relationships between box office success and other movie features, in order to see how much of an impact each feature has in predicting the financial performance of movies.

Table 6.1 and Figure 6.1 show the correlation coefficient and scatter plot between Budget, Runtime, Release Year, and Box Office respectively. Budget has the highest positive correlation with box office revenue among all the other features with a value of 0.7749. This indicates that movies with higher budgets tend to perform better at the box office. This is also demonstrated in the Budget and Box Office scatter plot in Figure 6.1, which shows a growth in box office revenue with an increase in the budget. There is also a positive correlation between a movie's runtime and its box office performance, though to a lesser degree than budget. Most movies released between 2010 and 2022 have a runtime of around 125 minutes to 150 minutes, as shown by the scatter plot in Figure 6.1. There is also a negligible positive correlation between a movie's release year and its box office revenue. This indicates that box office revenue for movie's have grown over the years, although very little.

Table 6.2 and Figure 6.2 show the correlation coefficient and scatter plot between the Success Score of different features and Box Office respectively. The feature with the highest positive correlation with Box Office Revenue is Production Company with a correlation coefficient of 0.2972. A close second are the producers of the film with a correlation coefficient of 0.2851. This indicates that a film's box office revenue is greatly dependent on the company and producers that produce it. The scatter plot in Figure 6.2 also show an increase in box revenue with an increase in the success score of production companies and producers. The success score of directors and writers also seems to have a great deal of impact on a movie's box office revenue with a positive correlation coefficient of 0.2521 and 0.2367 respectively. Every other feature to a lesser degree also has a positive correlation with box office revenue.

Table 6.3 and Figure 6.3 show the correlation coefficient and scatter plot between the Sentiment Score and Box Office revenue respectively. The value being very small at only -0.0142 indicates that there is very little correlation between box office success and sentiment score. This is unusual as the norm would suggest that audience sentiment plays a key role in the success of movies at the box office. This deviation from the norm may in part be due to the way in which sentiment scores were calculated from YouTube comments in this study.

Table 6.1: Correlation between Budget, Runtime, Release Year and Box Office

	Box Office
Budget	0.7749
Runtime	0.4237
Release year	0.0136

Table 6.2: Correlation between Success Score and Box Office

	Box Office
Cast Sucess Score	0.1895
Production Company Sucess Score	0.2972
Director Sucess Score	0.2521
Distributor Sucess Score	0.1776
Producer Sucess Score	0.2851
Cinematographer Sucess Score	0.1399
Writer Sucess Score	0.2367
Editor Sucess Score	0.1980

