BSc (Hons) in Computing
Level C/I/H

INDIVIDUAL ASSIGNMENT

Module Code & Title: Final Year Project

Prepared by: [Ashif Shakib] [(CB007534)] [HF20A1SEENG]

Date of Submission: 30th June 2021

Instructor: Mr. Priyantha Kumarawaduge

Supervisor: Mrs. Dimanthinie De Silva

Accessor: Mrs. Umanga Pilapitiya

Word Count:
18,516

Abstract

“IMDB Movie Review Analyzing System” is a system that will analyze Movie or TV show reviews in IMDB and predict the movie or TV show is Impressed the audience or not. Most of the people before watching a movie or a TV show they usually go through YouTube trailers. Even though they watch the trailer, trailer does not give a that much of idea of the movie/TV show. Sometimes the trailer is interesting but the movie/tv show is not that much interesting. Some people read reviews about the movie in social media and movie web platforms such as IMDB and Rotten Tomatoes. In social media and web review platforms, there will huge numbers of reviews. Reading that much of reviews will be hard thing to do and it will be a time-wasting process. For this issue, this system will be the best solution. This system will read the latest IMDB reviews about the Movie/Tv show that user wants to watch and recommend the user to watch the movie or not.

Acknowledgement

First, I would really thank to Mrs. Dimanthinie Silva for helping me to complete this assignment. Beginning of the Semester I really had no scope of this assignment and Mrs. Dimanthinie silva helped me a lot and made me to understand about the project. Then I would like to Thank my batchmates and parent for supporting and encouraging me.

Table of Contents

Table of Figures

Chapter 01: Introduction

Context

In recent years, a rising number of people have chosen to spend their free time watching movies since they give amusement and relaxation. It is a wonderful trend in my opinion, as long as it's done in moderation. To begin with, many people believe watching movies to be the ideal way to unwind after a long day at work. More specifically, most people these days have a demanding schedule and demanding work, and they look forward to finding ways to relax their mind and body. Movies, they believe, not only entertain them but also help them forget their troubles and stresses. There are many methods to watch movies these days that do not require going to a movie theater. People may now unwind and watch a movie or TV show in the privacy of their own home or yard! Watching movies can be a fantastic way to relieve stress in one's life. Stress is generated by a constant buildup of tension within a person, which is medically proved, and without a mechanism to remove it, stress is unavoidable. Watching movies is one of the most effective methods for releasing tension. Laughter and bonding time are provided by movies. That is what a good comedy film can accomplish. It is a method to bond as well as a means to lighten the mood.

Aside from that, watching movies requires no physical or mental work, which makes them the preferable option given that the bulk of occupations nowadays are demanding and take a lot of effort. It is not for everyone to go to the movies. Some have difficulty with sensory difficulties or crowds. Others simply like to watch movies at home. The good news is that it does not matter if you are at home watching Netflix or at a crowded theater. People might get so busy and agitated in today's fast-paced world that they have little time to enjoy moments and create memories. Watching movies is worthwhile for a variety of reasons, including entertainment, time with friends, and a simple way to reduce stress.

Problem Statement

To reduce the stress, people watch movies/Tv shows became a common thing in the modern world. But choosing a movie to watch on the free time is the hardest part. So, after a long day, people decide to unwind by watching a movie on preferred streaming service. The excitement of watching a movie begins with the selection of the film. Because there are so many movies available on the market today, it might be tough to make a decision. This is due to the fact that the websites have a large selection of movies to choose from. There are so many movies to pick from that making a decision gets tough. You look at a few selections and watch a trailer. It appears to be okay, but perhaps a different film would be preferable. As a result, you watch another trailer, and then another, and another, and so on.

Main reason for this struggle is the Movie/TV show trailer. When we want a new movie, we usually watch the trailer before watching the movie. We all know that the whole aim of movie trailers is to persuade us to pay to watch a film. And, because all is fair in love and battle, movie trailers are free to put their best, most exciting foot forward, no matter how much recutting or remixing is required to get the job done. Often, the correct combination of footage and music can persuade us that even the most implausible film has something exceptional to offer Because the primary purpose of a film trailer is to advertise and promote the picture to a wide range of viewers, trailers are also a sort of persuasive art and promotional storytelling, aimed to entice you to watch the film in the theater or in home (Jerrcik,2013).

Because of this reason, watching trailer makes us to watch the movie as soon as possible. Sometimes the trailer we are watching is so entertaining and interesting. But when we pay the movie or stream the movie, movie is not worth. Movie is overhyped by the trailer. So out free time will be ruined.

Some people read reviews about the movie in social media and movie web platforms such as IMDB and Rotten Tomatoes. In social media and web review platforms, there will huge numbers of reviews. Reading that much of reviews will be hard thing to do and it will be a time-wasting process. Even though, people are not that much interested in reading reviews in a free time because all they need is to Watch a Good Movie or a TV show.

To avoid this situation, “IMDB Movie Review Analyzing system” has come up. This Web Application Helps User to get a Recommendation before watching the movie/TV Show. So, they will be no time Wasting and Choosing a movie for a very long-time struggle.

Aims and Objectives

- Training Dataset

  Training data is a form of data that is used to train a new application, model, or system using a variety of approaches, depending on the practicality and requirements of the project. For this project, dataset should have to contain text and labeled polarity. Identifying a suitable dataset will be an objective.

- Machine Learning classifiers

  There are numerous categorization algorithms available today, and it is impossible to determine one is superior to the others. It is dependent on the application and the nature of the data collection supplied. Objective is to identify a suitable machine learning algorithm to the project.

- Existing Systems and Research papers

  To develop a system, must do lot of research on Available systems and research papers. Objective is to do more research and identify available research papers and currently existing solutions.

- Web scraping Techniques

  The process of collecting data from the internet is known as web scraping. There are various web scarping techniques are available. Objective is to identify a suitable and fast web scarping technique which fits the system.

- Selecting a programming language

  The structure and meaning of programming languages, like human languages, are determined by syntactic and semantic rules. Programming languages are used to communicate about the challenge of organizing and manipulating data and to precisely define algorithms. Objective is to identify a suitable programming language to implement the system.

Project Plan and Deliverables

The process of determining the scope of a project, as well as the objectives and actions to achieve them, is known as project planning. One of the most crucial aspects of project management is the scheduling process. A project management plan is the result of the project planning process.

Because this project was not started in response to a customer or client request, the requirement definition is inaccessible in a traditional software development environment. The requirements definition serves as the foundation for the entire procedure. In light of this, the project will be subjected to a thorough investigation with the goal of "developing a solution which recommends user a particular movie/Tv show worth to watch or not". Requirements are detailed in the requirements specifications part of the document.

Project Plan is done using Gantt chart presentation. A Gantt chart is a bar chart that shows tasks scheduled across time in a visual format. A Gantt chart is a handy technique of indicating what work is scheduled to be done on a certain day, and it is used for project planning of all sizes. It is also useful to see a project's start and finish dates in a single graph.

Gantt Chart is attached in the appendix part of the document.

Chapter 02: Literature Review

Introduction

The origins of motion pictures on film, like many other inventions, are a mystery. Around the same period, several persons are thought to have devised what we now call movies. In any case, there were a number of key figures in the early days of cinema whose combined contributions aided in the creation of what we have today. Like many other inventions, the beginnings of motion pictures on film are unknown. Several people are thought to have invented what we now call movies around the same time. In any event, in the early days of cinema, there were a number of notable personalities whose combined contributions contributed in the construction of what we have today (Cassidy,NA).

In recent years, a rising number of people have chosen to spend their free time watching movies since they give amusement and relaxation. It is a wonderful trend in my opinion, as long as it's done in moderation. To begin with, many people believe viewing movies to be the ideal way to unwind after a long day at work. More specifically, most people these days have a demanding schedule and demanding work, and they look forward to finding ways to relax their mind and body. People believe Movies not only entertain them, but also help them forget about their troubles and stresses. People can become so busy and agitated in today's fast-paced world that they do not take enough time to enjoy moments and create memories. Watching movies is worthwhile for a variety of reasons, including entertainment, time spent with friends, and an easy way to reduce stress. People's motivation can be sparked by watching movies. The transformation of ordinary people into heroes in the story can inspire or urge people to achieve the same in their own lives. Aside from that, watching movies necessitates little physical or mental exertion, making them the preferred option given that the bulk of occupations nowadays are demanding and take a great deal of effort (Nambiar,2019).

