Social Media Trend Analysis to Predict the Success of Products using Deep Learning Technique

by

Farden Ehsan Khan
19101418
Ahmed Mahir Ruhan
19101330
Rifat Shamsuddin
19101336

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
September 2022

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:

Farden

Farden Ehsan Khan
19101418

Ahmed Mahir Ruhan
19101330

Rifat Shamsuddin

Rifat Shamsuddin
19101336

Approval

The thesis report titled "Social Media Trend Analysis to Predict the Success of Products using Deep Learning Technique" submitted by

1. Farden Ehsan Khan (19101418)

2. Ahmed Mahir Ruhan (19101330)

3. Rifat Shamsuddin (19101336)

Of Summer, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on September, 2022.

Examining Committee:

Supervisor:
(Member)

Faisal Bin Ashraf
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
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

In recent times, social media usage has reached such heights that it has become a powerhouse in producing trends, bringing such topics that would have remained outside of popular consciousness. Our goal is to analyze how the success of products such as movies can be affected by people's shared thoughts and reactions about it on social media. From the data extracted from social media comments, we will study the sentiment of people regarding a certain movie. For our research, the work will be based on unreleased movies and predict the outcome after release. Accumulated reviews about a movie will be analyzed to decipher whether the public sentiment is positive or negative towards it and estimate the willingness to buy a specific film. From this we will find the correlation between how positive and negative attention can affect the success of a production.

Keywords: Social Media; Trend; Deep Learning; Trend Analysis; KNN; Text Mining; Random Forest; MLP; RoBERTa; BERT; DistilBERT; Sentiment Analysis; Decision Tree

Dedication

We dedicate our research to our parents who helped us in every step of our journey. They gave us encouragement & inspiration to work towards our goal. They deserve all the credit and more.

Acknowledgement

First and foremost, we express our gratitude to the Almighty Allah(SWT), the most benevolent and merciful, for allowing us and giving us the fortune of completing our thesis. With His blessings, we were able to finish our work in due time with due diligence.

Next, we would like to pay regards to our respected supervisor, Mr. Faisal Bin Ashraf who has provided us guidance and helped us overcome our obstacles. The completion of our research was made possible by his support, advice and counsel.

Last but not the least, we thank Brac University for providing us knowledge and giving us the opportunity to conduct the research.

Table of Contents

List of Figures

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

API	Application Programming Interface
$BERT$	Bidirectional Encoder Representations from Transformers
CBF	Content Based Filtering
CNN	Convolutional Neural Network
DT	Decision Tree
ICF	Item-based Collaborative Filtering
kNN	k-Nearest Neighbour
LDA	Latent Dirichlet Allocation
$LSTM$	Long Short-Term Memory
MLP	Multilayer Perceptron
NLP	Natural Language Processing
$NLTK$	Natural Language Toolkit
NMF	Non-negative Matrix Factorization
RFC	Random Forest Classifier
$RoBERTa$	Robustly Optimized BERT pre-training Approach
SVM	Support Vector Machine
$TF-IDF$	Term Frequency-Inverse Document Frequency
$VADER$	Valence Aware Dictionary and Sentiment Reasoner

Chapter 1

Introduction

1.1 Social Media Trend

Nowadays social media has become a crucial platform for interacting and expressing ideas, opinions and other forms of expressions. Millions of people dive into social media to convey their thoughts and point of view. Trend points to a certain action or expression that has solicited mass engagement at a said time. In today's world, social media and trends go hand in hand. Numerous events, activity and protests get thousands of responses in a short period of time, which results in trends. We see a new trend gaining popularity almost everyday. The current world is very much influenced by social media and its trends. People make choices, decisions and changes from being influenced by these online events. So the social media trends have turned into a prominent source of information to acquire public consensus. According to statistics provided by DATAREPORTAL[23], around 4.55 billion people use social media which equates to almost 57.6% of the global population. Their studies also show that Facebook was the most used social media platform in 2021. Other popular platforms are YouTube, Instagram, TikTok, Twitter etc.

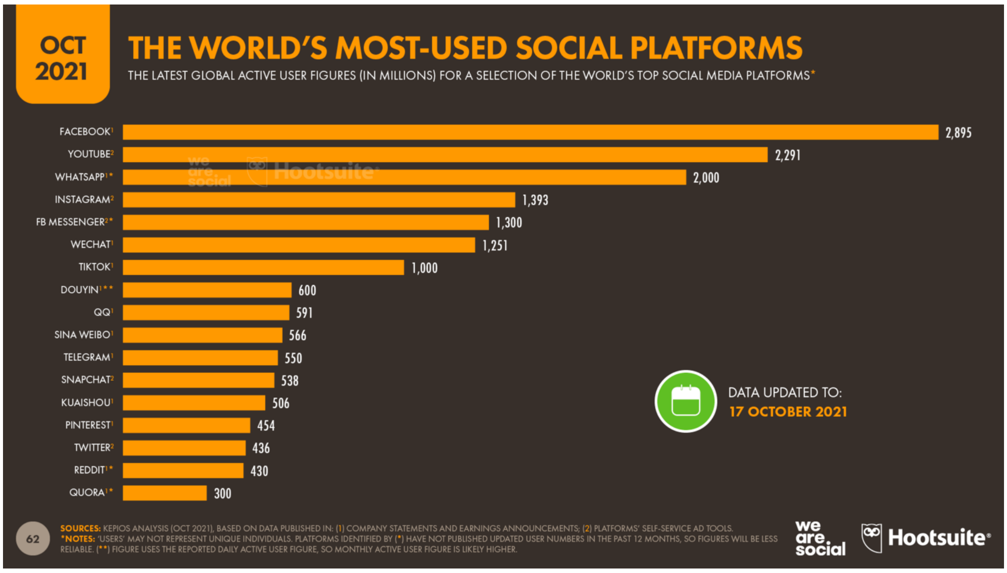


Figure 1.1: Social Media Usage[23]

According to another study done by DATAREPORTAL[19], it seems that online media has surpassed broadcast, print and radio media in terms of circulating news. Social media alone delivers more news than printed media such as newspapers. So considering the amount of social media consumption, it is certain that social media trends play a big part in human life.

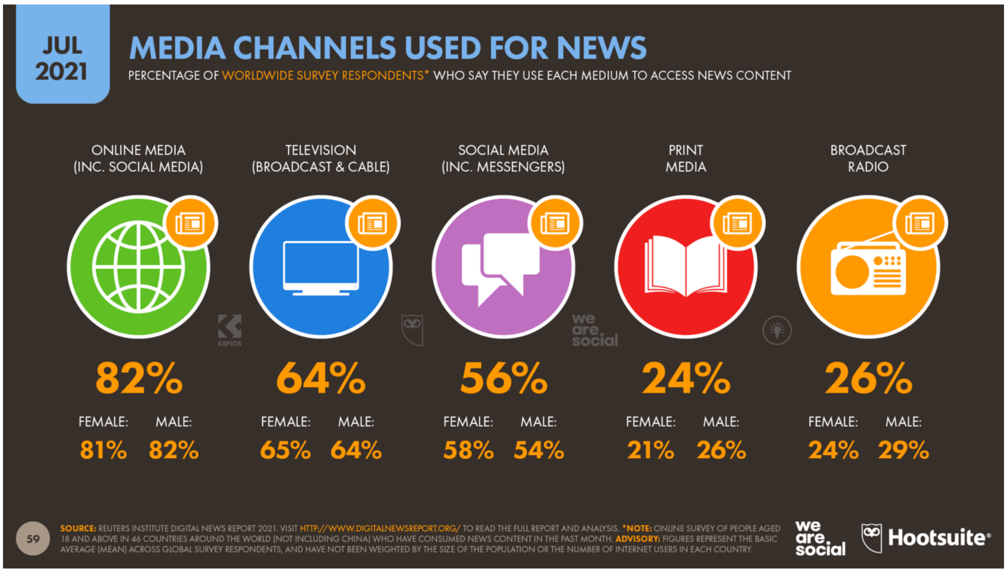


Figure 1.2: News Consumption[19]

With all of this data from social media, if it can be utilized to extrapolate information about certain events or response to such events and/or products. In any case, any topic in particular will help with finding out if a topic is trending or not in real time. We can find out what is trending but what will trend is what we are interested in. If we can study the initial response on a certain movie, we can use their initial response to evaluate the success of the movie itself. In order to do that, we will be using deep learning.

1.2 Deep Learning and Trend Prediction

Trend detection research has been done before and analyzing those trends with sentiment analysis from datasets, correlation between a topic's positive and negative response can be used to predict the longevity of that topic's trending period. To further analyze, this data can be used to build trend prediction mode with the help of deep learning. At first, we need to gather datasets, find trending topics, calculate the positive to negative response ratio and the time that topic was popular. After that, we can feed this data to our deep learning model for training and then pick trends to find and predict the outcome of its longevity. In this way, it can be useful for companies to market products, events. Governments can also benefit from this as they can take proper measures for certain topics.

1.3 Problem Statement

From our research, we would like to find out the association of trend analysis in the movie industry and find its effect on sales and revenue. Research has been done on event detection where after an event is getting traction on social media it can be used to find the public sentiment behind it[6]. Based on similar grounds, we will focus on the social media appeal of a movie and use the opinions as the key element for revenue estimation. Using the point of views generated after release of a trailer and the emotion of peoples' posts on social media about that topic during their popularity period as our training data, we will try to predict if a new movie will be able to gain profit or not. Social media contains huge amounts of information that can be used properly without pre-processing and further classification. We will use comments left on YouTube trailers as our base. It is necessary to remove the extra and irrelevant parts of the comments before we proceed with our research. Although many NLP modules exist to make sense of people's posts and views on social media, using that data to predict what will become a hit online will help with said film's market value. For this we will use BERT and RoBERTa model. On the other hand, if some movies were to not trend, with the data collected it can be studied on which parameters a topic's probability to succeed depends on.

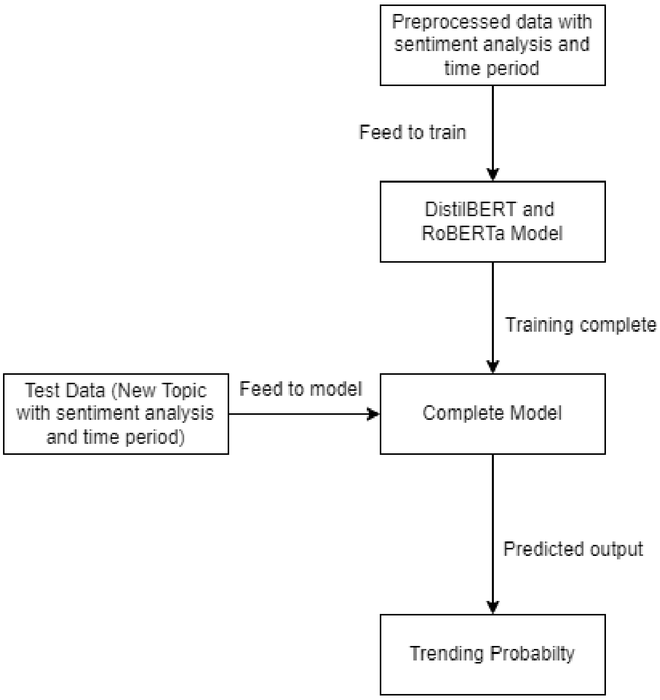

Preprocessed data with sentiment analysis and time period
Feed to train
DistilBERT and RoBERTa Model
Training complete
Test Data (New Topic with sentiment analysis and time period)
Feed to model
Complete Model
Predicted output
Trending Probabilty

Figure 1.3: Proposed Solution

For our research in other papers, topic prediction is done with respect to a user's preference[5]. But we are trying to do this with deep learning. With the help of NLP sentiment analysis, positive and negative comment ratio, we can find quantifiable measures to predict a topic's relation with the social media's user base. Our focus is to build a reliable model that can accurately estimate if the movie will be a success or not at the box office.