Table 6.3: Correlation between Sentiment Score and Box Office

	Box Office
Sentiment Score	-0.01423

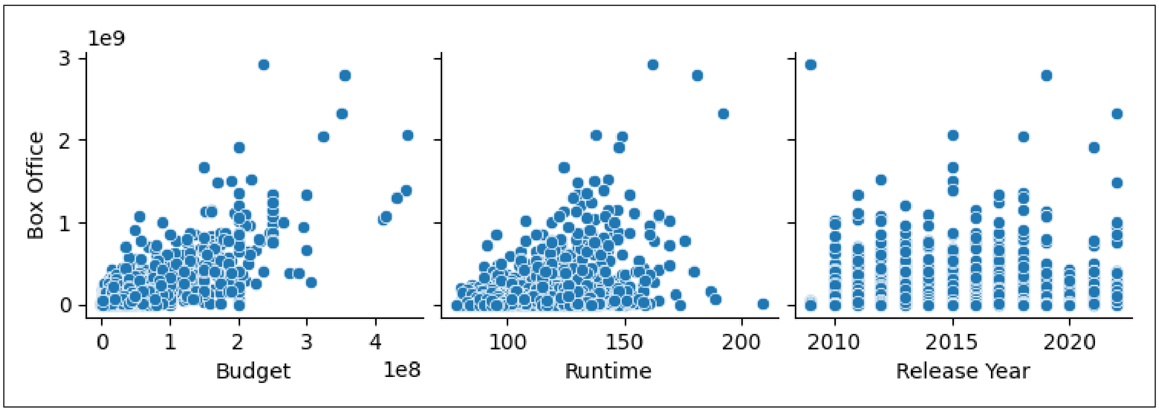


Figure 6.1: Scatter Plot of Budget, Runtime, Release Year and Box Office

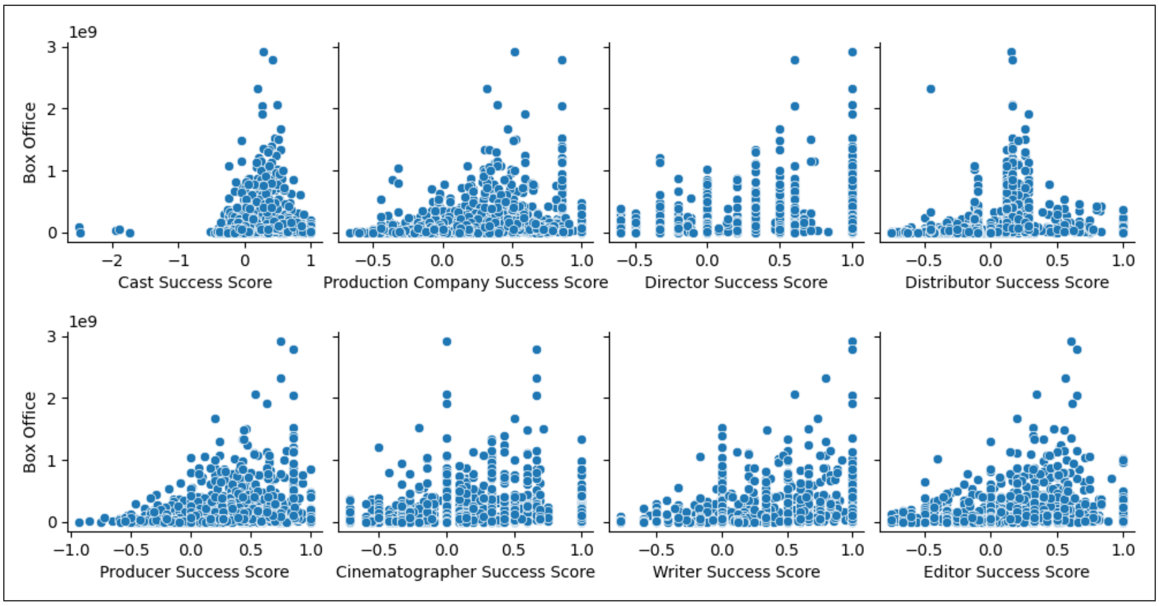


Figure 6.2: Scatter Plot of Success Scores and Box Office

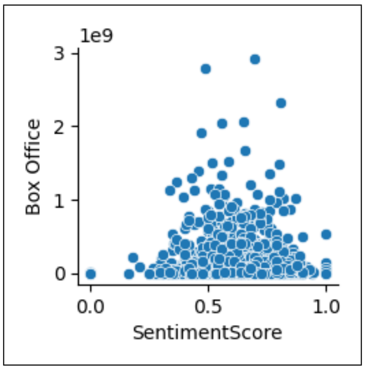


Figure 6.3: Scatter Plot of Success Scores and Box Office

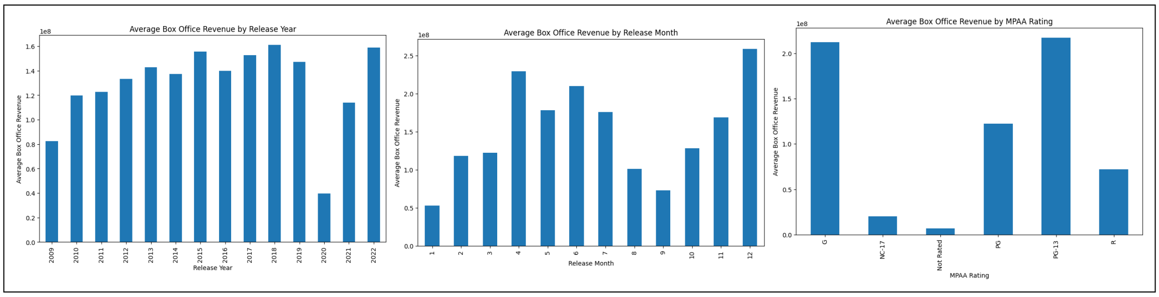


Figure 6.4: Average Box Office Revenue by Release Year, Release Month, and MPAA Rating

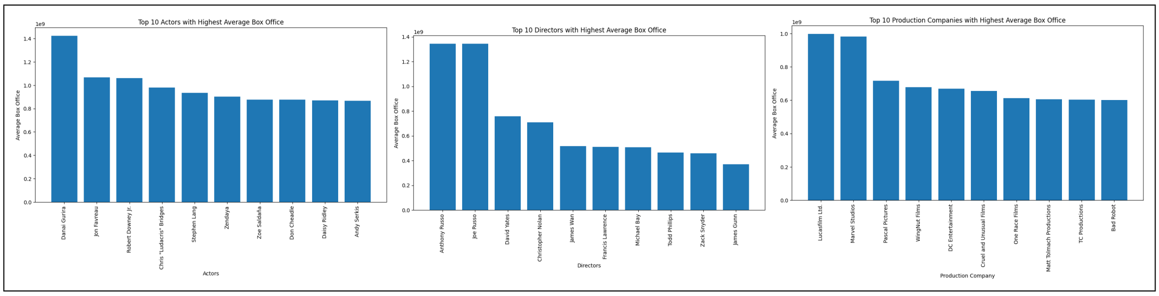


Figure 6.5: Top 10 Actors, Directors, and Production Companies with Highest Average Box Office Revenue

6.2 Average Box Office Revenue

Figure 6.4 shows the average movie box office revenue for different years, months, and MPAA ratings. It can be observed that 2018 and 2022 had the highest average box office revenues, for movies released between the years 2009 and 2022, with an average value of $160 million. The year 2020 saw a massive fall in average box office revenue, which can be attributed to less people going to movie theaters due to the restrictions put in place during the 2020 COVID-19 pandemic. For release months, the month of December seems to be the time when most people visit the movie theaters, as indicated by an average monthly box office value of over $250 million. The months of April and June also show respectable box office numbers with both having a monthly average box office revenue of over $200 million. On the other hand, the month of January seems to be the least popular time for people to see movies, with a average monthly box office of only $50 million. Movies with an MPAA rating of G and PG-13 have performed the best within the last decade. This is not surprising as most franchise films have PG-13 ratings and G films are usually family films watched by many families during Summer and Christmas holidays. Likewise, films with NC-17 and R ratings performed less well because of the niche nature of their audiences. Films that have no rating performed the worst.

Figure 6.5 shows the top 10 actors, directors, and production Companies with the highest average box office revenue. Danai Gurira, Jon Favreau, and Robert Downey Jr are the top 3 highest average box office grossing actors. Anthony and Joe Russo, along with David Yates are the top 3 highest average grossing directors. And the top 3 highest average grossing production studios are Lucasfilm Ltd, Marvel Studios, and Pascal Pictures.

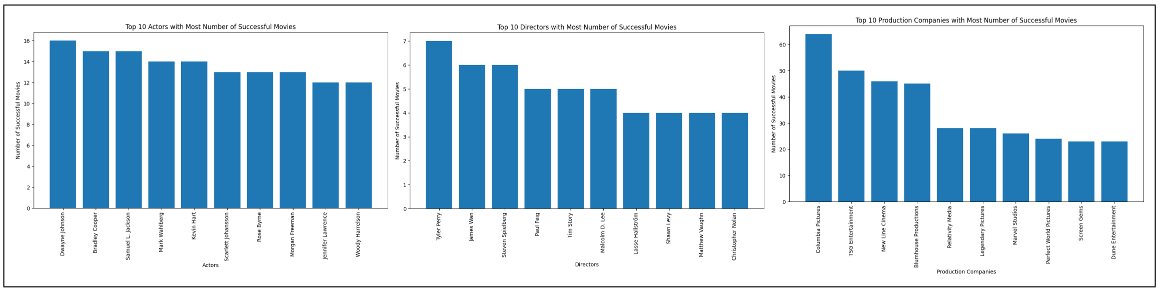


Figure 6.6: Top 10 Actors, Directors, and Production Companies with Most Number of Successful Movies

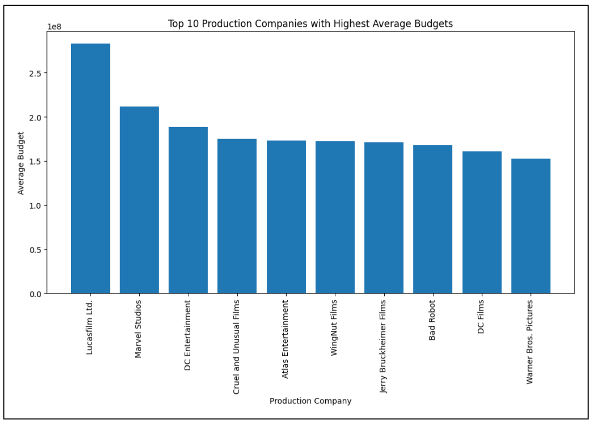


Figure 6.7: Top 10 Production Companies with Highest Average Budgets

6.3 Most Number of Successful Movies

Figure 6.6 shows the top 10 actors, directors, and production Companies with the most number of successful movies. Dwayne Johnson, Bradley Cooper, and Samuel L Jackson are the top 3 actors with most number of successful movies with 16, 15, and 15 successful movies respectively. Tyler Perry, James Wan and Steven Spielberg, are the top 3 directors with the most number of successful movies. And the top 3 production studios with the most successful movies are Columbia Pictures, TSG Entertainment, New Line Cinema.