Some People view a trailer on YouTube and think it is okay, but they think a different movie would be better. As a result, they watch yet another trailer, and then another, and so on. Movies and TV Shows are promoted using trailers. Trailers were created with the goal of generating enough curiosity in a film to entice people to see it when it was released. Most of the people before watching a movie or a Tv show, they usually watch the Trailer of it or read reviews about it in top web platforms like IMDB and Rotten Tomatoes. An analysis of the Movie or a TV show helps the spectator or anyone to know and appreciate the entire picture of the film. It benefits the audience too much to let them know if it really is worth watching the movie. Movie reviews are a widely used guide for audiences to understand whether a film is worth the admission price. Movie review talks about the perspective of the film's plot, the actors, their individual positions, history, scripts, and so on. That is why the viewers get a definite picture.

People watching movies/TV shows to relieve stress has become commonplace in the modern world. However, choose a movie to watch in your spare time is the most difficult task. As a result, individuals choose to unwind after a hard day by watching a movie on their chosen streaming service. The thrill of watching a movie begins with the film's choosing. It may be difficult to choose from the numerous films accessible today. This is owing to the high number of movies available on the websites. With so many films to choose from, making a decision can be difficult. You peruse a few options before watching a trailer.

With the motive of Promoting the movie, Movie Trailers shows the best scenes from the movie. Sometimes when people watch the movie, it does not worth the hype that trailer gave. So, when people watch a movie like that, even the free time they had also gets wasted. Sometimes trailer is not that good, but the movie is worth to watch.

Some individuals check out movie reviews on social media and on websites like IMDB and Rotten Tomatoes. There will be a large number of reviews on social media and web review platforms. It will be difficult to read so many reviews, and it will be a time-consuming procedure.

Domain Analysis

“IMDB Movie Review Analyzing System” will be the best solution to the issue which is discussed in the introduction part in the literature review. “IMDB movie Review Analyzing System” is a web application which will be easy to use because in the modern world, most of everyone has Internet connections. As other web applications, this will also need a Web Browser such as Google Chrome, Mozilla Firefox etc.

This part of the document will discuss about the available Movie Analyzing data sources. As well as this part will choose the Most reliable movie review data source which system will continue with.

Below are the Available methods which are people used to analyze a movie before watching it or purchasing it.

- YouTube Trailer comments and Movie reviews
  YouTube is the best social media network, with millions of users waiting in line to watch millions of videos every day. The production of a film is not enough to put money in the producer's pocket. The release of a picture, combined with the most effective promotion approach, will give the film new life and propel it to the top of the box office charts. The strategy of building a YouTube movie marketing campaign will allow for the creation of buzz. Only by narrating a spectacular plot in a teaser or trailer is it possible.
  The purpose of creating hype is to draw the audience's attention to the movie. It may have a better possibility of allowing the audience to talk about the film with their friends before it is released in theaters. The majority of film studios are using a new strategy, which is to collaborate with YouTube personalities known as YouTubers. Most films are now going viral as a result of the current trend of using YouTubers to promote them. They are a great way to catch people's attention and get them to buy movie tickets by breaking the suspense (RealnReel Team,2018).

- <u>IMDB Movie Reviews</u>
  IMDb (Internet Movie Database) is a massive database of movies, television shows, and video games. Its main function is to provide thorough information on any performer, producer, or piece of media material. Unlike the other sites, IMDb relies entirely on user reviews. Signing up for IMDb and leaving a review takes less than a minute, so there is little to no barrier to access. As a result, IMDb's greatest strength is that its ratings provide a solid indication of what ordinary people think of a film. IMDb scores are not influenced by professional critics (Stegner,2020). The IMDB has become one of the most popular online interaction forums. The establishment of the IMDB can be traced back to the late 1980s and early 1990s. Col Needham, the founder of IMDB, has created a website that seeks to offer valuable and up-to-date online film content across as many formats as possible (Donna,2015).

- <u>Rotten Tomatoes Movie Reviews</u>
  Rotten Tomatoes is a well-known website that gathers reviews from critics. To rate the quality of a film, the "Tomatometer" is used. A red tomato displays next to the critic's review if they like it. You will see a green splat instead if they do not like it. The popcorn bucket on Rotten Tomatoes also displays a user score. It displays a full bucket when at least 60% of users gave it a rating of 3.5 stars (out of 5) or above. A tipped-over bucket indicates that less than 60% of users gave it less than 3.5 stars (Stegner,2020).

With the methods discussed above, proposed system needs a data source to Web Scrape the reviews.

Some of the YouTube trailers does not allow to comment the opinions of the Viewers because some opinions will be Negative, and some will be positive. Reason is a straightforward question to answer. They are usually terrified of receiving negative comments and/or dislikes, and sometimes they deactivate the comments and/or rating because they post a video that tries to defraud people in various ways, and they are afraid that people will reveal the scam by disliking or commenting that it is a scam. So, for this reason YouTube comments will be not a valid data source.

The most serious flaw with Rotten Tomatoes is that it reduces nuanced opinions to a Yes or No rating. It gives the same score to a critic who thought the film was good but had flaws as one who thought it was terrible (Stegner,2020).

With the Availability of drawbacks, system will dismiss the YouTube Reviews, Rotten Tomatoes Reviews and System will choose IMDB Movie Reviews as the Data Source.

After selecting a Valid Data Source, now Deep research must be done to build the system workflow for the proposed Solution. “IMDB Movie Review Analyzing System” is based on Sentiment Analysis based Machine Learning technique. So, the research should be based on Sentiment Analysis based system research papers which will give an idea about how the system should be build and works. Below part will discuss about the research papers which used to study to develop this system. These research papers are so helpful to get a clear idea about the System workflow.

01.Sentiment Analysis of Hotel Reviews from Trip Advisor

Two students from ABES Engineering College in Ghaziabad, India, are conducting this research work. This is a System that Displays how much people satisfied with Hotels in Trip Advisor. As the outcome, System will visualize the Positive, Negative and Neutral Reviews in a Bar chart Plot view.

The System workflow of the system explained as below.

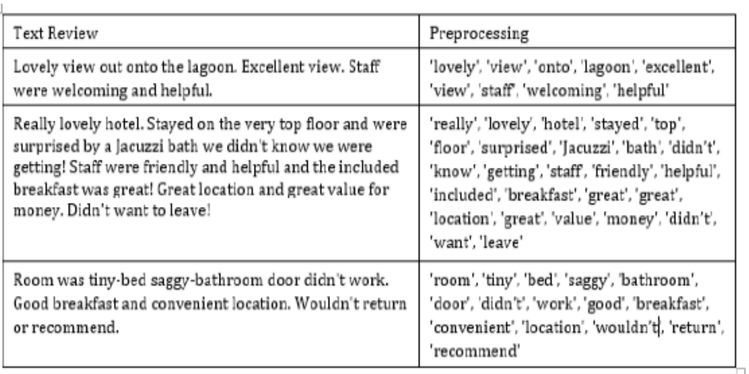
Data Collection → Data Preprocessing → Sentiment Detection → Sentiment Classification → Output

- Dataset – Reviews are Web Scraped from Trip Advisor website.to Scrape the Reviews, System have used BeautifulSoup Library. The Dataset has 738 rows of Records. Records have two columns. Columns are Review and Date.

- Data Preprocessing – Data preprocessing is a method for converting unclean data into a clean data set. In other words, anytime data is received from various sources, it is collected in raw format, which makes analysis impossible. The Steps of Data Cleaning which system has used, listed as below.

  Dataset → Case folding → Remove Punctuation → Remove Stop words → Lemmatization → Tokenization → Feature Extraction and Feature Selection.

  Data Preprocessing will Remove all the unwanted characters from each review and Then system will lemmatize the reviews. Lemmatization is the process of combining a word's several inflected forms into a single item for analysis. Stemming and lemmatization are both part of text preprocessing. Many individuals are perplexed by these two terms. Some people confuse the two. In fact, lemmatization is preferable than stemming since it does morphological analysis on the words (GeeksForGeeks,2018).
  Below figure displays the reviews after word lemmatization process.