1.4 Research Objective

Our objective is to foresee the success rate of a movie based on pre-release reviews. By utilizing metadata and relevant details such as purchase intention mined from the comments, sentiment of the comment and the gross collection would also be estimated. To get started we would need to extract our targeted data from social media (i.e. YouTube) without taking any bias into account. Research data needs to be unbiased to get the full prediction. Moreover, with taking in data we need to use the right tools to conduct sentiment analysis and with matching the sentiment from one movie to another. Hence, from input we can match the sentiment to predict the probability of the movie's outcome. We would need to implement BERT, RoBERTa for sentiment analysis, KNN, RFC, DT algorithm to estimate the approval of the mass. Thus, our model would portray an image of a movie even before its release and help the entertainment industry assess the potential profit or loss.

1.5 Paper Orientation

The first chapter contains necessary information regarding social media trends and deep learning. It also consists of our problem statement and research objectives, which explain our domain of research and end goal. In addition, chapter 2 includes a literature review of previously published papers related to our topic and some of the major findings from the papers in this field. Then comes chapter 3, which explains our dataset and how we have collected it from scraping social media comments. In addition, explaining our data pre-processing steps. Next in chapter 4, the methodology is given, where we have described the models that we have used for sentiment analysis with the parameters that we have used for our prediction algorithm. Along with that the evaluation of the model is discussed. Following is chapter 5, which is an analysis of our acquired results from the model. Chapter 6 consists of the conclusion and future scope of our research. Lastly, a bibliography detailing all the references we have used has been added at the end of the paper.

Chapter 2

Literature Review

Currently, social media like Facebook, YouTube and Twitter is the main platform for people to show their personal opinions and share information. As the biggest platform for expressing opinions, social media has an immense influence on people's perceptions and mold their cognition about the world itself. Although we can gather information regarding interests in movies using social media, it is hard to gauge people's opinions about them without manual review[18]. Using machine learning, we can classify the sentiment but it might not work with other trends and topics. Using deep learning, we can overcome this hurdle. It combines the machine learning algorithms with deep learning architecture which enables the processing of unstructured data.

2.1 Related Work

Although opinion monitoring about social conflicts like Nuclear power has been performed by surveys, that method is comparatively time-consuming, expensive, and inaccurate as many people might not be interested in such a topic. The study[3] showed that opinion mining using social media about Nuclear power plants can give us a complementary approach to gathering user sentiments. From the study, we can see that people in Korea's perception changed according to various nuclear power-related news and incidents. User opinion became more negative after disasters like the Fukushima nuclear accident. Moreover, the Number of tweets also increased after such incidents. The study was conducted by making a sentiment dictionary and classifying the tweets made by users during a certain period.

Similarly, another study[2] showed people's activity on Twitter increases after disasters like Earthquakes. In the study, the frequency of tweets increased along with earthquakes. According to the study, users act like sensors as they are more likely to tweet about a given situation. Using Twitter's search API, the time, location and device can be gathered. People's frequency of tweets also varied due to the availability of the internet and time of the day.

Although making a sentiment dictionary can work effectively in certain cases. We can increase the effectiveness of gauging the sentiment of a post or a tweet using techniques like machine learning and NLP. In a paper[1] about sentiment classification using machine analysis among the three standard machine learning algorithms, Naive Bayes classification, maximum entropy classification, and support vector machine. The support vector machine is the most accurate although the difference among them is not very large. Using these algorithms, different features such as using the unigrams, bigrams, adjectives, parts of speech, and their position and frequency are used to come to this conclusion.

The biggest hurdle of such analysis is the large amount of data it requires to perform precisely. In a paper[16], it shows that classifying tweets and posts by polarity or negative and positive may simplify the sentiment analysis process and help in further decision making. Using Twitter API, it is possible to gather numerous real-time data and process those tweets using NLTK or Pythons Natural Language Toolkit or other data processing language such R.

By using social media, the release of a product can be well monitored. Also, it can give results about how it is performing with its targeted demography. In this paper[6], from Twitter data, after a certain event has occurred it can be traced to find if it was trending or not. By using ontologies, data is first parsed from datasets and formatted then extracted. After that, it uses input from wiki or DBpedia and connects entities from keywords into relationships. This is what ontology is. They used Web Ontology Language (OWL), to make connections with keywords to Wikipedia to make the base that is used for querying. The result they found using nouns, named entities, and hashtags gave better results compared to verbs. Also, hashtag usage is used less but their appearance makes it more relevant to the query (after spam removal). In short, this paper uses Wikipedia to give reliable information and gauge the success of a marketing campaign, analyzing what is trending from Twitter data. Further analyzing this data if we can find the positive and negative side of the trend it will help much better with the analysis.

Here[5] after finding a trending event that event-related articles will be recommended to a user based on the users' preference. To achieve this, first a trend is determined using how many times that topic has been googled in a selected time frame. Next, that topic is then used to match with the user's preferences and give a recommendation accordingly. LDA (Latent Dirichlet Allocation) is used for detecting topics. This recommendation is given by CBF (Content-Based Filtering) and ICF (Item-based Collaborative Filtering). By finding the popularity degree of an article and the trends popularity score on google insight with relation to a user's preference, a trend-based recommendation system can be established.

This paper[20] collected tweets about the coronavirus vaccine and extracted the sentiment from them. Feelings regarding the vaccine were then analyzed and divided into categories. Here the author used selected keywords to gather the tweets about this issue and used NMF (Non-negative Matrix Factorization) for topic detection. This is how related tweets were singled out. Next, they had to do sentiment analysis. Here, polarity classification was done with VADER (Valence Aware Dictionary and Sentiment Reasoner) Python library. Also, emotion classification was achieved by BERT (Bidirectional Encoder Representations from Transformers) and cosine similarity.

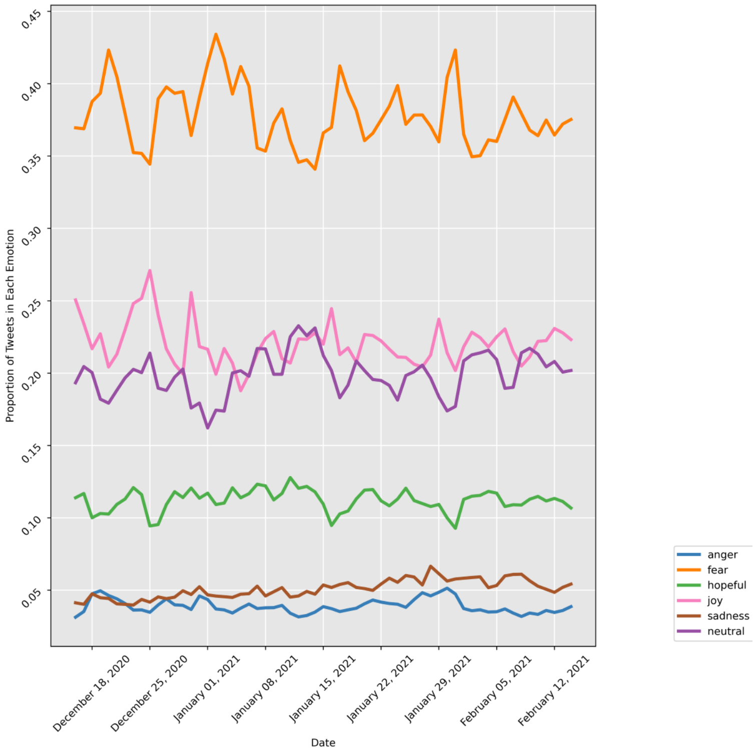


Figure 2.1: Emotion Classification[20]

As a result, they were able to track the emotion from the tweets over a 60-day period. With the help of this research, they were able to detect and find negative emotions about the vaccine ranked highest with fear being the most common. With the help of this analysis government officials can monitor the response of an event and take necessary actions to help mitigate the problems.

A study[7] on social media food trends uses data analysis and image recognition to analyse popularity and trend. Their proposed system works to classify food categories on social media and associate it to a location. Since users don't always mention the names of the food in their tweets, this system relies on image recognition. For data collection, they search based on certain food-related hashtags (e.g. #food,#foodporn, etc) and collect tweets using the streaming API provided by Twitter. From the collection of tweets only tweets with images are further analysed. Then comes Deep CNN which is a pre-trained GoogleNet CNN fine-tuned with ETHZ Food-101 dataset, to categorise the images. Next, food recognition is done by implementing a k-NN classifier which is trained by a union of ETHZ Food-101 dataset and UPMC Food-101 dataset. Their best accuracy rate was 95% and it was obtained by using 15 as the value of k. After categorising and recognising the food items, the system goes on to detect the location of the said item. As the primary source, they use GPS and Place fields in the tweet. If that information is not available, they resort to matching the user location provided in the user profile with the GeoNames dictionary. By detecting the food and location, they analyse the trend and popularity of a certain food in a certain region. Altogether, this study gives a decent way to identify location based food trends from twitter data and images.

In another report[8], trend analysis was carried out on a Mexico based event known as "Justice of the Marquesa". In this ominous event, a bus passenger took on 4 assailants and shot them dead, which generated a wave of tweets on social media. Similar to previous research, this study uses social media as its main source of information. By focusing on the hashtag #justicierodelamarquesa, they collected 10,000 tweets regarding the subject. Then using an algorithm based on NLP and programmed in R, they downloaded tweets and removed all the unnecessary information i.e. the hashtag, author's name and timestamp. Moreover, the extracted data was processed using text mining, which discovers patterns from textual dataset. Based on the frequency of words, a word cloud is formed. The generated cloud is divided into 3 layers(figure 2.2): the core, the middle and the border.

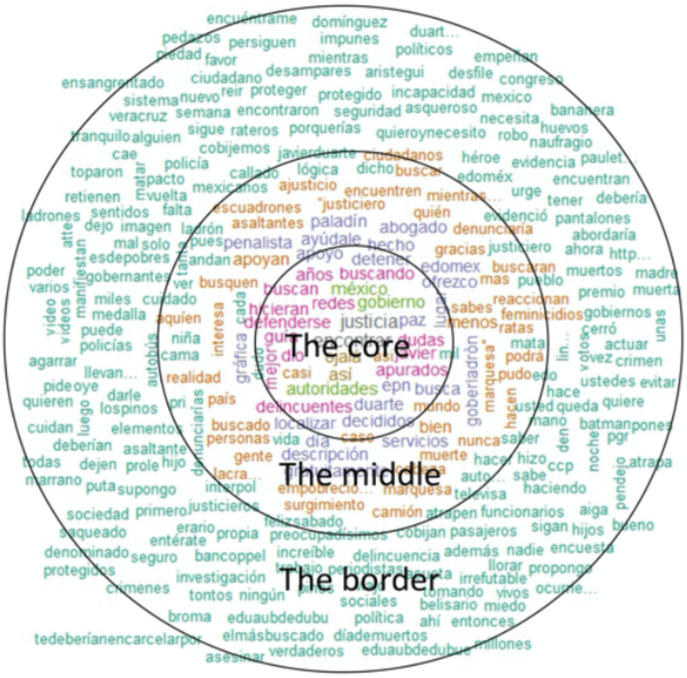


Figure 2.2: The Three Layers of the Radial Word Cloud[8]

Most frequently used words are located in the core layer, the less frequent ones are in the middle layer and the least found words are in the border. From the core, it was found that most people are talking about the justice served in the event, the middle layer shows people discussing the former governor who was corrupt and did not enforce the law. The border layer presents a portion of the mass who want the unknown passenger to be hired as police. So from the word cloud, they derived public reaction to a certain event, which can be further utilized in future events.