6.4 Highest Average Budgets

Figure 6.7 depicts the top 10 production companies with highest average budgets. Lucasfilm Ltd, Marvel Studios, and DC Entertainment are the top 3 production companies with highest average budgets. This can be attributed to the fact all three studios produce large scale blockbuster franchises such as Star Wars, Marvel, and DC, which have large crews and huge production budgets.

6.5 Most Number of Movies Released

Figure 6.8 shows the number of movies released in different years, months, and for different MPAA ratings. It can be observed that 2011 and 2010 had seen the most numbers of movie released, between the years 2009 and 2022, with more than 150

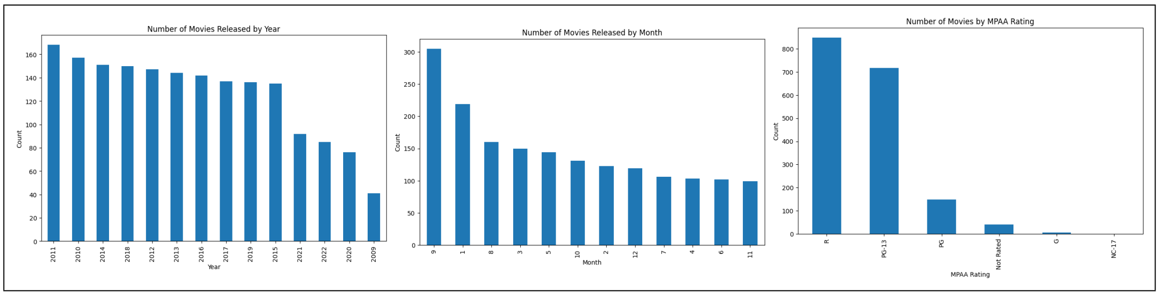


Figure 6.8: Top 10 Production Companies with Highest Average Budgets

major films being released. The year 2020 saw a massive fall in the number of movies released, which can be attributed to movie theaters being closed and film production coming to a halt due to the restrictions put in place during the 2020 COVID-19 pandemic. For release months, the month of September seems to be the time when most movies are released in movie theaters, as indicated by a release count of over 300 movies, between the years 2009 and 2022 . In terms of MPAA ratings, the vast majority of films released in the last decade have been rated either R or PG-13.

Chapter 7

Data Preprocessing

Before any machine learning algorithms could be implemented on the data, it needed to be preprocessed in order to remove irrelevant features, address null values, and to make it suitable for machine learning implementation. This was done in four steps. Firstly, the data was cleaned in order to remove irrelevant features and address null values. Sentiment analysis was then performed on the YouTube trailer comments. The success scores for cast, director, producers, etc was calculated. And finally, the MPAA rating column was encoded using both Label Encoding and One Hot Encoding.

7.1 Data Cleaning

During the data cleaning process, all irrelevant features were removed from the dataset. These features included: **Country**, **Story by**, **Music by**, **Based on**, **Trailer ID**, **Trailer Title**, **Language**, and **Title**. **Trailer ID** and **Trailer Title** were removed as they were only used to collect YouTube comments data using the YouTube Data API. **Country** and **Language** were removed because 99% of the movies in the dataset were produced in the United States and 99% of the languages spoken in the movies were English, which gave no significant diversity to the features. Part of the data cleaning process was also to address null values within the features and samples. All null were filled in manually by the authors by looking up the missing information corresponding to the movie's title on IMDB. For features such as **Cast**, **Producers**, **Directors**, etc., the features were converted to lists of strings from strings. This was done to make sure that the names of people or companies within those lists can be evaluated individually. The **Release Date** feature was split into three distinct **Release Year**, **Release Month**, and **Release Day** features.

7.2 Sentiment Analysis

YouTube Data API

7.3 Success Score Calculation

The success score is a measure that was formulated by the authors of this study to calculate the likelihood of an individual actor, director, producer, etc.'s movie being

Table 7.1: Final Dataset After Preprocessing

Feature	**Data Type**
Cast Success Score	float
Director Success Score	float
Producer Success Score	float
Cinematographer Success Score	float
Runtime	int
Distributor Success Score	float
Writer Success Score	float
Editor Success Score	float
Production Company Success Score	float
Release Day	int
Release Month	int
Release Year	int
Age Rating	int
SentimentScore	float
Budget	float
Box Office Status	float

successful. Each individual was assigned a success score which was calculated by subtracting the number box office failures that they appeared in from the number of box office successes that they appeared in, divided by the total number of movies that they appeared in. This is demonstrated in Equation 7.1. After the individual success score were calculated, the scores were assigned to all groups of individuals involved in each movie. The success score for the Cast feature was assigned with weights, with the lead actor having the most weight. The lead actor contributed to 20% of the total cast score, while the other actors jointly contributed to the other 80%. This was done because not all roles contribute to the movie's success equally, with lead actors traditionally contributing the most. A group success score was then calculated by summing up the success score of each individual in the group, and dividing that sum by the number of individuals in the group. This is demonstrated in Equation 7.2.

$$\text{Individual Success Score} = \frac{\text{No. of Successful Movies} - \text{No. of Failed Movies}}{\text{Total No. of Movies}} \tag{7.1}$$

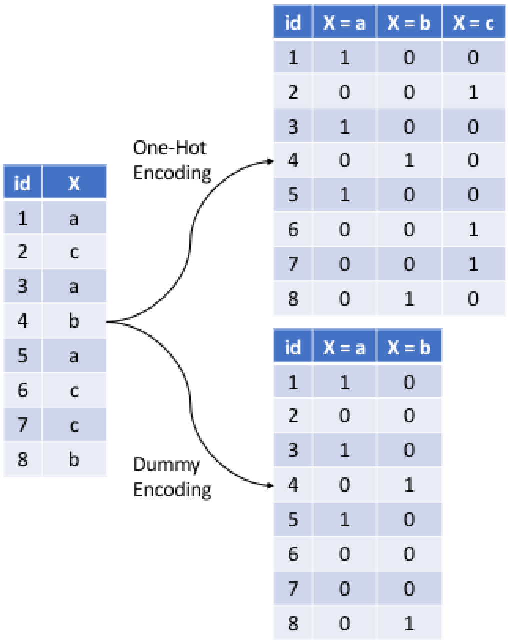
$$\text{Group Success Score} = \frac{\sum \text{Individual Success Scores}}{\text{No. of Individuals in the Group}} \tag{7.2}$$

7.4 Data Encoding

The **MPAA Rating** feature was composed of categorical values. To make it suitable for machine learning implementation, encoding techniques were used to convert the categorical data into numerical figures on which machine learning algorithms can be trained on. Label Encoding and One Hot Encoding are both techniques which help to convert categorical data into numerical data. Figure 7.1 illustrates the difference between the two encoding techniques. Both techniques were used in order to evaluate which one gave better results. The target column **Box Office Status** was also converted into numerical values through binary encoding, with *Success* being 1, and *Failure* being 0.