Text Review	Preprocessing
Lovely view out onto the lagoon. Excellent view. Staff were welcoming and helpful.	'lovely', 'view', 'onto', 'lagoon', 'excellent', 'view', 'staff', 'welcoming', 'helpful'
Really lovely hotel. Stayed on the very top floor and were surprised by a Jacuzzi bath we didn't know we were getting! Staff were friendly and helpful and the included breakfast was great! Great location and great value for money. Didn't want to leave!	'really', 'lovely', 'hotel', 'stayed', 'top', 'floor', 'surprised', 'Jacuzzi', 'bath', 'didn't', 'know', 'getting', 'staff', 'friendly', 'helpful', 'included', 'breakfast', 'great', 'great', 'location', 'great', 'value', 'money', 'didn't', 'want', 'leave'
Room was tiny-bed saggy-bathroom door didn't work. Good breakfast and convenient location. Wouldn't return or recommend.	'room', 'tiny', 'bed', 'saggy', 'bathroom', 'door', 'didn't', 'work', 'good', 'breakfast', 'convenient', 'location', 'wouldn't', 'return', 'recommend'

Figure 1 Word Lemmatization Preview (Singh, V., Mahajan, A., and Chaudhary, D, 2020).

- Sentiment Detection – Sentiment Detection Process is done using TextBlob Library. TextBlob is a Python package for text processing. It offers a basic API for doing standard natural language processing (NLP) activities like part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, and translation, among others. After Extracting the Tokenized Features, System will Apply the TextBlob Library. After Applying System will give the Sentiment Polarity value for each Review.

- Sentiment Classification – After getting the Polarity Value, System can Classify each review into Sentiment Categories. If the Polarity score is Above 1, Then the review will be classified as a Positive Review. If the Polarity score is below 0, Then the review will be classified as a Negative Review. If the Polarity score is above 0 and below 1, Then the review will be classified as a Neutral Review.

- Used Classifiers – System uses Naïve Bayes Classifier as the default Classifier. The Bayes Theorem is used to create a collection of classification algorithms known as Naive Bayes classifiers. It is not a single algorithm, but rather a group of algorithms that all follow the same premise. Classifier is used to get the Prediction Accuracy Score (Singh, V., Mahajan, A., and Chaudhary, D, 2020).

02. Twitter Sentiment Analysis on Movie reviews using Machine Learning Techniques

This system is also very similar to the above discussed system. This research paper is conducted by students at School of Computer Science and Engineering college, India. This System will Display How People Satisfied with the Movies as the Outcome.

- Dataset - Dataset is created using Twitter Posts which are Based on Movie Reviews. This Dataset Contains 21,000 Records. Out of the 21,000 records, system will use 1200 Records for Training. This Training Data contains 600 Positive Reviews, 600 Negative Reviews and 600 Neutral Reviews.

Data Preprocessing - In the Data Cleaning Process System will,

- Turn the Tweet text into Lower case.
- Remove URL
- Target Name @username will be replaced as AT_USER.
- Remove Hashtags
- Replacing the Repeated Characters with two Occurrences
- Remove White Spaces.

Vectorizing the Cleaned tweets - In this Process system will vectorize the Reviews into Numerical values. So, the Machine Learning classifiers can understand the tweets properly.

- Classifiers used – Machine Learning Classifiers the system using are Naïve Bayes, Support Vector Machine. Naïve Bayes Classifier is already discussed in the above System. Support Vector Machine (SVM) is a non-probabilistic, linear, binary classifier used in machine learning to categorize data by learning a hyperplane that separates the data (GeeksForGeeks,2019). Documents will further discuss about SVM in the upcoming chapters.

Accuracy Score of Naïve Bayes classifier is 65% and Accuracy Score of SVM is 75%. So, As Conclusion SVM Classifiers give the highest Accuracy Score, and it is the most suitable Classifier for this module (Baid, p., Chaplot, N., and Gupta, A, 2017).

03. Sentiment Analysis of Movie Reviews Using Machine Learning Techniques

This system is likewise extremely similar to the one mentioned previously. Students from the School of Computer Science and Engineering college in India conducted this study. As same as the Above discussed Systems, this projects scope is also to give a conclusion about the Movies using Movie Reviews by Visualizing the Positive, Negative and Neutral Reviews.

Dataset - System will use a CSV formatted Dataset which will contains Movie Reviews.

Data Preprocessing – In this Process System will remove Stop words and makes the words more meaningful using Porters Stemming Method.

Train and Test Splitting – System will split the Cleaned Data in to Training and Test data in a Ratio of 80:20. This means 80% of the Data will be used for Training and 20% of the Data will used for Testing.

- Training Data - The data that will be input into the model will be stored in the train set. A Regression model, for example, might utilize the instances in this data to locate gradients in order to lower the cost function. These gradients will then be used to cut costs and accurately anticipate data (GeeksForGeeks,2020).

- Testing Data - The data used to test the trained and validated model is contained in the test set. It indicates how effective our overall model is and how likely it is to anticipate something that is illogical. There are numerous assessment criteria (such as precision, recall, and accuracy) that may be used to assess our model's performance (GeeksForGeeks,2020).

Classifiers – System has used Three classifiers. There are K-Nearest Neighbor, Multinomial Naïve Bayes, and Logistic Regression. Accuracy Scores of Each Classifier are listed below.

- K- Nearest Neighbor – 98.6994%
- Multinomial Naïve Bayes – 98.9161%
- Logistic Regression – 99.3497%

In Conclusion, with an Accuracy score of 99%, Logistic Regression is the best and suitable classifier for this module. So, system will use Logistic Regression Classifier for the prediction (Amolik, A. et al, 2020).

04. A Sentiment Analysis of Food Review using Logistic Regression.

This system is likewise extremely similar to the one mentioned previously. Scope of this project is to predict the Food reviews which are posted online in to Positive, Negative and Neutral. In collusion user can get a conclusion about which food people most like and which food people do not like.

Below workflow will explain how the system works.

Review Data → Preprocessing → Feature extraction→ Classification → Result Interpretation

Dataset - User Food Reviews is a massive dataset that includes over 568454 surveys reviver food products written by commentators between 1999 and 2012. Every survey includes ten parameters: Id, Product Id, User Id, Profile Name, Helpfulness Numerator, Helpfulness Denominator, Score, Time Summary, and Text. Scores range from 1 to 5, with 1 being the highest and 5 being the lowest. As a result, all audits with a score of 3 are considered neutral, those with a score of less than 3 are considered negative, and those with a score of more than 3 are considered positive. 73% of the reviews are positive 33% of the reviews are Negative.

Data Preprocessing – First system removes the stop words from the dataset. Tokenization: In these systems, each string is assigned an integer token id. Counting: The amount of tokens in this system is primarily counted. Normalization: The tokens that appear frequently are given more weight in this scheme.

System will vectorize the cleaned reviews and creates a bag of words model.

Classifiers – System used three classifiers. They are Linear regression, perceptron, and Bernoulli Naïve Bayes classifiers. Accuracy Scores of Each Classifier are listed below.

- Linear regression – 89%
- Perceptron – 86%
- Bernoulli Naïve Bayes – 87%

Linear Regression has the highest accuracy score among the three classifiers. So, system will Linear Regression classifier as the default Machine learning classifier (Mamtesh. And Mehla, S, 2020).

The techniques which have been identified in the Similar system research papers are discussed in the below part of the document.

Technical Research and Analysis

This part will discuss about the technical aspects of the identified Techniques to develop this system.

Sentiment Analysis Based Machine Learning

Sentiment analysis is a type of machine learning that examines texts for polarity, ranging from positive to negative. Machine learning systems are taught how to detect sentiment without human input by providing them with examples of emotions in text. Computers can automatically process text input and analyze it using sentiment analysis, exactly like a person (Wolff,2020).

Machine learning is a branch of artificial intelligence (AI) that focuses on designing applications that learn from and increase data quality over time without being programmed to do so. In data science, the algorithm is a series of steps for predictive analysis. In machine learning, algorithms are 'trained' to identify patterns and characteristics in vast volumes of data in order to render new data-based decisions and predictions. The higher the algorithm, the more accurate the decisions and forecasts can become, as more data is processed (IBM,2020).

Supervised machine learning is trained on a labeled data collection. That is, the data is branded with knowledge that the machine learning algorithm is being designed to evaluate and that can also be categorized in ways that the model is intended to identify the data. Supervised machine learning needs fewer training data than other machine learning techniques and makes training simpler when the outputs of the algorithm can be correlated to the real results labelled. But correctly labelled data is costly to prepare, and there is a danger of over-fitting or having a model so tightly linked and skewed to training data that it does not treat changes in new data precisely (IBM,2020).

Machine Learning Classifiers

The process of classifying data points entails anticipating their class. Targets, labels, and categories are terms used to describe classes. Approximating a mapping function (f) from discrete input variables (X) to discrete output variables is the work of classification predictive modeling (y). There are numerous classification algorithms available today, however it is impossible to determine which is superior. It is determined by the application and the nature of the data set accessible. This part of the document will explains most commonly using classifiers in machine learning processes.