One of the most important aspects of digital marketing is social media trends. An analysis[9] conducted on social media trends shows how marketing can be focused on a targeted audience for better outcome. The system retrieves data from three major platforms: Facebook, Twitter, Instagram. Accumulated data from APIs of the three platforms are combined for a specific user. Then the user is categorised based on their preferences. To classify the user, the Multi-layer Perceptron (MLP) where the input is transformed using learnt non-linear transformation. The transformation converts the input data into distinguishable linear data. There can be multiple hidden layers between the input and output layer. But in this study a single layer has been used to create a universal approximator as shown in figure 2.3.

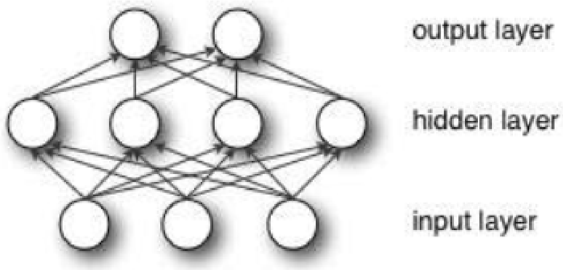
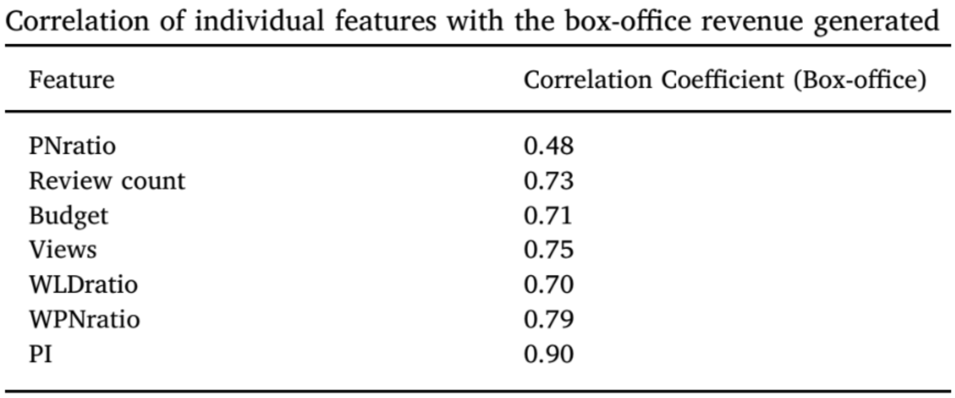


Figure 2.3: Multi Layer Perception[9]

Categorising users creates a personal engagement system which sends ads to users based on their preferences. The proposed system ensures meaningful advertisement for the user and a higher success rate from paid advertisements.

A paper[14] on the motion picture industry focuses on public opinion and movie data to predict the success rate and total revenue of a movie before its theatrical release. The study shows the use of YouTube comments on trailers to find out the purchase intention of the mass. Using the estimated purchase intention along with metadata of the movie, the prediction is completed. In this paper, extracted data from YouTube comment sections are pre-processed and then the TF-IDF algorithm finds the most frequent words. From the array of words, top 200 are manually annotated to be used to figure out the purchase intention. Applying sentiment analysis, the positive and negative reviews are identified among the dataset. The resulting categories are used to predict the revenue by applying SVM, MLR, MLP-NN, RF algorithms. This study is different from previous papers on this domain as most of the previous papers relied on the metadata and production value of a movie. But it can be observed that public opinions are being considered along with other factors in this case. It is also visible that the PI feature has a better correlation with box-office than other features.



Correlation of individual features with the box-office revenue generated

Feature	Correlation Coefficient (Box-office)
PNratio	0.48
Review count	0.73
Budget	0.71
Views	0.75
WLDratio	0.70
WPNratio	0.79
PI	0.90

Figure 2.4: Feature correlation[14]

Therefore, after comparing with 3 experiments, the paper seems to offer better accuracy than previous studies.

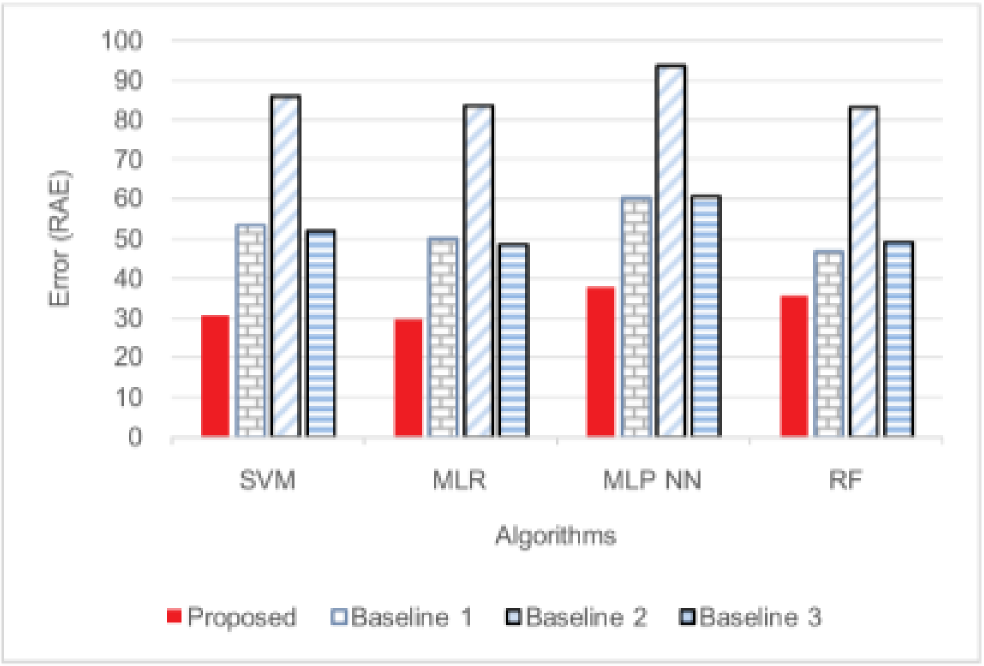


Figure 2.5: Error comparison[14]

2.2 Major Findings

First step of conducting research is to collect sufficient data. Most of the cases mentioned used social media as a primary source for training and test datasets. Twitter has been used the most among the platforms for its handy streaming API. Also data from Facebook and Instagram played a big part in the accumulation of public opinions which sets the stage for trend analysis. The previously referred studies focus on a certain domain and use some keywords & hashtags to collect ample amounts of data. Then comes the classification part. Collected data is classified and categorised based on frequency, similarity, impact and school of thought. Common methods of tabulating data include text mining, deep CNN, NLP and sentiment analysis. Text mining or text data mining is an area of artificial intelligence that is used to detect patterns, extract information, categorization from textual databases such as social media. One of the end results of text mining is word clouds that help determine the public consensus. For image classification, Convolutional Neural Network(CNN) is a popular choice among researchers. Deep learning techniques implemented with CNN can be used to find out the features and details from images. Using the retrieved data, images can be sorted according to their features. k-Nearest Neighbour(kNN) is another technique that can be applied to detect objects from categorized images. The success rate depends on the training dataset and the value of k which can defer for different cases. For detecting relevant topics and concentrating on one focus MF (Matrix Factorization) is used mostly, although LDA (Latent Dirichlet Allocation) is also used in old papers. In addition, TF-IDF algorithm has been applied to detect the words with most frequency. Then TextBlob is used for sentiment analysis. Beside sentiment analysis, python's TextBlob library is a very common and useful tool for pre-processing textual data. It can be also used for text correction, inflection, lemmatization, tokenization, POS tagging (Parts of speech tagging) and N-gram. Moreover, deep learning methods were more preferred than machine learning methods.

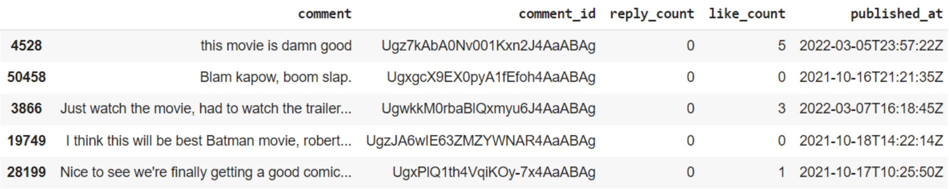
From all of this, we can see that NLP and sentiment analysis plays a big part in finding opinions and extracting emotions out of those data. As for trend analysis, an array of methods/algorithms can be used.

Chapter 3

The Dataset

3.1 Data Collection

For our dataset, we collected comments of movies from their respective YouTube trailers in their official channels from YouTube using the Youtube API. Using Youtube API we collected comment, comment id, reply count, like count and published date.



	comment	comment_id	reply_count	like_count	published_at
4528	this movie is damn good	Ugz7kAbA0Nv001Kxn2J4AaABAg	0	5	2022-03-05T23:57:22Z
50458	Blam kapow, boom slap.	UgxgcX9EX0pyA1fEfoh4AaABAg	0	0	2021-10-16T21:21:35Z
3866	Just watch the movie, had to watch the trailer...	UgwkkM0rbaBlQxmyu6J4AaABAg	0	3	2022-03-07T16:18:45Z
19749	I think this will be best Batman movie, robert...	UgzJA6wIE63ZMZYWNAR4AaABAg	0	0	2021-10-18T14:22:14Z
28199	Nice to see we're finally getting a good comic...	UgxPlQ1th4VqiKOy-7x4AaABAg	0	1	2021-10-17T10:25:50Z

Figure 3.1: Raw Dataset

After that we pre-processed the dataset. Since we are trying to figure out the purchase intention from their initial reaction from the trailer, we removed all the comments after the release of the movie from our datasets. The datasets of all the movies are kept separated. We collected data from 93 movies released in 2018 and 2022. The movies with a budget around 8,000,000 $ to 300,000,000 $ are selected. This large range is selected because in 2021 box office revenue was affected by the COVID 19 restrictions and many production companies relied on streaming services or just delaying the movie altogether which affected the box office performance of most of movies from 2020 and the ones that got delayed to 2021. Moreover, we are excluding movies that have their comments turned off in their official movie trailer. We also made another dataset that consists of movie box office revenue, budget, genre. This information has been taken from Box Office Mojo and cross referenced with The Numbers.