7.5 Data Splitting

Initially, the dataset was split into features and target variable. The features were then normalized using a MinMaxScaler which transforms the features by scaling each feature to a given range. The normalized values and the target variable was then split into three parts for training, validating, and testing, with a random state of 42. This was done in the ratio 0.7 : 0.06 : 0.24, which in numerical terms was 1232 : 106 : 423 respectively. The training set of 1,232 movies was used to train the five machine learning algorithms. The validation set of 106 films was used for hyper-parameter tuning. And, the test set of 423 movies as used to test the prediction capabilities of the algorithms.



id	X
1	a
2	c
3	a
4	b
5	a
6	c
7	c
8	b

id	X = a	X = b	X = c
1	1	0	0
2	0	0	1
3	1	0	0
4	0	1	0
5	1	0	0
6	0	0	1
7	0	0	1
8	0	1	0

id	X = a	X = b
1	1	0
2	0	0
3	1	0
4	0	1
5	1	0
6	0	0
7	0	0
8	0	1

Figure 7.1: Label Encoding vs. One Hot Encoding

Chapter 8

Methodology

8.1 Model Description

8.1.1 Sentiment Analysis Model

8.1.2 Random Forest Classifier

Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to make more accurate predictions by reducing overfitting and increasing model stability. We chose this model it can handle high-dimensional data like image pixels, capture complex patterns, and reduce overfitting by combining multiple decision trees, making it suitable for image classification tasks.

8.1.3 Decision Tree Classifier

Decision Tree is a supervised machine learning algorithm that represents a flowchart-like structure to make decisions by splitting data into branches based on feature values, helping in classification and regression tasks. We chose this algorithm because it can learn hierarchical image features and make decisions based on pixel values, which can be useful for image classification.

8.1.4 K Nearest Neighbors Classifier

K Nearest Neighbors is a supervised machine learning algorithm used for classification and regression, where an object is classified or predicted based on the majority class or average of its k nearest neighboring data points in the feature space. KNN can be used for image classification by considering pixel values as features and comparing them to neighbors, making it robust in capturing local patterns and textures in images.

8.1.5 Logistic Regression

8.1.6 Support Vector Classifier

Support Vector Classifier is a supervised machine learning algorithm used for classification and regression tasks, aiming to find the optimal hyperplane that maximizes the margin between different classes in the feature space. SVMs are powerful for image classification because they can find optimal decision boundaries in high-dimensional spaces, effectively separating different image classes while maximizing the margin, which enhances their generalization capability.

8.2 Model Training

8.2.1 Sentiment Analysis Model

8.2.2 Random Forest Classifier

The Random Forest was trained using 50 100 and 200 tress. All were tested on the validation set to identify the optimum number of tress. The results are compared in Table 8.1.

Table 8.1: Results from Different Number of Trees for Random Forest On Validation Set

No. Of Trees	**Accuracy**	**Precision**	**Recall**	**F1 Score**
50 Trees	0.959	0.959	0.959	0.959
100 Trees	0.960	0.960	0.960	0.960
200 Trees	0.960	0.960	0.960	0.960

The results were all very similar and highly accurate. It can be observed that increasing the number of tress over 100 has little to no impact on accuracy. We chose 100 trees for training our Random Forest model as it gives the best balance between accuracy and training time.

8.2.3 Decision Tree Classifier

The Decision Tree was trained using 2 criterions.

- **Gini Impurity:** A measure of the impurity or disorder in a set of data. Used in decision tree algorithms to evaluate the quality of a split in a dataset.

$$\text{Gini(D)} = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{8.1}$$

- **Entropy:** A measure of disorder or uncertainty in a dataset. Used in decision tree algorithms to assess the information gain from splitting a dataset based on a particular attribute.

$$\text{Entropy(D)} = -\sum_{i=1}^{c} p_i \log_2(p_i) \tag{8.2}$$

Both criterions were tested on the validation set to identify the best criterion. The results are compared in Table 8.2.

Table 8.2: Results from Different Criterions for Decision Tree On Validation Set

Criterion	Accuracy	Precision	Recall	F1 Score
Gini Impurity	0.947	0.947	0.947	0.947
Entropy	0.943	0.943	0.943	0.943

Both criterions were highly accurate at making predictions on the validation set. Since Gini Impurity gave more accurate results we chose Gini Impurity as the criterion for training our Decision Tree model.

8.2.4 K Nearest Neighbors Classifier

The KNN was trained using 1 2 5 10 15 and 20 neighbors. All number of neigbors were tested on the validation set to identify the optimum number of neighbors. The results are compared in Table 8.3.

Table 8.3: Results from Different Number of Neighbors for KNN On Validation Set

No. Of Neighbors	Accuracy	Precision	Recall	F1 Score
1 Neighbor	0.951	0.950	0.951	0.950
2 Neighbors	0.909	0.918	0.909	0.911
5 Neighbors	0.823	0.865	0.823	0.831
10 Neighbors	0.724	0.814	0.724	0.743
15 Neighbors	0.679	0.787	0.679	0.703
20 Neighbors	0.645	0.776	0.645	0.676

The results for each number of neighbors were very dissimilar to each other. It can be observed that increasing the number of neighbors has a negative effect on model accuracy. The accuracy fell as the number of neighbors increased. We chose 1 neighbor for our KNN model as it gave the most accurate results among all.

8.2.5 Logistic Regression

8.2.6 Support Vector Classifier

The SVM was trained using a linear and polynomial kernel. The kernels were tested on the validation set to identify the optimum number of neighbors. The results are compared in Table 8.4.

Table 8.4: Results from Different Kernels for SVM On Validation Set

Kernel	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Linear	0.955	0.956	0.955	0.956
Polynomial	0.959	0.959	0.959	0.959

The results were all somewhat similar and highly accurate. Since a Polynomial Kernel gave more accurate results we chose a Polynomial Kernel as the kernel for training our SVM model.

8.3 Performance Metrics

The results of our machine learning models were analyzed using 5 performance metrics and visualized using a Confusion Matrix and ROC Curve.

- **Accuracy:** A measurement that determines the percentage of predicted instances, including both true positives and true negatives out of the total instances, in a dataset. It is commonly used to evaluate classification models.

- **Precision:** Assesses the accuracy of predictions by measuring the proportion of positive predictions, among all the positive predictions made by the model. This metric is particularly important when minimizing errors is crucial.

- **Recall:** Sensitivity or Recall measures how accurately the model identifies all instances. It calculates the ratio of predictions, to all actual positive instances. Recall is crucial when prioritizing the avoidance of missing any cases.

- **F1 Score:** This combines precision (the accuracy of predictions) and recall (the ability to find all positive instances) into a single score. It's particularly valuable when working with imbalanced datasets or aiming for a balance between precision and recall.

- **ROC AUC Score:** A representation of the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

- **Confusion Matrix:** A representation used to evaluate the performance of a classification model. It provides counts, for true positives, true negatives, false positives and false negatives allowing assessment of the models accuracy and error rates.