- <u>Decision Tree</u>

  In the shape of a tree structure, a decision tree constructs classification or regression models. For classification, it employs an if-then rule set that is mutually exclusive and exhaustive. The rules are learned one by one, one by one, utilizing the training data. The tuples covered by the rules are eliminated each time a rule is learned. On the training set, this process is repeated until a termination condition is met. The tree is built in a recursive divide-and-conquer fashion from the top down. All of the characteristics should be categorical in nature. Otherwise, they should be separated ahead of time. The information gain concept is used to identify attributes at the top of the tree that have a greater impact on classification. A decision tree can easily be over-fitted, resulting in an excessive number of branches, which can reveal anomalies due to noise or outliers. The performance of an over-fitted model on unseen data is bad, despite its great performance on training data. Pre-pruning, which stops tree growth early, or post-pruning, which removes branches from a fully grown tree, can help to avoid this (Asiri,218).

- <u>Naïve Bayes</u>

The Bayes theorem inspired Naive Bayes, a probabilistic classifier that works under the basic assumption that the qualities are conditionally independent.

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i)$$

Figure 2 Naive bayes Classifier Formula (Asiri,2018)

With the above assumption applied to Bayes theory, the classification is done by obtaining the maximum posterior, which is the maximal P(Ci|X). By merely counting the class distribution, this assumption drastically minimizes the computational cost. Even though the assumption is not valid in most circumstances because the qualities are dependent, Naive Bayes has been able to perform well (Asiri,2018).

Naive Bayes is a very simple algorithm to develop, and it has produced good results in the majority of applications. Because it requires linear time rather than the expensive iterative approximation employed by many other types of classifiers, it can quickly scale to larger datasets. The zero-probability problem can be a difficulty with naive Bayes. When the conditional probability for a given property is zero, the prediction is invalid. Using a Laplacian estimator, this must be addressed explicitly (Asiri,2018).

- <u>Artificial Neural Networks (ANN)</u>

In order to construct and evaluate computational analogs of neurons, psychologists and neurobiologists created Artificial Neural Networks, which are a series of connected input/output units with each link having a weight associated with it. The network learns by modifying the weights during the learning phase so that it can anticipate the right class label of the input tuples. Feed-forward, Convolutional, Recurrent, and more network designs are now available. The model's application determines the proper architecture. In most circumstances, feed-forward models produce reasonably accurate results, however convolutional networks outperform feed-forward models in image processing applications (Asiri, 2018).

Depending on the complexity of the function to be mapped by the model, there may be numerous hidden layers in the model. Modeling complex relationships, such as deep neural networks, will be easier with more hidden layers. However, training and adjusting weights takes a long time when there are several hidden layers. The model's interpretability is also weak when compared to other models such as Decision Trees, due to the unknown symbolic meaning behind the acquired weights. Artificial Neural Networks, on the other hand, have done admirably in the majority of real-world applications. It has a high tolerance for noisy input and can classify patterns that have not been trained. Artificial Neural Networks usually function better when the inputs and outputs are continuous (Asiri,2018).

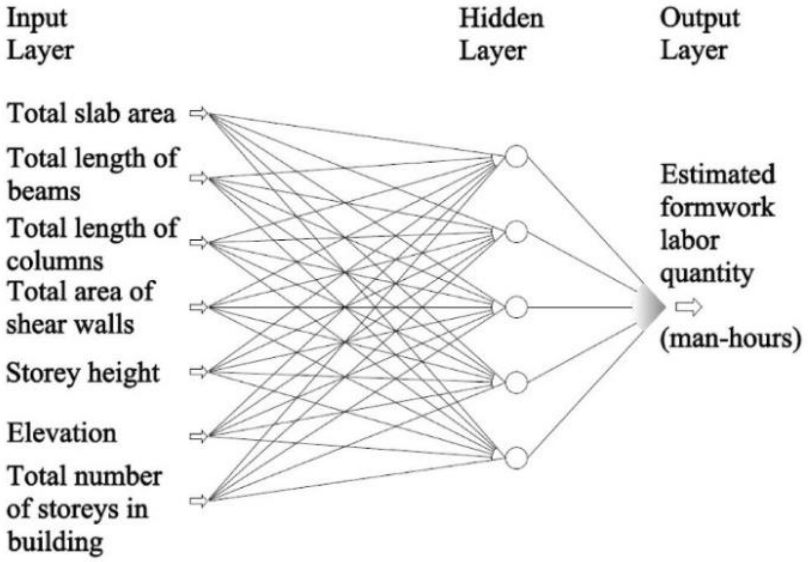


Figure 3 Artificial Neural Network (Asiri,2018)

- <u>k-Nearest Neighbor (KNN)</u>

The k-Nearest Neighbor algorithm is a lazy learning technique that stores all instances in n-dimensional space that correspond to training data points. When an unknown discrete data is received, it examines the nearest k number of saved instances (nearest neighbors) and returns the most common class as the prediction, whereas real-valued data returns the mean of k nearest neighbors (Asiri,2018).

The distance-weighted nearest neighbor method uses the following query to weight the contributions of each of the k neighbors based on their distance, providing greater weight to the closest neighbors. Because it averages the k-nearest neighbors, KNN is usually resistant to noisy data (Asiri,2018).

- Linear SVC

Support vector machines (SVMs) are supervised machine learning methods for classification, regression, and identification of outliers that are both powerful and adaptable. SVMs are commonly employed in classification tasks because they are particularly efficient in high dimensional spaces. Because SVMs use a subset of training points in the decision function, they are popular and memory efficient. SVMs' main purpose is to partition datasets into a large number of classes in order to discover a maximum marginal hyperplane (MMH), which can be done in two steps:

1. Support Vector Machines will initially iteratively build hyperplanes that best distinguish the classes.

2. After that, it will select the hyperplane that best separates the classes.

SVM has three classes that can do multiclass-class classification: SVC, NuSVC, and LinearSVC. Linear Support Vector Classification is what it is called. It is analogous to having kernel ='linear' in SVC. The distinction between the two is that LinearSVC is written in liblinear, whereas SVC is written in libsvm. That is why LinearSVC gives you more options when it comes to penalties and loss functions. It also handles a larger number of samples better (Asiri,2018).

- Logistic Regression

In the early twentieth century, the biological sciences adopted logistic regression. It went on to be employed in a variety of social science applications. When the dependent variable is categorical, logistic regression is utilized. A binary outcome, a positive or negative conclusion, is predicted by a logistic regression algorithm: Yes/No, Existence/Non-existence, Pass/Fail. It simply means that something occurs or does not occur. To determine the 0/1 outcome (one of two categories), variables are compared to one another (Mesevage,2020).

Although the independent factors can be numeric or categorical, the dependent variable is always categorical: the likelihood of dependent variable Y given independent variable X. This can be used to discern the object in a photo or video image by assigning a probability between 0 and 1 to each object, or to compute the likelihood that a phrase has a good or negative meaning (Mesevage,2020).

Text Vectorization

The technique of translating text into numerical form is known as text vectorization. Here are a few popular ways for text vectorization:

- <u>Bag of Words (BoW) Term Frequency</u>

  Bag of Words (BoW) Term Frequency captures frequency of term in document. Text modeling using a bag of words is a technique used in Natural Language Processing. It is a method of feature extraction with text data, to put it another way. This method of extracting features from documents is straightforward and adaptable. A bag of words is a text representation that describes the frequency with which words appear in a document. We only keep track of word counts and do not pay attention to grammatical subtleties or word arrangement. Because any information about the sequence or structure of words in the document is deleted, it is referred to as a "bag" of words. The model simply cares about whether or not recognized terms appear in the document, not where they appear. (Chen,2020).

- <u>TF-IDF</u>

  The TF-IDF is a statistical measure that assesses the relevance of a word to a document in a set of documents. This is accomplished by multiplying two metrics: the number of times a word appears in a document and the word's inverse document frequency over a collection of documents. It has a variety of applications, including automatic text analysis and scoring words in machine learning techniques for Natural Language Processing (Chen,2020). For document search and retrieval, the TF-IDF (term frequency-inverse document frequency) algorithm was developed. It works by growing in proportion to the number of times a word appears in a document but offset by the number of papers containing the word. As a result, words like this, what, and if, which appear frequently in all documents, rank low since they do not mean much to that document (Chen,2020).