Table 3.1: Movie List

Movie List	
Movie Name	Release date
A Quiet Place part 2	2021-05-28
The Batman	2022-03-04

Continuation of same Table 3.1	
Movie Name	Release date
Candyman	2021-08-27
Chaos Walking	2021-02-24
Death on the Nile	2022-02-11
Dr Strange The Multiverse of Madness	2022-05-06
Dune	2021-10-22
Eternals	2021-11-05
Fantastic Beasts: The Secrets of Dumbledore	2022-04-08
Free Guy	2021-08-13
Ghostbusters: Afterlife	2021-11-19
Halloween Kills	2021-10-15
House of Gucci	2021-11-24
Jackass Forever	2022-02-04
Jungle Cruise	2021-07-30
The Lost City	2022-03-25
Moonfall	2022-02-03
Everything Everywhere All At Once	2022-03-25
Old	2021-07-19
Resident Evil: Welcome to Raccoon City	2021-11-25
Scream	2022-01-14
Snake Eyes: G.I. Joe Origins	2021-07-23
Space Jam A New Legacy	2021-07-16
The Conjuring the devil made me do it	2021-05-26
The King's man	2021-12-22

Continuation of same Table 3.1	
Movie Name	Release date
The Last Duel	2021-10-15
The Matrix: Resurrections	2021-12-22
Uncharted	2022-02-18
Venom: let there be Carnage	2021-10-01
Wrath of Man	2021-04-22
The Northman	2022-04-22
Top Gun Maverick	2022-05-27
King Richard	2021-11-19
West Side Story	2021-12-10
Shang-Chi and the Legend of the Ten Rings	2021-09-02
John Wick 3	2019-05-09
Once Upon a Time... In Hollywood	2019-07-26
Alita: Battle Angel	2019-02-14
Shazam	2019-04-05
Jumanji: The Next Level	2019-12-13
Maleficent: Mistress of Evil	2019-10-18
Godzilla: King of the Monsters	2019-05-31
Knives Out	2019-09-07
Terminator: Dark Fate	2019-11-01
Dark Phoenix	2019-06-07
Us	2019-03-22
Ford v Ferrari	2019-08-30
Rocketman	2019-05-22

Continuation of same Table 3.1	
Movie Name	Release date
Glass	2019-01-18
Dumbo	2019-03-29
Tolkien	2019-05-03
Ad Astra	2019-09-20
Little Women	2019-12-25
A Dog's Way Home	2019-01-11
Bombshell	2019-12-20
Cats	2019-12-20
Escape Room	2019-01-04
Midsommar	2019-07-03
Rambo Last Blood	2019-09-20
Doctor Sleep	2019-10-31
Angel Has Fallen	2019-08-23
Don't Breathe 2	2021-08-12
F9	2021-06-25
Hellboy 2019	2019-04-12
Huslters	2019-09-13
Men in Black	2019-06-11
Nightmare Alley	2021-12-17
No Time to Die	2021-09-30
Nobody	2021-03-26
Replica	2018-10-25
What Men Want	2019-02-08

Continuation of same Table 3.1	
Movie Name	Release date
The Happytime Murders	2018-08-22
12 Strong	2018-01-19
Solo: A Star Wars Story	2018-05-10
Den of Thieves	2018-01-19
A Wrinkle in Time	2018-02-26
First Man	2018-10-12
HOLMES AND WATSON	2018-12-20
Mortal Engines	2018-12-06
Overlord	2018-09-22
Robin Hood	2018-11-11
A Star is Born	2018-10-05
Crazy Rich Asian	2018-08-15
Fantastic Beasts: The Crimes of Grindelwald	2018-11-16
Ready Player One	2018-03-29
Blockers	2018-04-06
Pacific Rim: Uprising	2018-03-23
Skyscraper	2018-07-01
The Girl in the Spider's Web	2018-10-26
Ocean's Eight	2018-06-05
The Darkest Minds	2018-08-01
The Nutcracker and the Four Realms	2018-10-29
The Commmuters	2018-01-12
End of this Table	

We collected the comments from the first trailer released by each of the respective movies, since it is likely to leave a more lasting impression on the viewers. The first dataset is used to create purchase intention from extracting sentiment from the comments while the dataset consisting of the movies box office information is used to compare with the result of our research.
Two types of datasets have been used for this research. The first dataset consists of the comments and the date of the comments called the movie comment dataset. The first dataset is used for determining the overall sentiment toward a given movie. While the second dataset contains some general information regarding the movies.

3.1.1 Movie Comment Dataset

This dataset contains all the comments retrieved from the YouTube comment sections. The dataset contains the names of the movie, publish_at which holds the date of the comment, comment_tokenized which holds the tokenized comment and two sentiment labels, sentiment_label which holds the sentiment labels computed using DistilBert and sentiment_label_roberta which holds the sentiment label computed using RoBerta. The dataset is all separated according to the movies they belong to. This is done so that one-by-one each of the datasets of the movies can be processed and not overburden the system.

3.1.1.1 publish_at

The publish_at column of the dataset contains publish time of the comments of each of the movies. This is utilized for removing comments that were made before the release date of the movie taken from the Movie information Dataset during the preprocessing. This will also contain additional after-processing of the comments using the pre-trained Transformers.

3.1.1.2 sentiment_label

This sentiment label column holds the result from using the DistilBert Transformer. The labels are either 0 or 1 indicating negative or positive.

3.1.1.3 sentiment_label_roberta

This sentiment label column holds the result from using the Roberta Transformer. The labels are either 0, 1 or 99 indicating negative or positive or neutral.

3.1.2 Movie Information Dataset

This dataset contains relevant information regarding the movies. This includes name, date of release, views, the budget of the movie, domestic revenue, total revenue, classification using domestic revenue and classification using total revenue. All the movies that we choose for our dataset have a wide release in the USA.

3.1.2.1 Release Date

This data regards the worldwide release date of the movie. Using this data all the comments from the Movie Comment Dataset are dropped if they contain comments after the release date. All the selected movies are from 2018-2022. And there are no movies from 2020 that have been selected because of their revenue might be affected by the COVID-19 pandemic restrictions.

3.1.2.2 Views

This is the view count of the trailer before the release of the movie. This view count is not available by conventional means without using Youtube Analytics and that is only available if it is the user's own video. Therefore, this information was gathered using the Wayback Machine. However, records of some movies were not available right before their release day. For those movies, a record of the view is taken from any day that was available before the release day.

3.1.2.3 Budget

This is the budget for the movie. This is collected from Box Office Mojo and various other sources. This data along with the budget will determine if the movie is a flop or a hit.

3.1.2.4 Domestic Revenue and Global Revenue

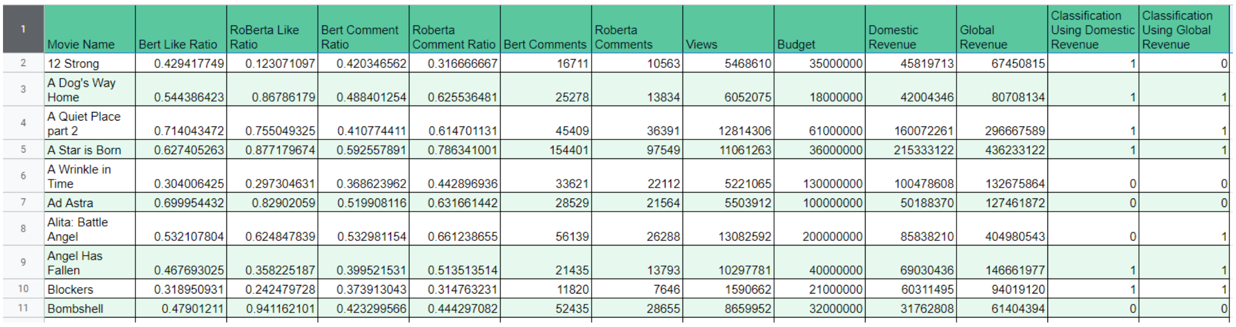
These two columns of data are also collected from Box Office Mojo and cross-referenced with The Numbers. The domestic box office revenue is the revenue that is generated in the home country. In all the cases in our dataset, it is the United States. And the Global Revenue indicates the Total Box Office Revenue. These will be utilized to find the classification of each of the data points.

3.1.2.5 Classification Using Domestic Revenue and Classification Using Global Revenue

The movies were classified as being a hit or a flop based on two features, movie budget and box office revenue. We also considered the domestic box office revenue of the movie as well. It is quite difficult to tell whether a movie is successful or not just from the available data. Since most production companies do not disclose the full amount they spent during the production of a movie. Moreover, a movie just grossing more than its budget may not make it successful. There is the case of overseas box office numbers despite being a major factor in being successful, the cut the production companies receive from the overseas box office revenue is not significant compared to the domestic box office revenue. We utilized the insider rule of thumb of a movie is successful when they make twice as much as its budget in its box office revenue. One feature, global success is made utilizing this logic. Another feature is made utilizing the fact that the movie makes most of its money in its own country or from the domestic box office revenue. If its domestic box office revenue exceeds its budget it is successful. We utilized these two features, one from domestic box office revenue and another from global box office revenue to classify the movies. [4]

3.1.2.6 Additional Features

More features like DistilBert Like Ratio, DistilBert Total Comments, RoBerta Like Ratio and RoBerta Total Comments will be added after running sentiment analysis on the Movie Dataset. Along with the views these four columns will be our main feature for our experiments.



1	Movie Name	Bert Like Ratio	RoBerta Like Ratio	Bert Comment Ratio	Roberta Comment Ratio	Bert Comments	Roberta Comments	Views	Budget	Domestic Revenue	Global Revenue	Classification Using Domestic Revenue	Classification Using Global Revenue
2	12 Strong	0.429417749	0.123071097	0.420346562	0.316666667	16711	10563	5468610	35000000	45819713	67450815	1	0
3	A Dog's Way Home	0.544386423	0.86786179	0.488401254	0.625536481	25278	13834	6052075	18000000	42004346	80708134	1	1
4	A Quiet Place part 2	0.714043472	0.755049325	0.410774411	0.614701131	45409	36391	12814306	61000000	160072261	296667589	1	1
5	A Star is Born	0.627405263	0.877179674	0.592557891	0.786341001	154401	97549	11061263	36000000	215333122	436233122	1	1
6	A Wrinkle in Time	0.304006425	0.297304631	0.368623962	0.442896936	33621	22112	5221065	130000000	100478608	132675864	0	0
7	Ad Astra	0.699954432	0.82902059	0.519908116	0.631661442	28529	21564	5503912	100000000	50188370	127461872	0	0
8	Alita: Battle Angel	0.532107804	0.624847839	0.532981154	0.661238655	56139	26288	13082592	200000000	85838210	404980543	0	1
9	Angel Has Fallen	0.467693025	0.358225187	0.399521531	0.513513514	21435	13793	10297781	40000000	69030436	146661977	1	1
10	Blockers	0.318950931	0.242479728	0.373913043	0.314763231	11820	7646	1590662	21000000	60311495	94019120	1	1
11	Bombshell	0.47901211	0.941162101	0.423299566	0.444297082	52435	28655	8659952	32000000	31762808	61404394	0	0

Figure 3.2: Dataset sample

3.2 Data Pre-processing

Data Pre-processing is a crucial step for enhancing the performance of our model. Since comments directly extracted from the YouTube comment section are usually filled with misspelling, errors and jargons. There are also comments from other languages that would make it hard to accurately predict the sentiment since the translations of the comments may not be accurate. Therefore, we are only considering the English comments. By doing so, this will help us better pre-process the dataset and later on it will reduce the errors and improve the accuracy of our predictive algorithm as we are using the comments to analyze the sentiment of the comment, which is the base of our research. After cleaning the comments we start our process of pre-processing the data for accurately classifying the comments. For pre-processing, we are only focusing on cleaning up the data so that it can be properly processed by the encoders for labeling.