- **ROC Curve:** A representation of the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

Chapter 9

Results

All models were trained with the most suitable hyper-parameters that were determined using the validation test set. All models performed well in predicting test set results with an average accuracy of over 90%.

9.1 Random Forest Classifier Results

Table 9.1: Random Forest Classifier Accuracy Measures On Test Data

Model	F1 Score	Accuracy	Precision	Recall	ROC AUC Score
Random Forest	0.9598	0.9598	0.9598	0.9598	0.9598

As shown in Table 9.1 the Random Forest model performed well on the test set. The model was able to accurately classify 95.98% of movies, suggesting that the model is effective at making correct predictions. The model achieved a precision of also 95.95%, which indicates that the the model doesn't often make false positive errors, which is crucial in film industry in order to avoid major loses. The model achieved a recall of 94.3%, indicating that it correctly identified 95.98% of all actual positive cases. And an F1 Score is also 95.98%, suggests that the model strikes a good balance between making accurate positive predictions (precision) and identifying all positive cases (recall).

9.2 Decision Tree Classifier Results

Table 9.2: Decision Tree Classifier Accuracy Measures On Test Data

Model	F1 Score	Accuracy	Precision	Recall	ROC AUC Score
Decision Tree	0.9409	0.9409	0.9411	0.9409	0.9406