- <u>Word2Vec</u>

  Word2Vec is a program that converts words into vectors. Word2Vec starts with a single representation of all words in the corpus and uses a big corpus of data to train a NN (with one hidden layer). The two most common approaches for training the NN are as follows:

  1. Continuous Bag of Words (CBOW) – Use a window of context words to predict the vector representation of the center/target word.
  2. Skip-Gram (SG) — Based on the center/target word, predict the vector representation of a window of context words (Chen,2020).

Web Scraping

Web scraping (also known as screen scraping, web data extraction, web harvesting, and so on) is a method of extracting huge amounts of data from websites and saving it to a local file on computer or to a database in table (spreadsheet) format. Web Scraping is the practice of collecting content from the Internet Most websites' data can only be viewed using a web browser. They do not allow to save a copy of this information for personal use. The only other alternative is to manually copy and paste the data, which is a time-consuming task that might take hours or even days to complete. Web scraping is a method of automating this process so that, instead of manually downloading data from websites, the Web Scraping software may do so in a fraction of the time. The terms "internet scraping" normally refer to a method involving automation. Some websites do not like it when automated scrapers collect their results, while others do not mind it. There are two types of web scarping methods (Perez,2019).

- Using software

  There are two types of web scraping software. The first can be installed locally on the computer, and the second is cloud-based and browser-based. Web scraping applications such as WebHarvy, OutWit Hub, Visual Web Ripper, and others can be installed on a PC, although cloud data extraction services such as import.io, Mozenda, and others are available online (Perez,2019).

- Implementing code

  can employ a programmer to create custom data extraction software to meet a specific need. Web scraping APIs can then be used by the developer to make the product easier to develop. Apify.com, for example, makes it simple to obtain APIs for scraping data from any website (Perez,2019).

Text preprocessing

We normally think of massive datasets with a big number of rows and columns when we talk about data. While this is a likely scenario, it is not necessarily the case because data can take numerous forms: Tables with structure, images, audio files, and videos, for example. Machines do not comprehend free text, image, or video data; instead, they comprehend 1s and 0s. So, putting on a slideshow of all our photographs and expecting our machine learning model to learn from it is probably not going to work.

Data Preprocessing is the step in any Machine Learning process in which the data is changed, or encoded, to make it easier for the machine to parse it. In other words, the algorithm can now easily interpret the data's features (Pandey,2019).

Web Application Development

The benefit of Web Applications is that they are autonomous platforms and can be run by anybody who has access to the Internet. Their programming is installed on a back-end server where the application handles incoming requests and responds to a shared protocol that is recognized by all browsers.

Chapter 03: Requirement Specifications

Functional Requirements

Load the Pre-Trained Model and TF-IDF Vectorizer

System will load the pre trained Linear SVC model which will predict the Reviews into Positive and Negative. And also, system will load the TF-IDF Vectorizer to vectorize the New Reviews.

Copy and Pasting the Movie Link

Users must Copy the Movie review link from IMDB Site and Paste it in the Area which is provided in the System.

Viewing Instructions

User can view the instructions page to get to know about how to copy the link in proper way.

Viewing Video Demo

User can view a quick video demo of how to copy the link properly.

Scrapes the User Reviews of Selected Movie from IMDB

System will scrape the User Reviews from the IMDB Movie Review link which User has given.

Movie Analysis Using User Reviews

System will analyze the User Reviews and Display a report about the movie including percentage of Positive and Negative Reviews with Bar chart Presentation and Movie Recommendation Status.

Non – Functional Requirements

- User Friendly GUI – System will have a User friendly and eye-catching Web Application.
- Reliability - Reliability is the degree to which the specified functions are continuously executed by the software system without loss.
- Maintainability - Maintainability is the simplicity with which bugs can be detected and corrected in the software code.
- Usability – user will be a very easy one for use. System will have an instruction page where every step will be displayed.
- Modifiability - Modifiability is the degree to which improvements to a software framework can be developed and executed reliably and cost-effectively.

Chapter 04: Methodology

Research Methods

Web Scraping

Web Scraping is the practice of extracting information from the Internet. The terms "Web Scraping" typically apply to a method requiring automation. Some websites do not like it when automated scrapers extract their data, while others do not mind it.

Dataset

The data set is a data array. In other words, the data set refers to the contents of a single database table, or a single statistical data matrix, where each column in the table represents a certain element, and each row corresponds to the data set in question. The system has a training data collection in Machine Learning projects. It is the exact data collection used to teach the algorithm to execute different behaviors. Machine Learning is highly dependent on Data (Gonfalonieri,2019). To Train the Model, dataset is collected from Kaggle. This Dataset has 50000 rows and two columns. The Two Columns are Review and Sentiment.

Data Cleaning and Preprocessing

Data preprocessing plays a major role in a number of supervised learning algorithms. The quality of data and the valuable knowledge that can be learned from it directly affects the model's ability to learn, so data preprocessing is an important step in Machine Learning.

TD-IDF vectorizer

TF-IDF is called "Term Frequency, Inverse Document Frequency." It is a way to measure the value of words in a text depending on how much they appear in different documents. TF-IDF allows one to equate each word in a document with a number that reflects the importance of each word in that document. TF-IDF represents the relative importance of words in a text depending on a collection of documents (Gupta,2019).

Machine Learning Algorithms

For the classification of film reviews, the system applies machine learning approaches to textual reviews. Various text classifiers using machine learning methods have been proposed, such as Naïve Bayes and Linear SVC.

Sentiment Analysis based Machine Learning.

Sentiment analysis is a type of machine learning that examines texts for polarity, ranging from positive to negative. Machine learning systems are taught how to detect sentiment without human input by providing them with examples of emotions in text. Computers can automatically process text input and analyze it using sentiment analysis, exactly like a person (Wolff,2020).

Development Platform

Since the proposed system is a Web Application, it needs to have a developing language, a proper IDE and a framework to develop.

- Programming language
  The ways for controlling particular components of the software or machine are known as commands. Classes and functions are used in programming languages to control commands. The importance of programming stems from the fact that it instructs a computer to repeat these commands so that individuals do not have to repeat the process. Instead, the software can perform it for you in an automated and precise manner (Singh,2021).

- IDE
  An integrated development environment (IDE) is a program that aids in the creation of applications. IDEs are programs that combine all programming tasks into a single application. As a result, IDEs provide a centralized interface with all of the tools a developer requires. Improved developer productivity is the overall goal and primary benefit of an integrated development environment. By minimizing setup time, boosting the speed of development tasks, keeping developers up to date, and standardizing the development process, IDEs increase productivity (Singh,2021).

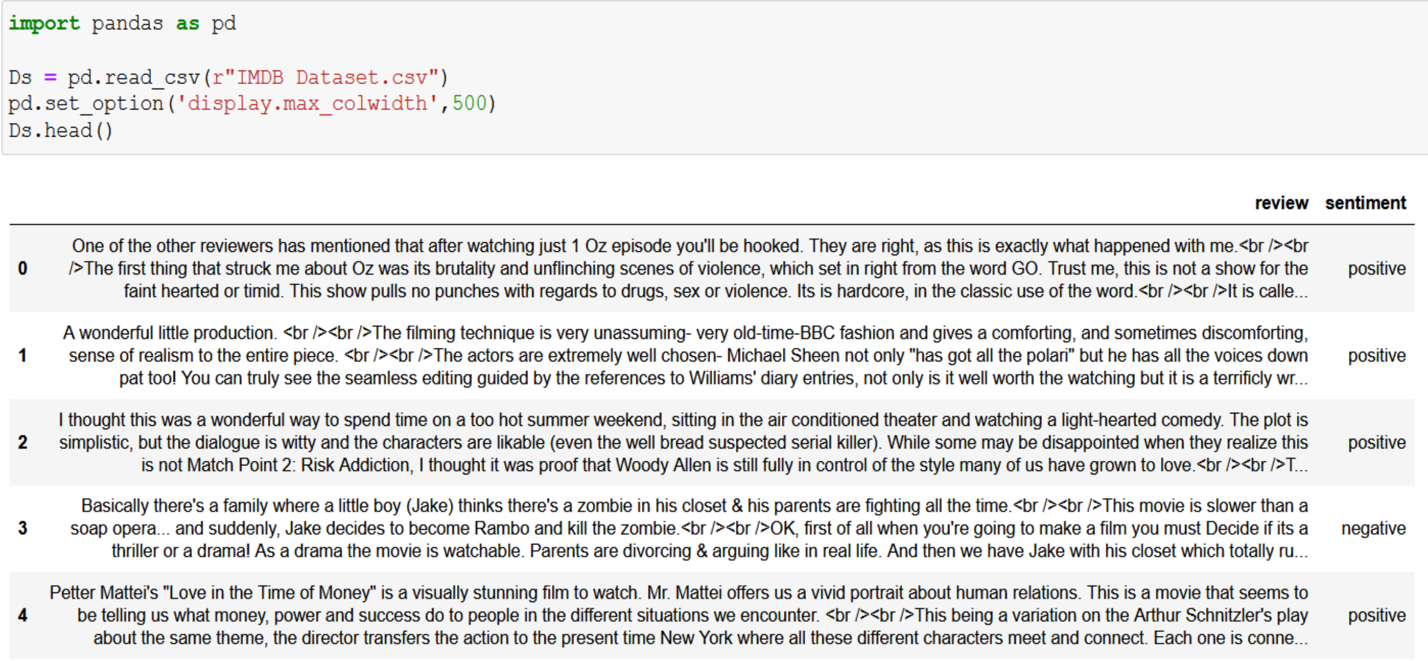
- Framework
  Frameworks are pieces of software that developers create and utilize to create applications. Developing apps with a software framework allows you to concentrate on the application's high-level functionality. This is because the framework handles all of the low-level functionality (Singh,2021).