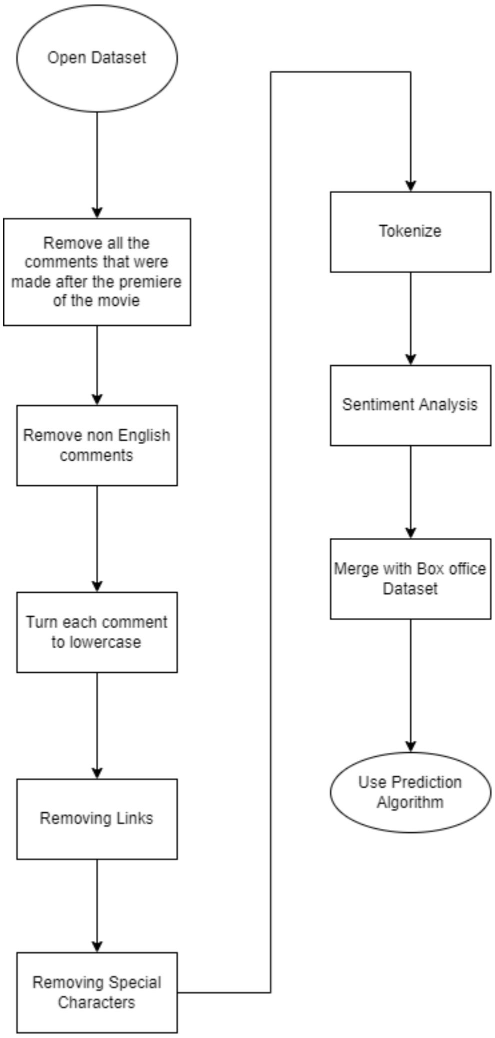

Open Dataset
Remove all the comments that were made after the premiere of the movie
Remove non English comments
Turn each comment to lowercase
Removing Links
Removing Special Characters
Tokenize
Sentiment Analysis
Merge with Box office Dataset
Use Prediction Algorithm

Figure 3.3: Data Pre-processing

3.2.1 Data Reduction

All the extracted comments may not be useful for our research. Therefore, first, we need to remove all the comments that were made after the release of the movie. The comments that were made after the release of the movie are more likely to comment that the users made after they have watched the movie. Since we are trying to evaluate whether people are having positive or negative reactions after watching a given trailer and before the release of the movie itself, we have decided to remove those comments from our dataset. Although those comments might give a better insight into what people thought after watching the movie, we are only considering the comments before the release date to gauge whether the initial reaction to the trailer translates to the final success of the movie itself. We also dropped all the comments that are not in English. Since even if it is possible to translate the comments, it might not be an accurate translation due to how complex translating language can be and it might not be adequate enough to capture the user's sentiment of the comment. Moreover, comment sections like these are usually filled with spam which has links to other websites. So we used regex to find these comments with such links and mark them for removal.

3.2.2 Data Cleaning

Data cleaning is the process of removing irrelevant aspects of the data and fixing errors in the data. We convert all the comments to lowercase which will help us with pre-processing. Since YouTube comments allow the use of emoticons, however, the reason for using emoticons might not be clear for everyone, so we have removed all the emoticons and special characters from the comments. Many of the comments utilize multiple punctuations or spaces to emulate conversations to express their sentiment the comments. However, this type of comment needs to be encoded different way to capture its actual sentiment. Moreover, there could be other types of sentence structure used in the comments. Employing different models to classify each of the sentences depending on their structure might be very complex. So we employed a more simple approach of preserving most of the comment structure but removing unnecessary padding or spaces used in the comments. So we removed special characters and punctuation. With that, the comments are ready for pre-processing.

3.2.3 Pre-processing

For pre-processing, we only going to lowercase all the characters in the comments and tokenize the comment to break it down. We opted for removing the stop words for preventing the loss of data and moreover removing stopwords does not change much in the case of the final result. Finally, tokenization is done. Utilizing models like Bert and RoBerta, we will evaluate the sentiment of the comment. We will also be truncating the size of the comments to 512 words to comply with the limit of the encoders.

3.2.4 Tokenization

The function to tokenize consecutive words in the length of n is called n-gram. Since we are trying to calculate the sentiment analysis of each of the comments, we tokenize each of the words. By tokenizing the comments it makes it easier to analyze comments and apply them to our models. We also use the length of the tokenized list and remove them if they are empty or only consist of one item. Moreover, this token list can be utilized to truncate the size of the comment for that exceed the 512 length of the encoder.

Chapter 4

Methodology

4.1 Sentiment Analysis

After our data is pre-processed and removed of unnecessary data we need to classify the comments with sentiment analysis and give them a positive or negative score. If they give a positive comment we assume that they are more likely to watch the movie while the negative comments. We will use DistilBert and RoBerta to analyze the sentiment of the comments.

4.1.1 BERT Model

For further accuracy, we will also use the BERT model to find out the sentiment of the comments column so we can further predict which will perform better in both of these options. BERT stands for Bidirectional Encoders Representations from Transformers. As the name suggests this model is made up of the encoder part of the transformer which consists of an encoder and decoder. In the BERT base it consists of 12 layers and the BERT large model contains 24 layers[22].

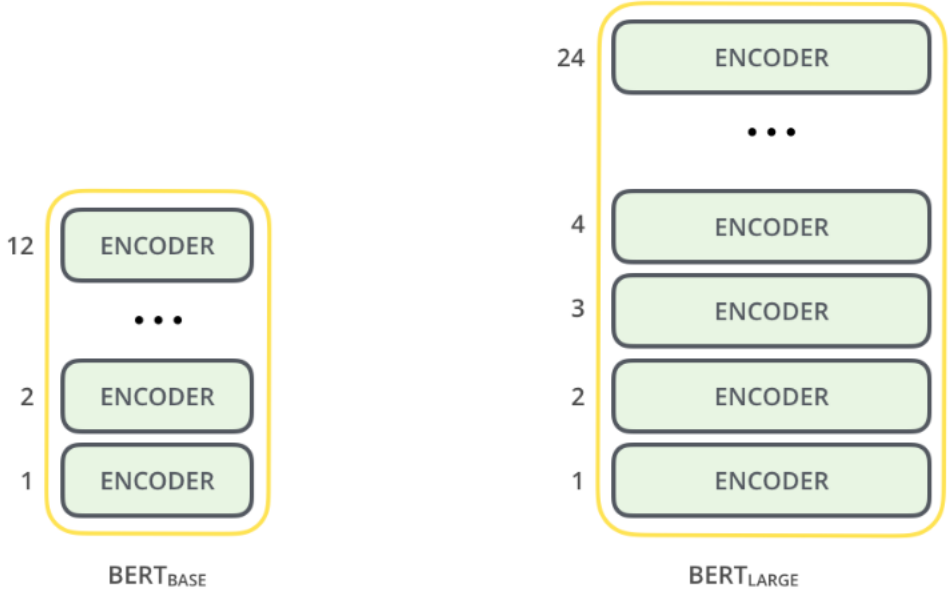


Figure 4.1: BERT base and large model[22]

Using a pre-trained BERT model that is next fine tuned to predict results. In the pre-training phase the model is fed masked sentences and the BERT model will try to predict the masked words. In this Masked Language Model (MLM) it trains the model to predict the masked words in context with the sentence and as it is bidirectional it reads from both left-to-right and right-to-left context[21].

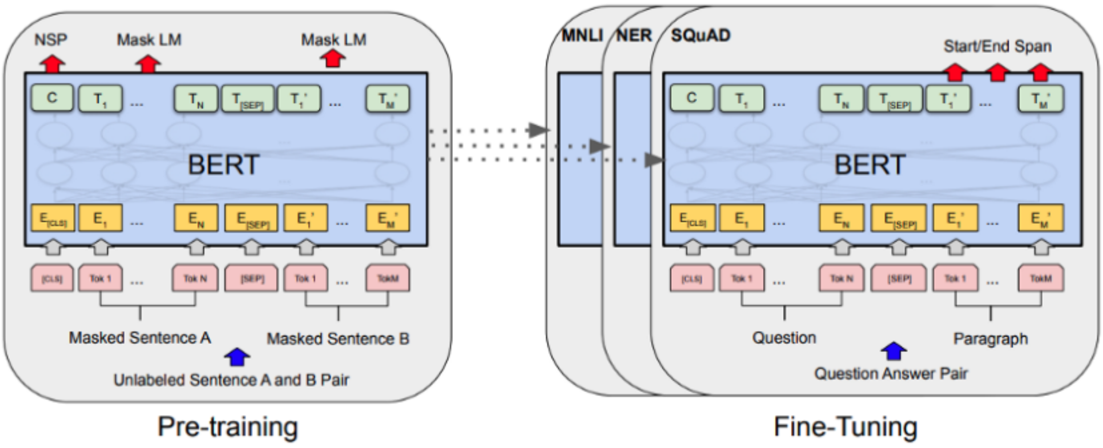
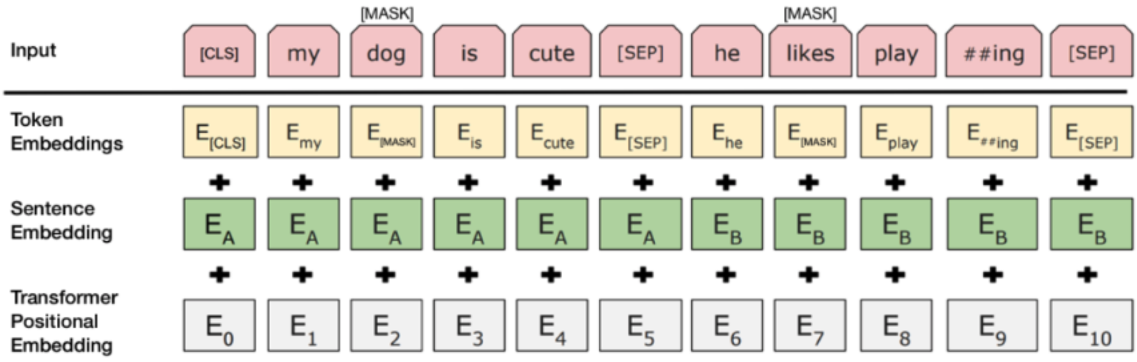


Figure 4.2: BERT pre-training and fine-tuning[21]

Also, at the same time, the BERT model will take a pair of sentences and try to predict if sentence A is followed by sentence B. This is done in the training phase and added with random sentences to further predict the output. This is called Next Sentence Prediction (NSP). In this process, both input sentences are tokenized and are added with a starting [CLS] and ending [SEP] tokens. The token embeddings are then passed to sentence embedding to mark them to sentence A or B. Finally, they are passed to positional sentence embedding to find their position in the sequence of the sentence[10].



			[MASK]					[MASK]			
Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	$E_{[MASK]}$	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	$E_{[MASK]}$	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Sentence Embedding	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Transformer Positional Embedding	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

Figure 4.3: BERT NSP[10]

For our sentiment analysis using the BERT model we will use the NSP and add a top layer for classification of the sentences for the cls token[10].

4.1.1.1 DistilBert

DistilBert is build upon the BERT model but is much cheaper, lighter, and faster, Still using the same amount of knowledge it uses less space and is much faster in its process. This is notable for its lower consumption on power. DistilBERT uses a knowledge distillation method where it learns from the large teacher model to help train the student model[13]. In this way, without using the power of the full scale model it can try to replicate its results with lesser resources.