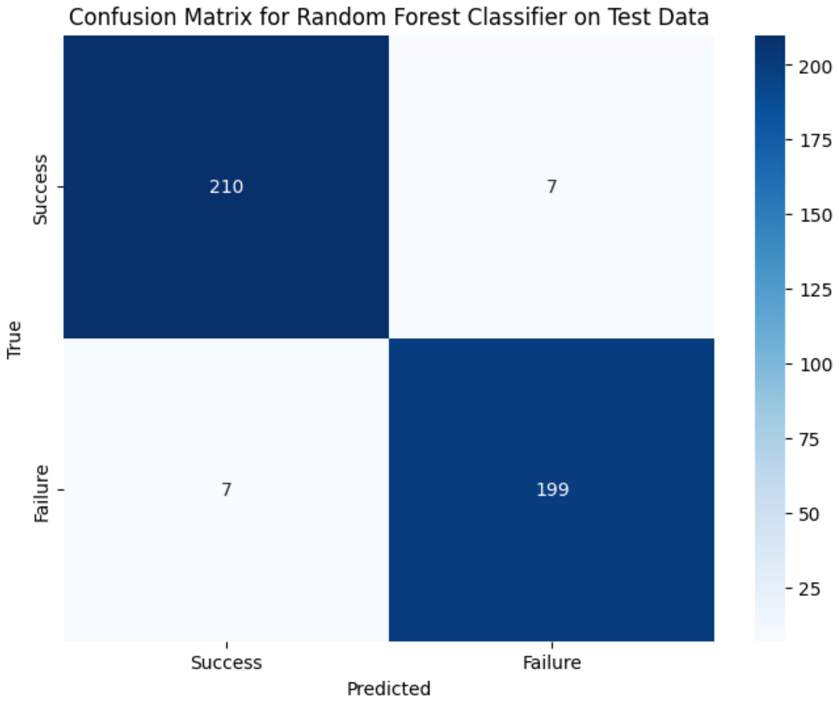


Figure 9.1: Confusion Matrix for Random Forest Classifier On Test Data

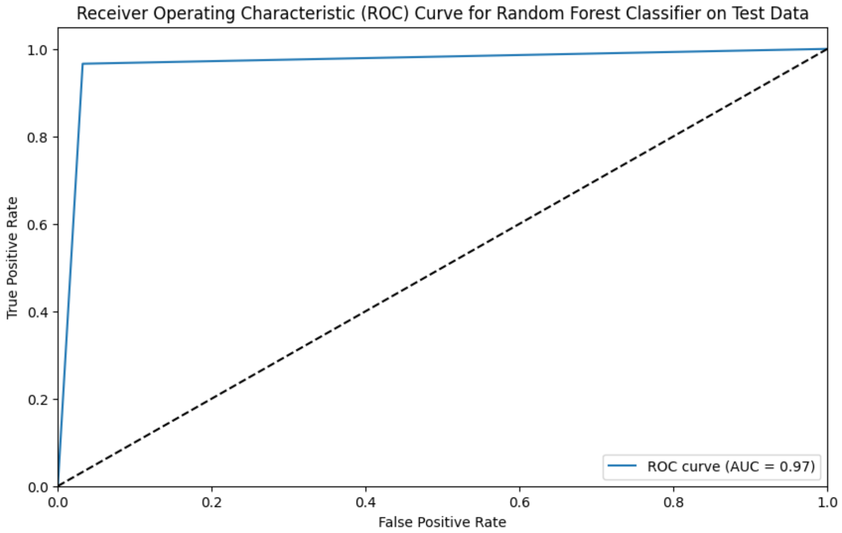


Figure 9.2: ROC Curve for Random Forest Classifier On Test Data

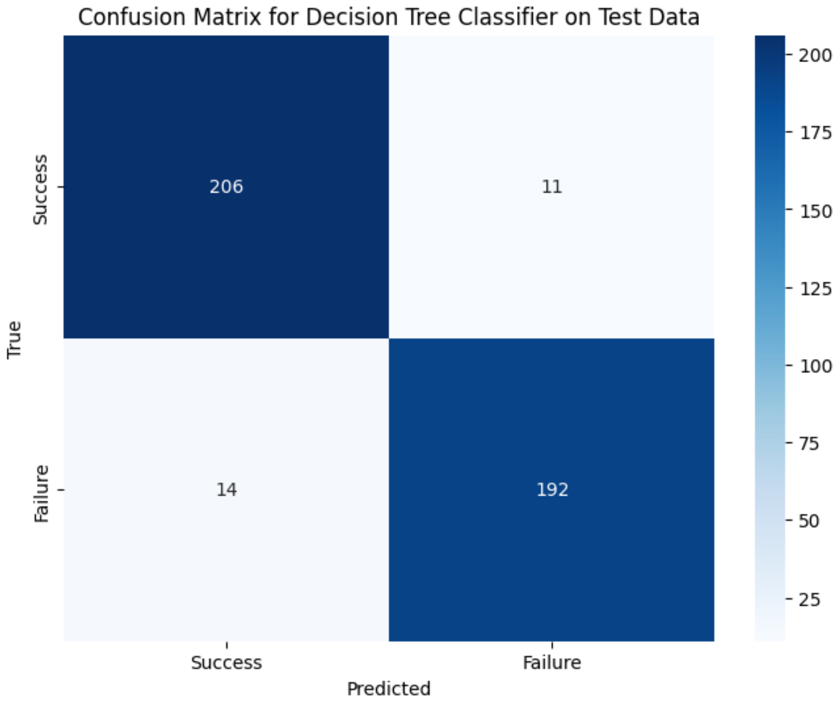


Figure 9.3: Confusion Matrix for Decision Tree Classifier On Test Data

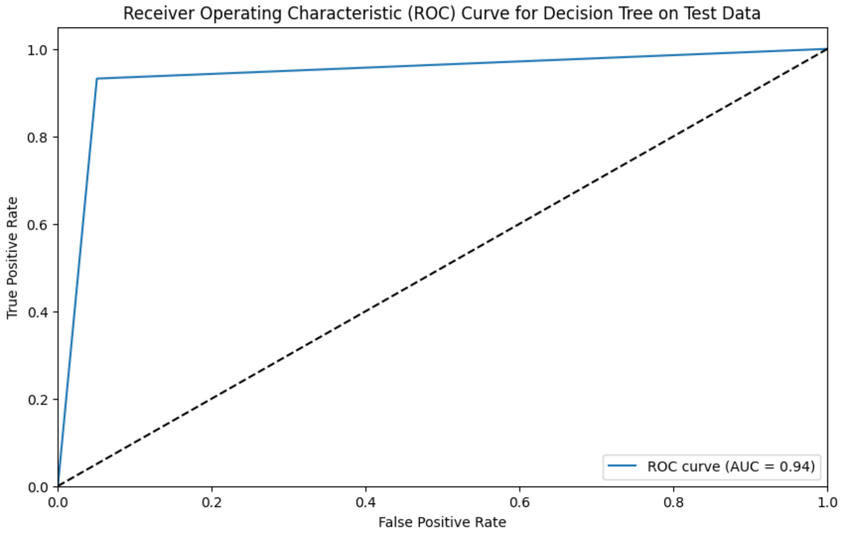


Figure 9.4: ROC Curve for Decision Tree Classifier On Test Data

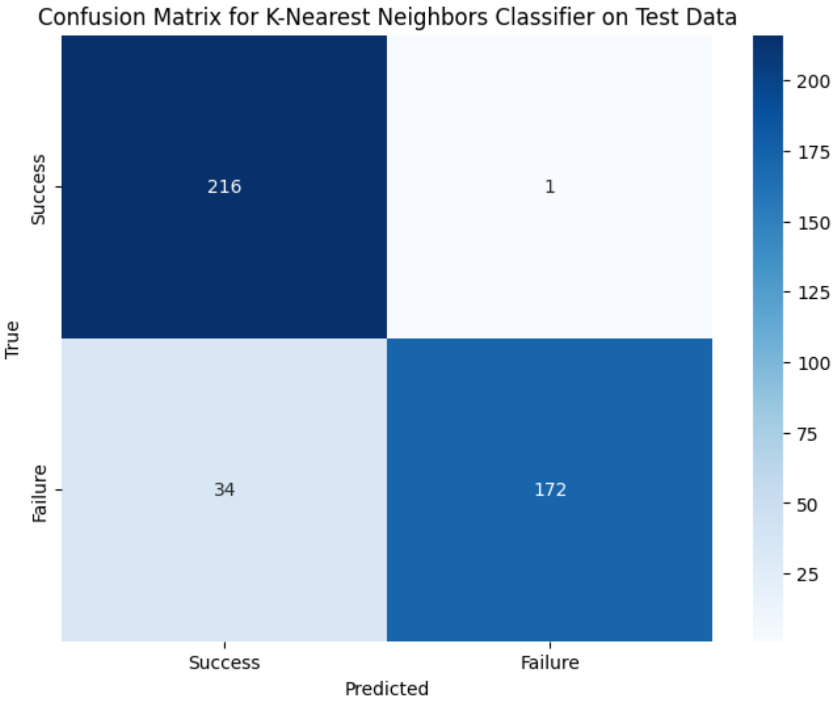


Figure 9.5: Confusion Matrix for K Nearest Neighbors Classifier On Test Data

As shown in Table 9.2 the Decision Tree model performed well on the test set. The model was able to accurately classify 94.09% of movies, suggesting that the model is effective at making correct predictions. The model achieved a precision of also 94.11%, which indicates that the the model doesn't often make false positive errors, which is crucial in film industry in order to avoid major loses. The model achieved a recall of 94.09%, indicating that it correctly identified 95.98% of all actual positive cases. And an F1 Score is also 94.09%, suggests that the model strikes a good balance between making accurate positive predictions (precision) and identifying all positive cases (recall).

9.3 K Nearest Neighbors Classifier Results

Table 9.3: K Nearest Neighbors Classifier Accuracy Measures On Test Data

Model	F1 Score	Accuracy	Precision	Recall	ROC AUC Score
K Nearest Neighbors	0.9166	0.9173	0.9274	0.9173	0.9152

As shown in Table 9.3 the K Nearest Neighbors model performed well on the test set. The model was able to accurately classify 91.73% of movies, suggesting that the model is effective at making correct predictions. The model achieved a precision of also 92.74%, which indicates that the the model doesn't often make false positive errors, which is crucial in film industry in order to avoid major loses. The model achieved a recall of 91.73%, indicating that it correctly identified 95.98% of all

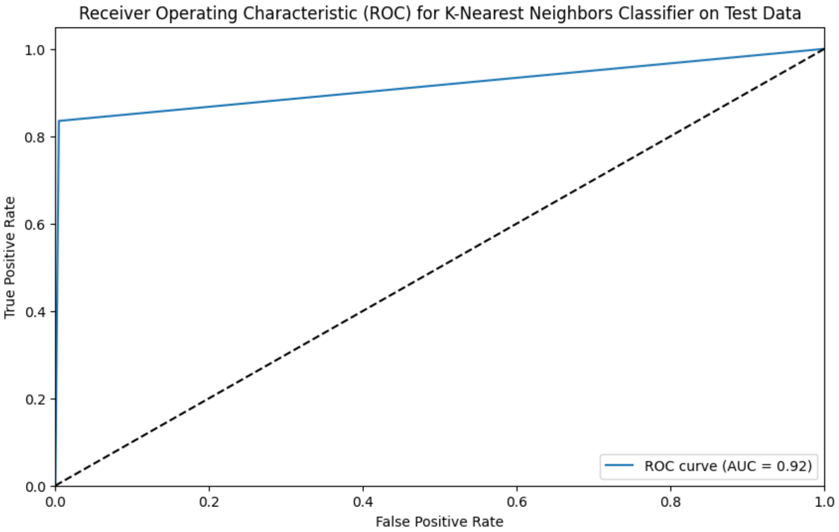


Figure 9.6: ROC Curve for K Nearest Neighbors Classifier On Test Data

actual positive cases. And an F1 Score is also 91.66%, suggests that the model strikes a good balance between making accurate positive predictions (precision) and identifying all positive cases (recall).