So, in the next part of the document will explain all the techniques which are used to develop the System.

Application of Methods

Dataset

To train the Machine Learning Model, System uses a Dataset called “IMDB Dataset” which is collected from Kaggle Website. This Dataset has 50,000 record which are labeled. The dataset contains two fields which are Review and Sentiment Type. Review field contains the reviews and sentiment field contains whether the Review is Positive review or Negative Review. The Reviews, which are in the dataset, are not preprocessed and Reviews are stored as raw texts.

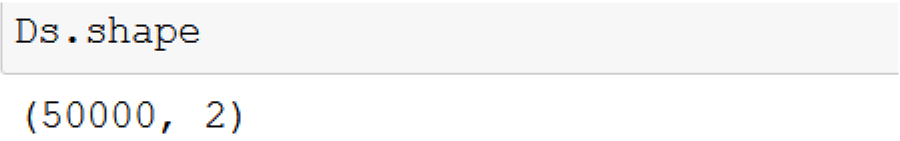


```python
import pandas as pd

Ds = pd.read_csv(r"IMDB Dataset.csv")
pd.set_option('display.max_colwidth',500)
Ds.head()
```

	review	sentiment
0	One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happened with me.<br /><br />The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in the classic use of the word.<br /><br />It is calle...	positive
1	A wonderful little production. <br /><br />The filming technique is very unassuming- very old-time-BBC fashion and gives a comforting, and sometimes discomforting, sense of realism to the entire piece. <br /><br />The actors are extremely well chosen- Michael Sheen not only "has got all the polari" but he has all the voices down pat too! You can truly see the seamless editing guided by the references to Williams' diary entries, not only is it well worth the watching but it is a terrificly wr...	positive
2	I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air conditioned theater and watching a light-hearted comedy. The plot is simplistic, but the dialogue is witty and the characters are likable (even the well bread suspected serial killer). While some may be disappointed when they realize this is not Match Point 2: Risk Addiction, I thought it was proof that Woody Allen is still fully in control of the style many of us have grown to love.<br /><br />T...	positive
3	Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his parents are fighting all the time.<br /><br />This movie is slower than a soap opera... and suddenly, Jake decides to become Rambo and kill the zombie.<br /><br />OK, first of all when you're going to make a film you must Decide if its a thriller or a drama! As a drama the movie is watchable. Parents are divorcing & arguing like in real life. And then we have Jake with his closet which totally ru...	negative
4	Petter Mattei's "Love in the Time of Money" is a visually stunning film to watch. Mr. Mattei offers us a vivid portrait about human relations. This is a movie that seems to be telling us what money, power and success do to people in the different situations we encounter. <br /><br />This being a variation on the Arthur Schnitzler's play about the same theme, the director transfers the action to the present time New York where all these different characters meet and connect. Each one is conne...	positive

Figure 4 Training Dataset Preview

Shape of the Dataset is shown in the below figure.



```
Ds.shape

(50000, 2)
```

Figure 5 Training Dataset Shape

This figure explains how many records are in the Dataset. There are 50,000 rows and 2 columns in the dataset.

This Dataset is a Well-Balanced Dataset which have Same Number of Positive Reviews and Negative Reviews. 50000 of dataset records has divided in to 25000 Positive Reviews and 25000 Negative Reviews. The Below figure explains that the dataset is well balanced.

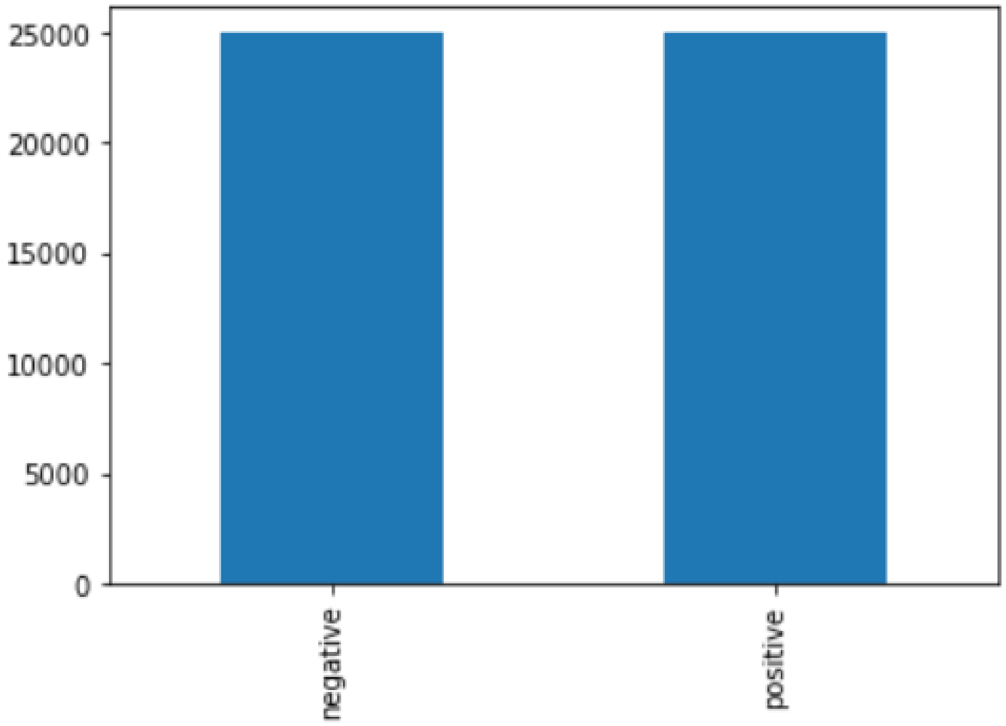
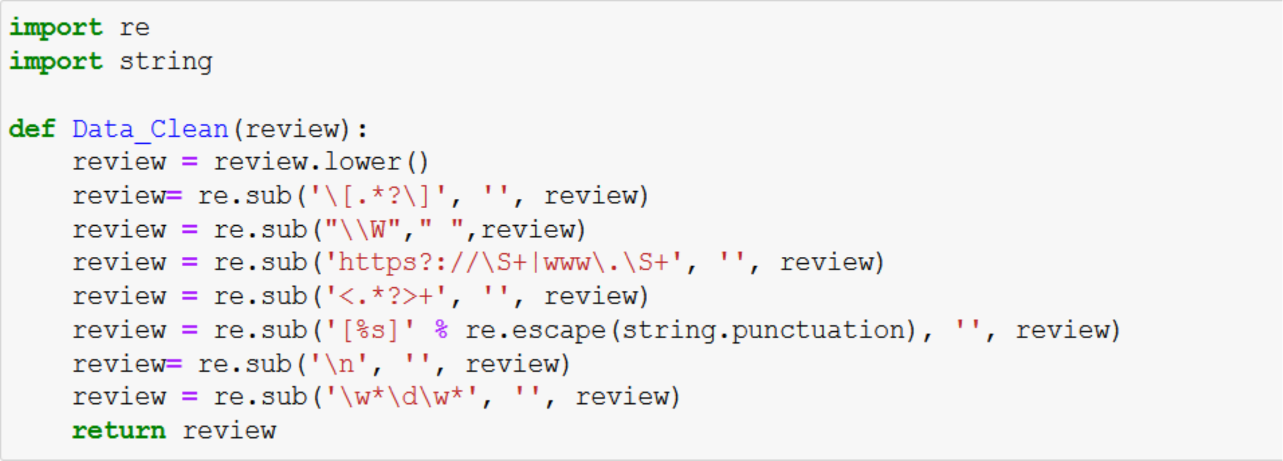


Figure 6 Reviews Count Bar Chart

Data Preprocessing

The quality of data and the valuable knowledge that can be learned from it directly affects the model's ability to learn, so data preprocessing is an important step in Machine Learning. The Review in the collected dataset is in raw format. So, system need to clean the reviews before training the model. The Main Purpose of Data Preprocessing is to remove the unwanted characters from the reviews. So, the Trained Model can learn proper meaning of the review. After cleaning the Unwanted characters, there will be only important characters which gives a proper meaning.