4.1.1.2 RoBERTa

RoBERTa is another type of BERT model with optimized pretraining approach. It uses byte level tokenizer and changes some hyper-parameters but has the same architecture as BERT. Compared to BERT, it is trained using bigger batches and longer sequences. It was made for optimizing the training of BERT and imporving the performance of the model[12]. This model was made using Facebook AI.

4.2 Prediction Algorithm

For our prediction algorithm we have picked Logistic Regression, MLP, KNN, DT and RFC. We have decided to use a non linear classifier to implement on our dataset as it yields better results for NLP classification[15]. We will use the positive comments and their likes as weight for each of the comments which includes the number of people who are anticipating to watch the movie and those who are not and the result of the sentiment analysis of the comments. We will also use the box office dataset in conjunction with the dataset containing the result of the sentiment analysis and apply it to our predictive algorithms which are Logistic Regression, MLP, KNN, DT and RFC to predict if the resulting movie is going to be a flop or a hit.

4.2.1 Multilayer Perceptron

MLP stands for multi-layer perceptron. For movie revenue prediction this is the most used algorithm[17]. This algorithm uses hidden layers to classify the data and can be used to predict the movie revenue. MLP has input layers and output layers. From the input layers, it takes the weight of the inputs and passes them through a hidden layer to calculate the weight of the inputs through activation functions and give an output. Using this, we can classify and predict the movie revenue for our case.

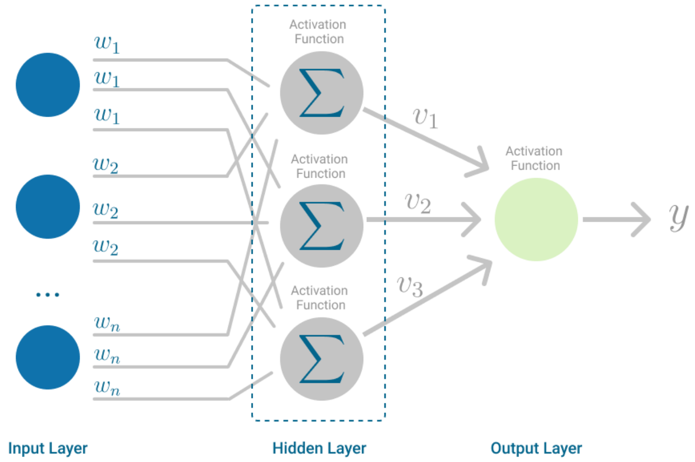


Figure 4.4: Multilayer Perceptron[17]

4.2.2 K-Nearest Neighbors

KNN is a supervised machine learning algorithm which is trained on labeled data and is used to assign an appropriate output to unlabeled data[24]. This algorithm depends on the value of K which indicates the number of neighbors. When a new datapoint is given as input, the number of nearest neighbors are considered for classifying the new input. In other words, there are no weights associated with the inputs but rather we check to see if the Euclidean distance of a new datapoint is closest to which class and classify the new datapoint to its nearest neighbor[24].

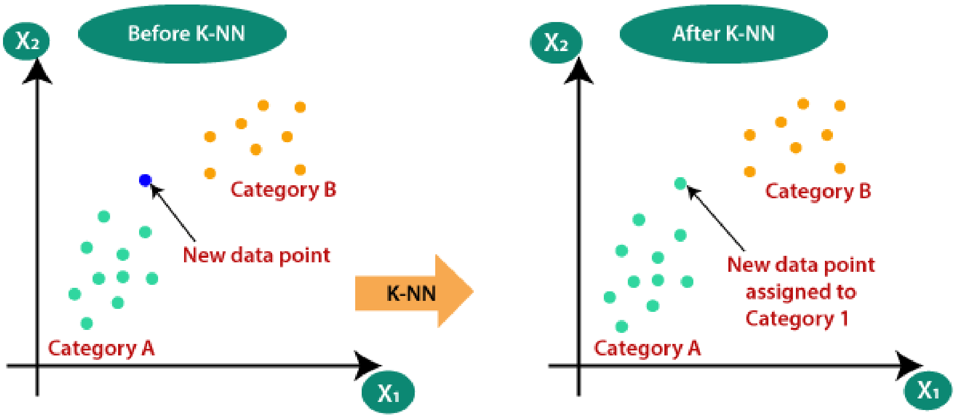


Figure 4.5: KNN[24]

For our prediction algorithm, KNN is used to detect the success probability of a certain motion picture. Using the output from sentiment analysis, a new datapoint, in this case a new movie, will be compared using the trained model. Based on the value of K, classification will be conducted.

4.2.3 Decision Tree

Decision tree is a machine learning algorithm which can be applied to design a predictive model and perform classification. It falls on the supervised learning category meaning the features and labels have to be predefined for the decision tree to learn and predict. Based on the attributes, the algorithm builds a tree structure of the model. The tree is constructed of leaf nodes and non-leaf nodes. Non-leaf nodes represent the attributes/features of the dataset. Information derived from the features are used to branch out the formulated tree. Leaf-nodes are results or outputs. So, the outcome or decision is found through leaf nodes. The first or starting node is known as the root node of the tree. Now the algorithm picks the feature with highest information purity to be the root node. Information purity refers to the conciseness of the information provided by a feature. An attribute is deemed more pure if it has more information purity. The feature with highest purity is chosen as the root node. Information purity is calculated using Entropy and Information Gain.

Entropy: It is the mathematical procedure to calculate information purity. Entropy is inversely proportional to information purity. So, higher value of entropy means that feature is less pure.

$$Entropy(S) = \sum_{i=1}^{n} Pi \log_2 Pi \tag{4.1}$$

Where S refers to the attribute/label, the value of i depicts the number of labels. So, the number of labels can be from 1 to N. Pi is the probability of i-th term. Here, the sum of the negative value of Pi*log2Pi is taken to calculate entropy.

Information Gain: Using the values of entropy, information gain is evaluated. Information gain is used to assess the attributes and the one with highest information gain is picked to be a non-leaf/decision node.

$$InformationGain(B) = Entropy(S) - \sum_{v \epsilon A} Px * Entropy(x) \tag{4.2}$$

Here, B is the attribute, S is the label, A is the feature, x is the number of unique outcomes under A, Px is the probability of each outcome. To calculate information gain, the summation of Px * Entropy(x) is subtracted from Entropy(S). This process is repeated for each feature to find out the highest information gain. If we do not get unambiguous outputs after picking a feature, we have to redo the whole process under the condition that previously picked features are occurring. This process might need to be repeated till all the features are exhausted. Even after going through all the features in a branch if a definitive answer is not found, the majority of the outcomes is considered the result.

4.2.4 Random Forest

Random forest is a machine learning algorithm that utilizes decision trees. Therefore it is also a supervised learning technique. Since Random Forest is an ensemble learning algorithm, it combines multiple decision trees to produce results with higher precision. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting[11].

In this algorithm, the use of numerous trees leads to the accurate prediction instead of an estimation. The predictions from each tree is taken into account and then the average outcome is chosen. It works for large datasets and does not overfit.

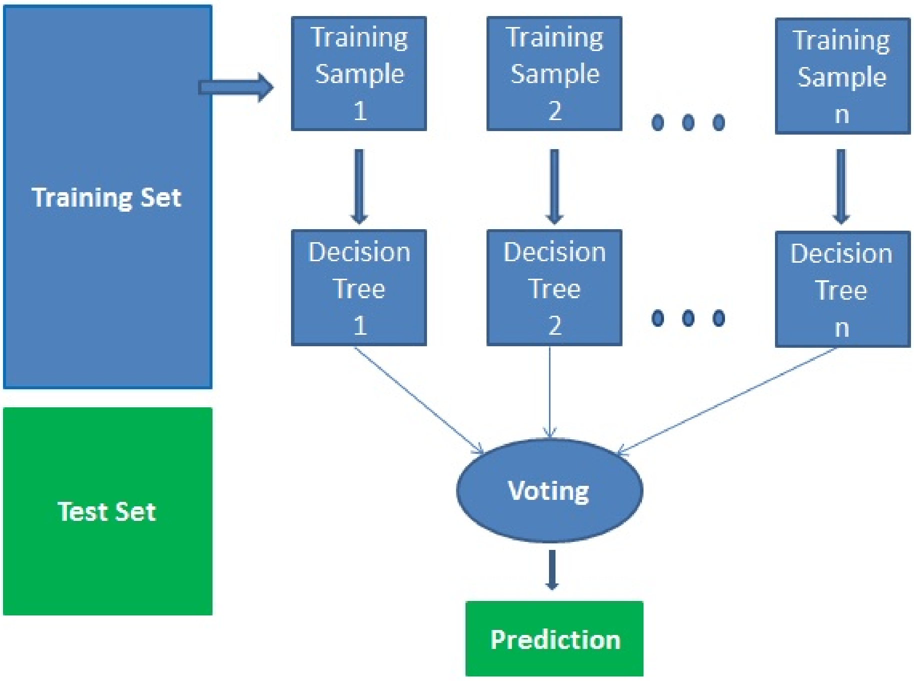


Figure 4.6: Random Forest[11]

4.3 Evaluation of the Model

4.3.1 Accuracy

Accuracy is the result of how well the model predicted the values correctly here. It is given with the equation:

$$\textbf{Accuracy:}\ \frac{TP+TN}{TP+TN+FP+FN}$$

Here:
TP: stands for true positive. When both prediction and outcome are true.
TN: stands for true negative. When both prediction and outcome are negative.
FP: stands for false positive. When prediction is positive but outcome is negative.
FN: stands for false negative. When prediction is negative but outcome is positive.

4.3.2 Precision

Precision is the ratio of positive values predicted correctly out of all the positive values. This shows us how well our predictions are predicting positive values.

$$\textbf{Precision: } \frac{TP}{TP+FP}$$

4.3.3 Recall

Recall is the ratio of predicted positive values out of all values and it is given by the equation:

$$\textbf{Recall: } \frac{TP}{TP+FN}$$

4.3.4 F1 Score

F1 score takes into account precision and recall to give us the combined result. We want to minimize the number of false labeling so f1 score can give us a better picture.

$$\textbf{F1 Score: } 2 * \frac{Precision * Recall}{Precision + Recall}$$

Chapter 5

Result Analysis

5.1 Sentiment Labeling

First, we label each of the datasets that consist of comments on the movies utilizing Bert models like DistilBERT and RoBerta. For implementing DistilBERT and RoBerta we utilized the pre-trained Hugging Face transformers. We implemented DistillBERT we used “distilbert-base-uncased-finetuned-sst-2-english” pre-trained transformer from Hugging Face. This model was trained using glue and sst-2 dataset and classifies each of the sentences it processes into either positive or negative depending on the assigned score. For implementing RoBerta “cardiffnlp/twitter-roberta-base-sentiment” was used. This transformer was mainly used for evaluating tweets on Twitter. However, since the structure of comments under YouTube trailer is somewhat similar we decided to use this transformer to classify comments using this pre-trained transformer. This transformer classifies each of the comments as positive, neutral or negative. We classified the sentiment for all the comments in the dataset and then added up all the positive comments and negative comments. We used the likes for each of the comments as their weight. So, likes in this case act as support to the comment.

$$\begin{aligned}\textbf{Positive Comments} = {} & \text{Total Number Positive Comments} + \\ & \text{Number of Likes for the Positive Comments}\end{aligned}$$

We mainly got two features from this process, DistilBERT positive comment ratio and RoBERTa positive comment ratio. We only used the positive like ratio as the negative comment ratio is just the inverse of that features.

$$\textbf{Positive Like Ratio} = \frac{\text{Positive Comments}}{\text{Total Likes} + \text{Total Comments}}$$

Both of these features are then added to the Movie Inforamtion Dataset. On which we will run our prediction algorithms to estimate whether a given movie is going to be a hit or a flop.