9.4 Logistic Regression Results

Table 9.4: Logistic Regression Accuracy Measures On Test Data

Model	F1 Score	Accuracy	Precision	Recall	ROC AUC Score
Logistic Regression	0.9383	0.9385	0.9416	0.9385	0.9374

As shown in Table 9.4 the Logistic Regression model performed well on the test set. The model was able to accurately classify 93.85% of movies, suggesting that the model is effective at making correct predictions. The model achieved a precision of also 94.16%, which indicates that the the model doesn't often make false positive errors, which is crucial in film industry in order to avoid major loses. The model achieved a recall of 93.85%, indicating that it correctly identified 95.98% of all actual positive cases. And an F1 Score is also 93.83%, suggests that the model strikes a good balance between making accurate positive predictions (precision) and identifying all positive cases (recall).

9.5 Support Vector Classifier Results

As shown in Table 9.5 the Support Vector Classifier model performed well on the test set. The model was able to accurately classify 95.74% of movies, suggesting that the model is effective at making correct predictions. The model achieved a

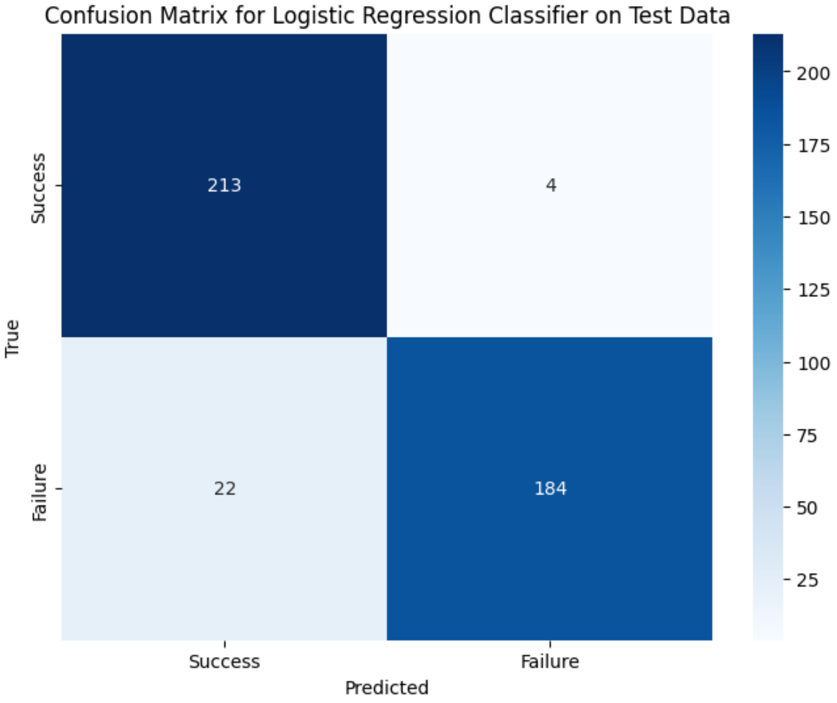


Figure 9.7: Confusion Matrix for Logistic Regression On Test Data

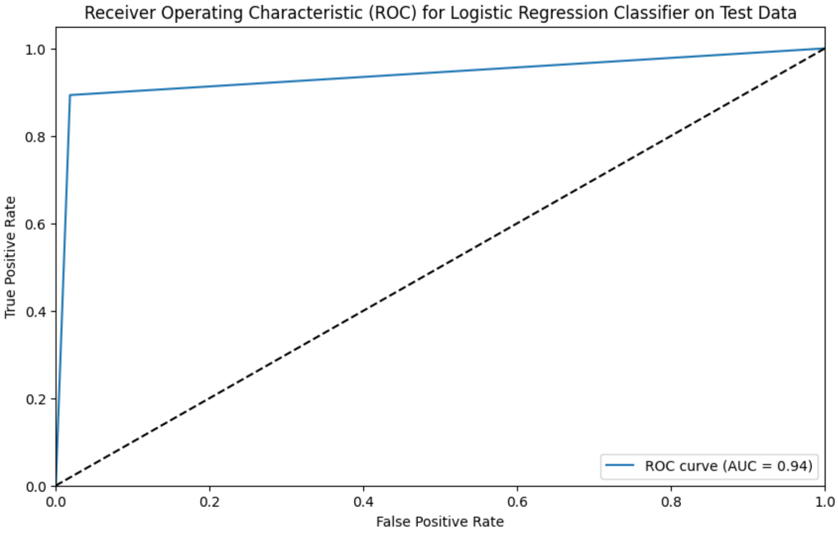


Figure 9.8: ROC Curve for Logistic Regression On Test Data

Table 9.5: Support Vector Classifier Accuracy Measures On Test Data

Model	**F1 Score**	**Accuracy**	**Precision**	**Recall**	**ROC AUC Score**
Support Vector Classifier	0.9574	0.9574	0.9594	0.9574	0.9566