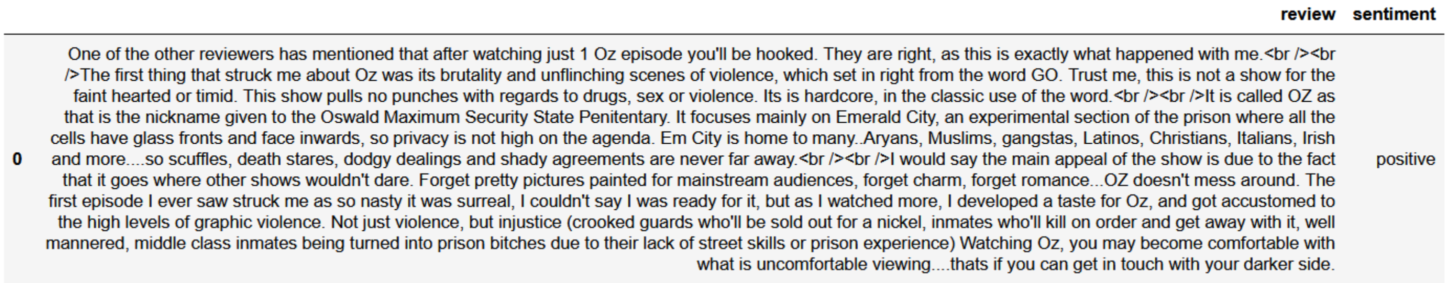
```
import re
import string

def Data_Clean(review):
    review = review.lower()
    review= re.sub('\[.*?\]', '', review)
    review = re.sub("\\W"," ",review)
    review = re.sub('https?://\S+|www\.\S+', '', review)
    review = re.sub('<.*?>+', '', review)
    review = re.sub('[%s]' % re.escape(string.punctuation), '', review)
    review= re.sub('\n', '', review)
    review = re.sub('\w*\d\w*', '', review)
    return review
```

Figure 7 Data Preprocessing Method Preview

This is the method System uses to Clean the Reviews which are in the Dataset. Each Review will have to go through these Steps. When Reviews are going through this method, System will turn reviews in to Lower case and Removes arrays Lists, Special Characters, Web links, punctuations, and numbers. Then then method will reformat the reviews by matching the white spaces which were created when cleaning the reviews.

According to the Above figure, System has Cleaned the unwanted characters and left the meaningful and wanted text for machine learning process. For Further clarification, the below figures compare the reviews before and after data preprocessing.

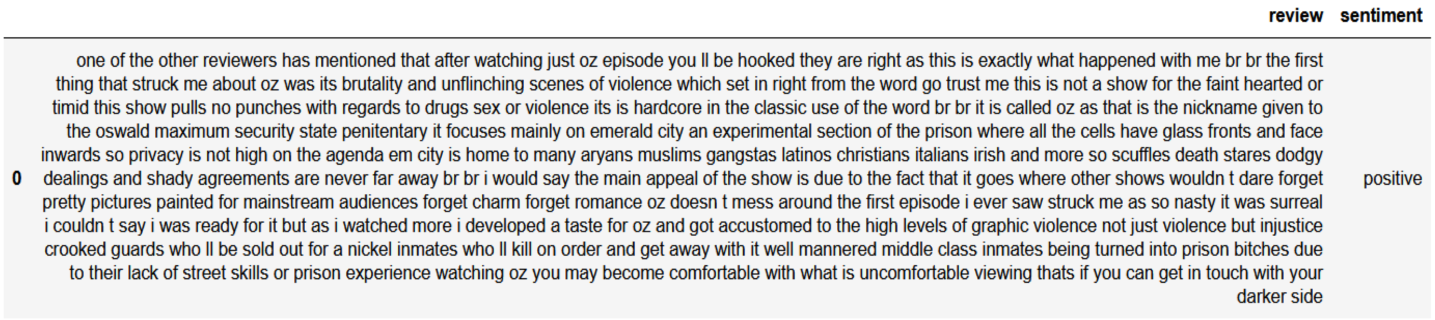
Before Data Preprocessing



	review	sentiment
0	One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happened with me.<br /><br />The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in the classic use of the word.<br /><br />It is called OZ as that is the nickname given to the Oswald Maximum Security State Penitentary. It focuses mainly on Emerald City, an experimental section of the prison where all the cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many..Aryans, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings and shady agreements are never far away.<br /><br />I would say the main appeal of the show is due to the fact that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget charm, forget romance...OZ doesn't mess around. The first episode I ever saw struck me as so nasty it was surreal, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to the high levels of graphic violence. Not just violence, but injustice (crooked guards who'll be sold out for a nickel, inmates who'll kill on order and get away with it, well mannered, middle class inmates being turned into prison bitches due to their lack of street skills or prison experience) Watching Oz, you may become comfortable with what is uncomfortable viewing....thats if you can get in touch with your darker side.	positive

Figure 8 Reviews Before Data Pre-Processing

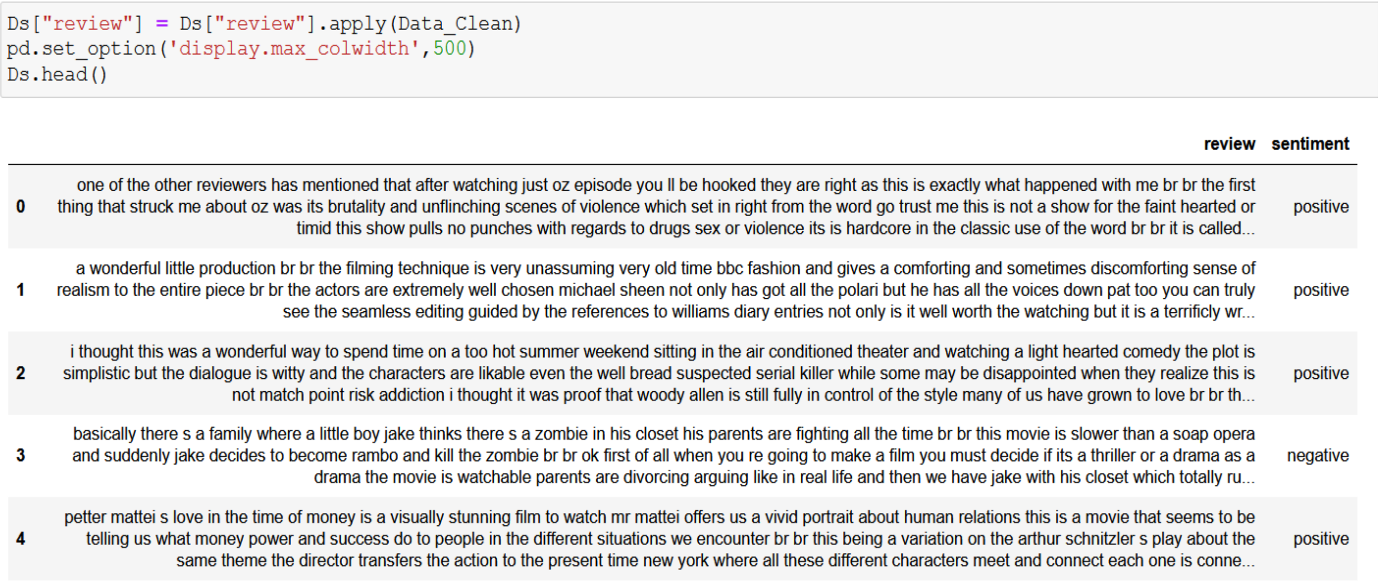
After Data Preprocessing



	review	sentiment
0	one of the other reviewers has mentioned that after watching just oz episode you ll be hooked they are right as this is exactly what happened with me br br the first thing that struck me about oz was its brutality and unflinching scenes of violence which set in right from the word go trust me this is not a show for the faint hearted or timid this show pulls no punches with regards to drugs sex or violence its is hardcore in the classic use of the word br br it is called oz as that is the nickname given to the oswald maximum security state penitentary it focuses mainly on emerald city an experimental section of the prison where all the cells have glass fronts and face inwards so privacy is not high on the agenda em city is home to many aryans muslims gangstas latinos christians italians irish and more so scuffles death stares dodgy dealings and shady agreements are never far away br br i would say the main appeal of the show is due to the fact that it goes where other shows wouldn t dare forget pretty pictures painted for mainstream audiences forget charm forget romance oz doesn t mess around the first episode i ever saw struck me as so nasty it was surreal i couldn t say i was ready for it but as i watched more i developed a taste for oz and got accustomed to the high levels of graphic violence not just violence but injustice crooked guards who ll be sold out for a nickel inmates who ll kill on order and get away with it well mannered middle class inmates being turned into prison bitches due to their lack of street skills or prison experience watching oz you may become comfortable with what is uncomfortable viewing thats if you can get in touch with your darker side	positive

Figure 9 Reviews After Data Pre-Processing

This Figures explains that System has removed all the unwanted characters from review and made the reviews Meaningful.