5.2 Evaluation

Four prediction algorithms were utilized, specifically MLP, K-Nearest Neighbors, Decision Tree, and Random Forest to perform our experiments. Four different tests have been performed depending on the input dataset.
After each prediction Recall, Precision, Accuracy and F1 score are calculated for each prediction algorithm to evaluate their performance.

Test No.	Input Features	Output Feature
1	RoBerta Like Ratio, Roberta Comments, Budget, Views	Classification using Domestic Revenue
2	DistilBert Like Ratio, Bert Comments, Budget, Views	Classification using Domestic Revenue
3	RoBerta Like Ratio, Roberta Comments, Budget, Views	Classification using Global Revenue
4	DistilBert Like Ratio, DistilBert Comments, Budget, Views	Classification using Global Revenue

Table 5.1: Features and labels

Since, the created dataset only has 94 datapoints. The dataset has been split into 8:2 ratio for training and testing. We utilized Standardization for pre-processing this data and upsampled the dataset. For utilzing classification using Domestic Revenue as output we upsampled 1 or movies that are hit from 41 to 52 and for classification using Global Revenue we needed to upsample 0 from 39 to 54 to avoid under-fitting.

5.3 Results

We have chosen a few standards on which to establish our findings. By choosing specific features to obtain different results, such as choosing attributes we obtained from one model in one and utilizing a different model for the rest, we have built four test cases with changeable features. We choose from among our characteristics to achieve the greatest outcomes, and the results are contrasted below. To obtain findings that are not under-fitted, we up-sampled and standardized our original dataset of 93 data points.

For our test case 1 we have picked only the RoBerta like ratio and comment counts with budget and views as our feature. Our label is classification with domestic revenue. We have used RoBERTa model here to get the sentiment of our comments. Next, we applied different algorithms to our dataset with an 8:2 train and test split. Table 5.2 contains the algorithms' results.

In test 1, we found that RFC had the highest accuracy. Additionally, it has a recall value of 1, making it possible to identify all movies that were genuine box office successes. The precision value, however, is 91%, meaning that 91 percent of the test results were accurately detected. Additionally, the f1 score is 95%, indicating that

	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.81	0.889	0.727	0.8
MLP	0.857	0.833	0.909	0.87
KNN	0.857	0.786	1	0.88
Decision Tree	0.857	1	0.727	0.842
Random Forest Classifier	0.952	0.917	1	0.957

Table 5.2: Accuracy, Precision, Recall and F1 Scores of Test 1

there were not many false positives and false negatives. The accuracy of all other algorithms is less than 85%, hence they are not the best fit. DistilBert Like Ratio, DistilBert Comment Count, Budget, Views, and label was classified using domestic revenue were our features for test 2. Table **??** contains the results we obtained from this.

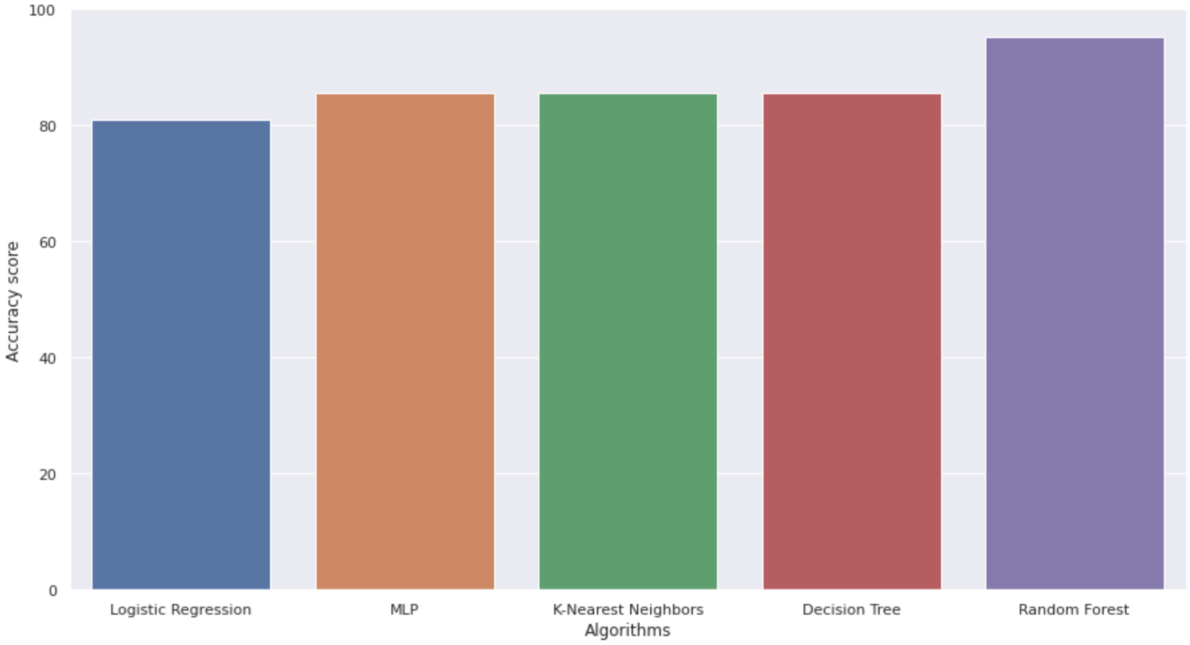


Figure 5.1: Algorithms Accuracy for Test 1

For test 2, our features are DistilBert Like Ratio, DistilBert Comment Count, Budget, Views and label was classification using domestic revenue. The results are comparable to our test 1 as seen in the 5.3.

	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.81	0.889	0.727	0.8
MLP	0.905	0.846	1	0.917
KNN	0.905	0.909	0.909	0.909
Decision Tree	0.952	0.917	1	0.957
Random Forest Classifier	0.952	1	0.909	0.952

Table 5.3: Accuracy, Precision, Recall and F1 Scores of Test 2

We used DistilBert model for our sentiment labeling here and with that we calculated the ratios. As we can see from the table the result is also similar to our test 1 with RFC having 95% accuracy. Additionally, since this precision is 100%, no false positives were found. However, Decision Tree also offers us superior precision, recall, and f1 score along with the same accuracy. Decision Tree algorithm is the best fit for this test since it has a higher f1 score.

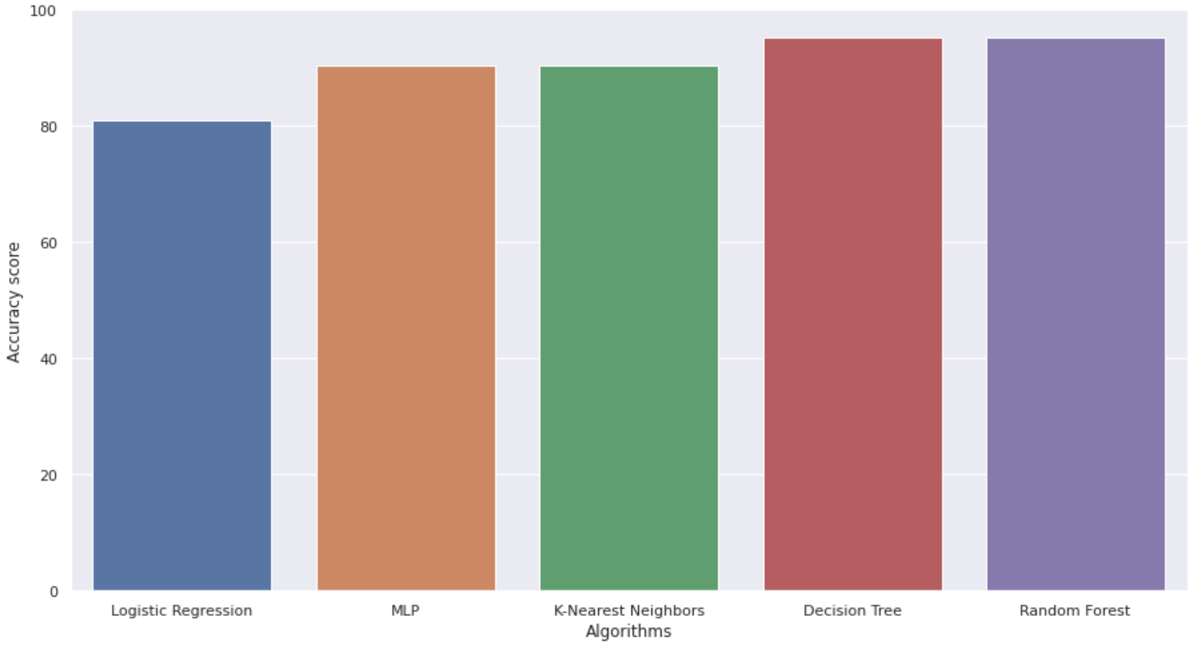


Figure 5.2: Algorithms Accuracy for Test 2

We repeated test 1's features for test 3, however this time we utilized a different label. The results are shown in Table 5.4. The classification is made based on global revenue.

	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.682	0.727	0.667	0.696
MLP	0.727	0.714	0.833	0.769
KNN	0.773	0.818	0.75	0.783
Decision Tree	0.864	0.909	0.833	0.87
Random Forest Classifier	0.818	0.833	0.833	0.833

Table 5.4: Accuracy, Precision, Recall and F1 Scores of Test 3

As we can see, every accuracy score is below 81%, with RFC having the maximum accuracy of 81%. In comparison to earlier test cases, all of the algorithms had poor precision, recall, and f1 scores.

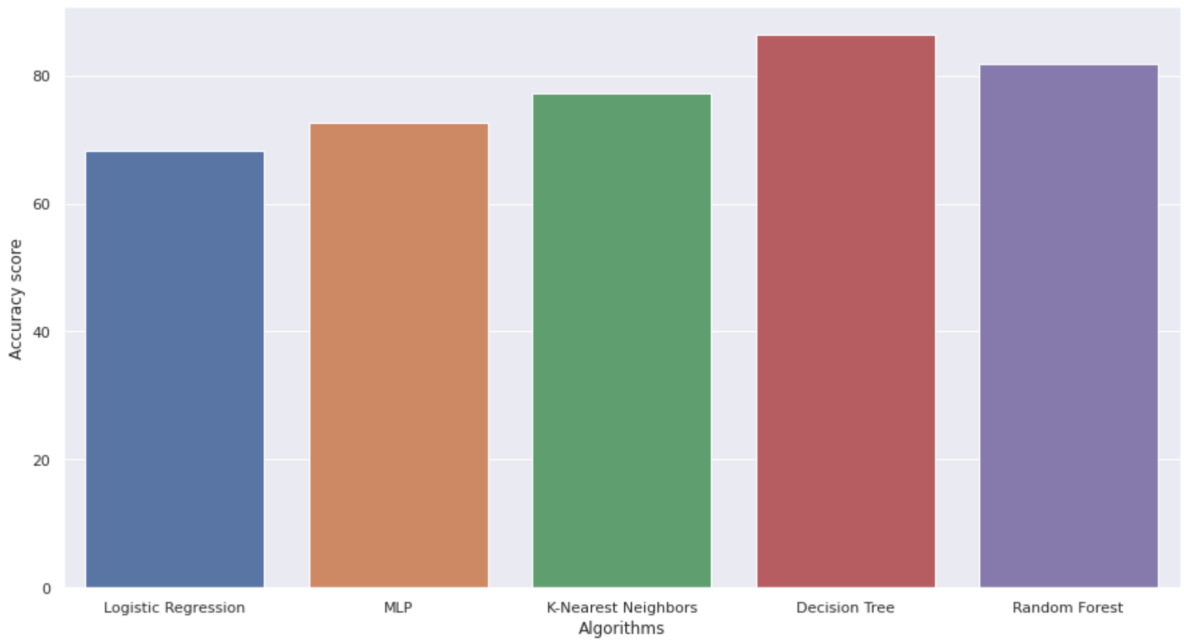


Figure 5.3: Algorithms Accuracy for Test 3

We utilized the same features as test case 2 for test case 4, but we classified thelabels differently using global revenue. For this test instance, the outcomes of the various algorithms are shown in Table 5.5.