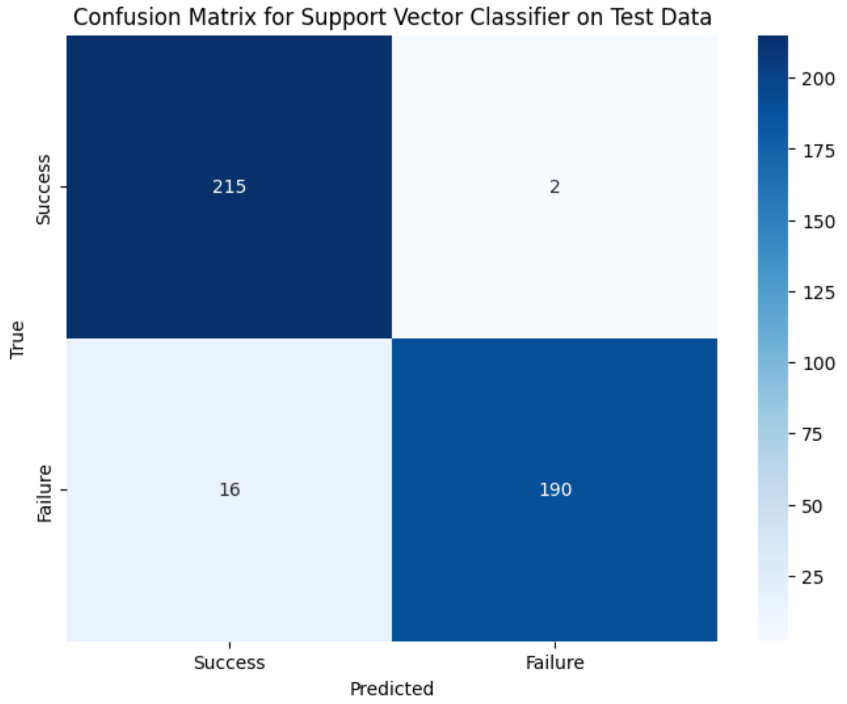


Figure 9.9: Confusion Matrix for Support Vector Classifier On Test Data

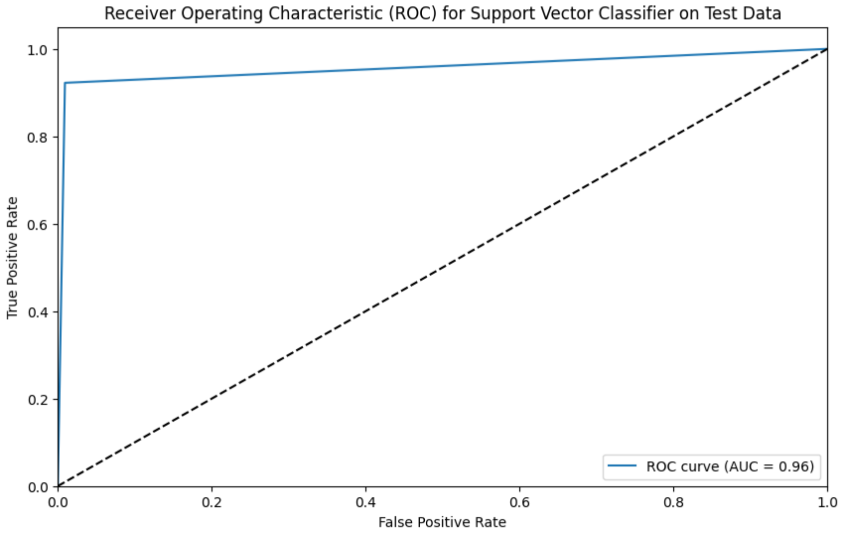


Figure 9.10: ROC Curve for Support Vector Classifier On Test Data

precision of also 95.94%, which indicates that the the model doesn't often make false positive errors, which is crucial in film industry in order to avoid major loses. The model achieved a recall of 95.74%, indicating that it correctly identified 95.98% of all actual positive cases. And an F1 Score is also 95.74%, suggests that the model strikes a good balance between making accurate positive predictions (precision) and identifying all positive cases (recall).

Chapter 10

Conclusion

In conclusion, this study aims to contribute to movie box office prediction by integrating YouTube trailer comments and Wikipedia data. Our approach offers an enhanced understanding of audience sentiments and expectations, enabling us to predict movie box office success accurately. The research outcomes provide valuable insights into the film industry, assisting stakeholders in making informed decisions to maximize box office performance and overall success. However, market trends and unforeseen events can significantly influence predictions' accuracy. Future research can build upon these findings to make further advancements in this field.

10.1 Limitations

While continuing the process of our analysis, we noted multiple limitations regarding the data collection process and strategies. Due to these limitations, the ultimate findings from this analysis may be impacted compromising their stability and robustness.

The limitations that we could identify are as follows:

- **Compact Dataset:** The model's potential for scaling to a wider spectrum of movies may be limited due to assembling only 1780 samples. Such a small sample size can impose limits on the model's ability to provide accurate prediction while tested on fresh unseen data resulting in overfitting and mediocre outcomes. Data augmentation techniques along with opting for transfer learning using pre-trained models might help us deal with this issue.

- **Credibility of Wikipedia Data:** Since Wikipedia is an open-source platform, there is a huge risk of having inaccurate and outdated data. Solely relying on Wikipedia data imposes a huge risk on the validity of the model's prediction compromising its credibility. Cross-referencing of key information with other reliable sources like official datasets, websites, and reports could be a great solution to this problem.

- **Biased Data:** Having considered only American theatrical releases, the dataset used in the analysis is inflicted with several unfortunate biases. Since the selected dataset may not represent a wide portion of international or other non-theatrical movies, it may limit the generalizability of the prediction analysis. Moreover, movies without sufficient info and trailers were excluded from this research which certainly raises a question of its impartiality towards less prominent films.

- **Confined Sentiment Analysis:** Since only 100 comments were chosen from each movie trailer it may not represent the sentiment of the broader audience. Furthermore, relying only on YouTube comments may not provide a comprehensive overview of public opinion since a huge part of the audience uses social media rather than the YouTube comment section to express themselves. Combining other social media platform comments and reactions could provide a more generic overview while adding an extra dimension to this analysis.

10.2 Future Work

Our goal was to provide a better prediction model to the film industry and with our predictive research, we have been able to conclude a movie's box office performance successfully. Still, there is ample scope for improvement in the future that will make the model capable of ensuring a much more robust, unbiased, and reliable outcome.

The future enhancements that we anticipate include:

Bibliography

[1] K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. Liszka, and F. Gregorio, "Prediction of movies box office performance using social media," Aug. 2013, pp. 1209–1214. DOI: 10.1145/2492517.2500232.

[2] V. Jain, "Prediction of movie success using sentiment analysis of tweets," *The International Journal of Soft Computing and Software Engineering [JSCSE]*, vol. 3, no. 3, 2013, ISSN: 2251-7545.

[3] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," Jul. 2013, pp. 1–5. DOI: 10.1109/ICCCNT.2013.6726818.

[4] M. S. Usha and M. I. Devi, "Analysis of sentiments using unsupervised learning techniques," Feb. 2013, pp. 241–245. DOI: 10.1109/ICICES.2013.6508203.

[5] V. Nithin, M. Pranav, P. Sarathbabu, and A. Lijiya, "Predicting movie success based on imdb data," *International Journal of Data Mining Techniques and Applications*, vol. 3, pp. 365–368, Jun. 2014. DOI: 10.20894/IJBI.105.003.002.004.

[6] A. Bhave, H. Kulkarni, V. Biramane, and P. Kosamkar, "Role of different factors in predicting movie success," Jan. 2015, pp. 1–4. DOI: 10.1109/PERVASIVE.2015.7087152.

[7] R. Dhir and A. Raj, "Movie success prediction using machine learning algorithms and their comparison," Dec. 2018, pp. 385–390. DOI: 10.1109/ICSCCC.2018.8703320.

[8] W. Lu, "Research on prediction of movie box office based on internet comments," Dec. 2019, pp. 11–14. DOI: 10.1109/ISCID.2019.00010.

[9] G. Verma and H. Verma, "Predicting bollywood movies success using machine learning technique," Feb. 2019, pp. 102–105. DOI: 10.1109/AICAI.2019.8701239.

[10] M. D. Athira and K. S. Lakshmi, "Movie success prediction using ensemble classifier," Jan. 2020, pp. 1–5. DOI: 10.1109/ICCCI48352.2020.9104183.

[11] N. Darapaneni, C. Bellarmine, A. R. Paduri, *et al.*, "Movie success prediction using ml," Oct. 2020. DOI: 10.1109/UEMCON51285.2020.9298145.

[12] P. Sivakumar, V. P. Rajeswaren, K. Abishankar, E. Ekanayake, and Y. Mehendran, "Movie success and rating prediction using data mining algorithms," *Journal of Information Systems & Information Technology (JISIT)*, vol. 5, no. 2, pp. 72–80, Feb. 2021, ISSN: 2478-0677. DOI: 10.13140/RG.2.2.18052.86402.

[13] D. M. Qaseem, N. Ali, W. Akram, A. Ullah, and K. Polat, "Movie success-rate prediction system through optimal sentiment analysis," *Journal of the Institute of Electronics and Computer*, vol. 4, pp. 15–33, 2022. DOI: 10.33969/JIEC.2022.41002.