	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.773	0.733	0.917	0.815
MLP	0.773	0.733	0.917	0.815
KNN	0.864	0.846	0.917	0.88
Decision Tree	0.909	0.857	1	0.923
Random Forest Classifier	0.955	0.923	1	0.96

Table 5.5: Accuracy, Precision, Recall and F1 Scores of Test 4

As can be seen, RFC outperforms all other compared algorithms in this set of test instances, with a recall score of 1 and an overall f1 score of 0.96.

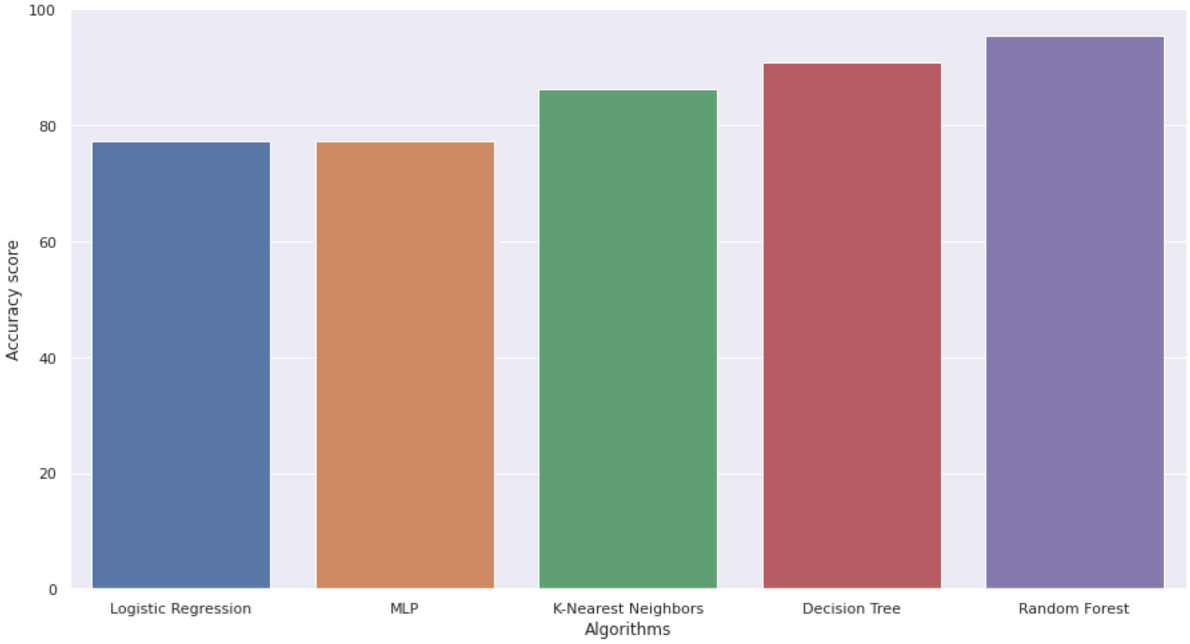


Figure 5.4: Algorithms Accuracy for Test 4

These four experiments demonstrate that the Random Forest Classifier performs the best across the range. When comparing the outcomes of all the prediction algorithms, experiments using Classification with Domestic Revenues as their output data revealed higher forecasts overall. Despite the greatest RFC forecasts in Test #4. This statistic may be distorted since we up-sampled the data to match the amount of hits and flops. And for the reason for the prediction algorithms not faring better in predicting the classification of Classification using Global Revenues is that we only considered comments that are English. Additionally, there are trailers that were released expressly for other nations in various languages that we did not take into account during our tests. So, the proposed model fares better in predicting the success of the movie in their Domestic Region.

Chapter 6

Conclusion

In this study, we used sentiment analysis to classify a product's success using comments from social media sites like YouTube. In our case, we tested this using YouTube movie trailers. Since using data from YouTube comments and their sentiment analysis to predict whether a movie may be a hit or a flop produced satisfying results, this approach can be applied to the case for other items as well.
For pre-processing we only focused on data cleaning and data reduction to prevent losing textual data. Random Forest gave the best results among all the prediction algorithms used in our research. Moreover, Our model gave better results while classifying the movies that will do well in the domestic region of the movie utilizing distilBERT features. Our research affirms that the success of movies can be accurately predicted by utilizing only data from YouTube comments and Budget.

6.1 Future Scope

The findings of this research showcase how effective utilizing YouTube data can be for predicting whether or not a movie will be a box office success simply by using comments and their sentiment analysis. This can be expanded to accurately predict the box office revenue of a movie considering that there are enough data points in the dataset. Moreover, training a model using labelled data on a specific genre of a movie will yield better results rather than using pre-trained transformer models. For example, in the case of horror movies, many of the comments under their trailer may be labelled as negative when utilizing pre-trained transformers. As most of the people are either commenting about how scary, gory or frightening the trailer is. Even if the comments showcase anticipation toward the movie it might be classified as negative in the result of sentiment analysis. Therefore, if the emotion of the comment is included as a feature in training the model for categorizing comments from each movie genre, that method is better for predicting box office revenue. However, Finding labelled data, can be challenging in this situation. This issue can also be solved by employing other transformers that highlight emotions in place of positive-negative sentiment analysis. This can be used to any subject that relates to what our research is trying to accomplish.

Furthermore, rather than just classifying hit or flop, by increasing the data points in the dataset we might be able to predict the overall box office revenue. We can also take into account the popularity of particular topics on other social media websites, such as Reddit, Twitter, TikTok etc. In our research, we only considered the first released trailer by the production company on YouTube. For designing a model specifically predicting box office revenue the overall trend in the top sites may be used as features as well. Moreover, additional features such as the director, writer, producers and actors involved in a given movie may be used when creating a model that explicitly forecasts box office income. Since our project is only focusing on utilizing social media comments to find how successful a product can be we only opted only using the sentiment in the comments and the budget.

Lastly, utilizing comments from other languages can give better results in predicting the global box office revenue. Since nowadays, they release videos for different languages. Analyzing the comments from those videos may give more insight to how well the movie will do in each country's box office.

Bibliography

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," en, in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, vol. 10, Not Known: Association for Computational Linguistics, 2002, pp. 79–86. DOI: 10.3115/1118693.1118704. (visited on 01/19/2022).

[2] T. Sakaki, F. Toriumi, and Y. Matsuo, "Tweet trend analysis in an emergency situation," in *Proceedings of the Special Workshop on Internet and Disasters*, 2011, pp. 1–8.

[3] D. S. Kim and J. W. Kim, "Public Opinion Sensing and Trend Analysis on Social Media: A Study on Nuclear Power on Twitter," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 9, no. 11, pp. 373–384, Nov. 2014. DOI: 10.14257/ijmue.2014.9.11.36. (visited on 01/19/2022).

[4] J. M. III and J. M. III, *Building the perfect bomb: The numbers behind box office flops, popmatters*, Jul. 2015. [Online]. Available: https://www.popmatters.com/192562-box-office-flops-or-building-the-perfect-bomb-2495539022.html.

[5] D.-R. Liu, H. Omar, C.-H. Liou, H.-C. Chi, and C.-H. Hsu, "Recommending blog articles based on popular event trend analysis," *Information Sciences*, vol. 305, pp. 302–319, 2015.

[6] R. Kaushik, S. A. Chandra, D. Mallya, J. Chaitanya, and S. S. Kamath, "Sociopedia: An interactive system for event detection and trend analysis for twitter data," in *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, Springer, 2016, pp. 63–70.

[7] G. Amato, P. Bolettieri, V. Monteiro de Lira, C. I. Muntean, R. Perego, and C. Renso, "Social media image recognition for food trend analysis," in *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 2017, pp. 1333–1336.

[8] D. Valle-Cruz, J. E. Vega-Hernández, and R. Sandoval-Almazán, "Justice of the marquesa: A twitter trend analysis using text mining and word clouds," in *Proceedings of the 18th Annual International Conference on Digital Government Research*, 2017, pp. 592–593.

[9] H. N. Bhor, T. Koul, R. Malviya, and K. Mundra, "Digital media marketing using trend analysis on social media," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, IEEE, 2018, pp. 1398–1400.

[10] R. Horev, *Bert explained: State of the art language model for nlp*, Nov. 2018. [Online]. Available: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.

[11] A. Navlani, *Sklearn random forest classifiers in python tutorial*, May 2018. [Online]. Available: https://www.datacamp.com/tutorial/random-forests-classifier-python.

[12] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[14] I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, "Movie revenue prediction based on purchase intention mining using youtube trailer reviews," *Information Processing & Management*, vol. 57, no. 5, p. 102 278, 2020.

[15] I. S. Ahmad, A. A. Bakar, M. R. Yaakub, and S. H. Muhammad, "A survey on machine learning techniques in movie revenue prediction," *SN Computer Science*, vol. 1, no. 4, pp. 1–14, 2020.

[16] "Sentiment Classification System of Twitter Data using Python," en, *International Journal of Recent Technology and Engineering*, vol. 8, no. 6, pp. 1042–1044, Mar. 2020, ISSN: 2277-3878. DOI: 10.35940/ijrte.F7348.038620. (visited on 01/19/2022).

[17] C. Bento, *Multilayer perceptron explained with a real-life example and python code: Sentiment analysis*, Sep. 2021. [Online]. Available: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141.

[18] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," en, *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[19] S. Kemp, *Digital 2021 july global statshot report - datareportal – global digital insights*, Oct. 2021. [Online]. Available: https://datareportal.com/reports/digital-2021-july-global-statshot.

[20] M. Monselise, C.-H. Chang, G. Ferreira, R. Yang, C. C. Yang, *et al.*, "Topics and sentiments of public concerns regarding covid-19 vaccines: Social media trend analysis," *Journal of Medical Internet Research*, vol. 23, no. 10, e30765, 2021.

[21] O. G. Yalçın, *Sentiment analysis in 10 minutes with bert and hugging face*, Feb. 2021. [Online]. Available: https://towardsdatascience.com/sentiment-analysis-in-10-minutes-with-bert-and-hugging-face-294e8a04b671.

[22] J. Alammar, *The illustrated bert, elmo, and co. (how nlp cracked transfer learning).* [Online]. Available: http://jalammar.github.io/illustrated-bert/.

[23] *Global social media stats - datareportal – global digital insights.* [Online]. Available: https://datareportal.com/social-media-users.

[24] *K-nearest neighbor(knn) algorithm for machine learning - javatpoint.* [Online]. Available: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.