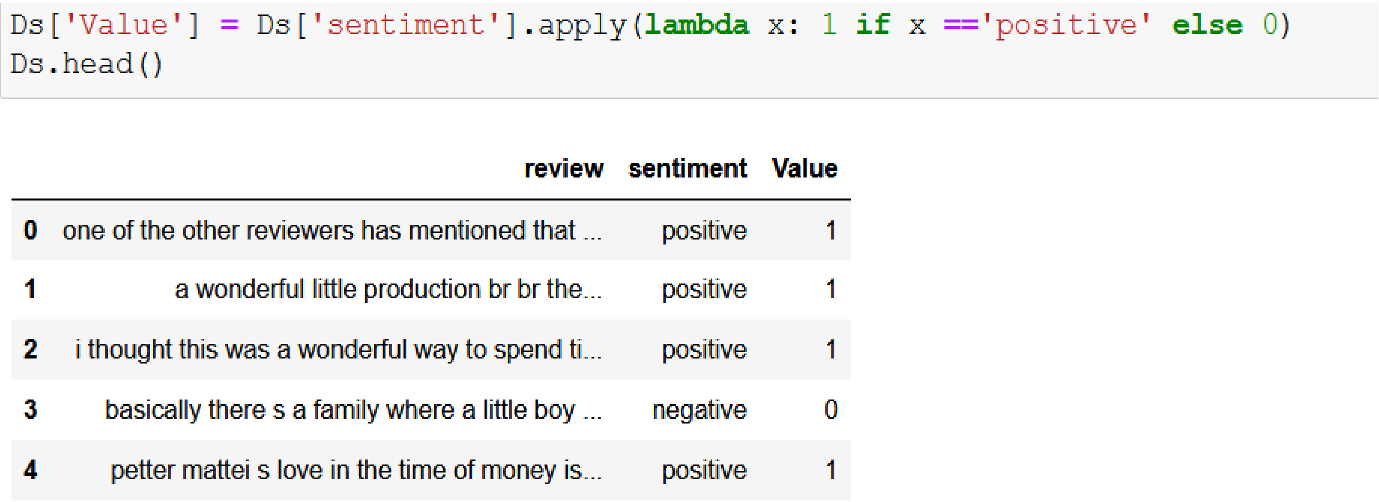
```
Ds["review"] = Ds["review"].apply(Data_Clean)
pd.set_option('display.max_colwidth',500)
Ds.head()
```

	review	sentiment
0	one of the other reviewers has mentioned that after watching just oz episode you ll be hooked they are right as this is exactly what happened with me br br the first thing that struck me about oz was its brutality and unflinching scenes of violence which set in right from the word go trust me this is not a show for the faint hearted or timid this show pulls no punches with regards to drugs sex or violence its is hardcore in the classic use of the word br br it is called...	positive
1	a wonderful little production br br the filming technique is very unassuming very old time bbc fashion and gives a comforting and sometimes discomforting sense of realism to the entire piece br br the actors are extremely well chosen michael sheen not only has got all the polari but he has all the voices down pat too you can truly see the seamless editing guided by the references to williams diary entries not only is it well worth the watching but it is a terrificly wr...	positive
2	i thought this was a wonderful way to spend time on a too hot summer weekend sitting in the air conditioned theater and watching a light hearted comedy the plot is simplistic but the dialogue is witty and the characters are likable even the well bread suspected serial killer while some may be disappointed when they realize this is not match point risk addiction i thought it was proof that woody allen is still fully in control of the style many of us have grown to love br br th...	positive
3	basically there s a family where a little boy jake thinks there s a zombie in his closet his parents are fighting all the time br br this movie is slower than a soap opera and suddenly jake decides to become rambo and kill the zombie br br ok first of all when you re going to make a film you must decide if its a thriller or a drama as a drama the movie is watchable parents are divorcing arguing like in real life and then we have jake with his closet which totally ru...	negative
4	petter mattei s love in the time of money is a visually stunning film to watch mr mattei offers us a vivid portrait about human relations this is a movie that seems to be telling us what money power and success do to people in the different situations we encounter br br this being a variation on the arthur schnitzler s play about the same theme the director transfers the action to the present time new york where all these different characters meet and connect each one is conne...	positive

Figure 10 Applying Data Pre-Processing Method and Outcome Preview

Numerical Values for Sentiment Types

Machine learning Algorithms deals with numerical values. So, system needs the independent variable to continue the machine learning process. The input for a process under investigation is independent variables. In this case independent variable will be the sentiment type. Dataset has a separate column called **"Value"** which contains the sentiment type of each review. But sentiment types in text format and machine learning only understand numerical format. For that system must convert the positive statement in to 1 and negative statement in to 0. So, machine learning can understand the process. Also, the Sentiment Type will the Independent Variable of this model.



```
Ds['Value'] = Ds['sentiment'].apply(lambda x: 1 if x =='positive' else 0)
Ds.head()
```

	review	sentiment	Value
0	one of the other reviewers has mentioned that ...	positive	1
1	a wonderful little production br br the...	positive	1
2	i thought this was a wonderful way to spend ti...	positive	1
3	basically there s a family where a little boy ...	negative	0
4	petter mattei s love in the time of money is...	positive	1

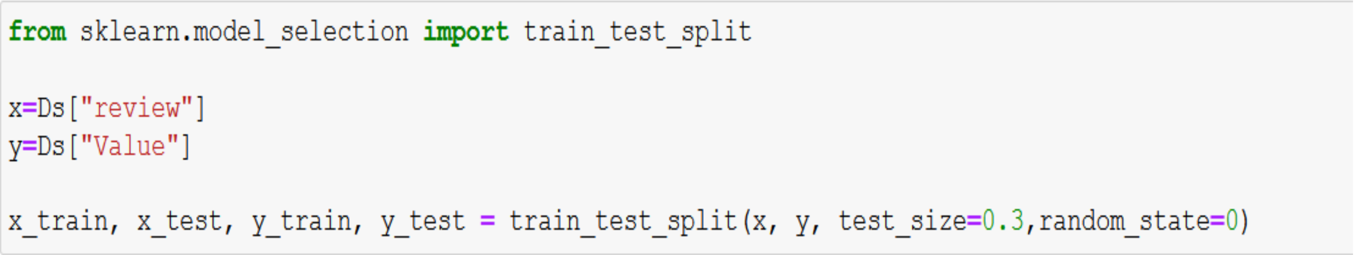
Figure 11 Converting Sentiment Polarity into Numerical Values

The above figure explains that system have convert sentiment type "Positive" into numerical value "1" and sentiment type "Negative" into numerical value into "0". System will create a new field called "Value" to store the numerical values of each sentiment type.

Data Splitting

Dataset will be split into Training data and Testing data. To train every machine learning algorithm, regardless of the type of dataset used, System must divide the dataset into training data and testing data. System may use this method to determine the model hyper-parameter as well as estimate the generalization efficiency.

The **“train test split()”** method in the scikit-learn Python machine learning package implements the train-test split evaluation technique. The function accepts a supplied dataset and splits it into two subgroups as an output. Developer should separate the original dataset into input (X) and output (y) columns, then provide both arrays to the function, which should split them into train and test subsets correctly. The **“test size”** input accepts a percentage of the dataset size between 0 and 1 to specify the size of the split (Brownlee,2020).



```
from sklearn.model_selection import train_test_split

x=Ds["review"]
y=Ds["Value"]

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3,random_state=0)
```

Figure 12 Data Splitting Preview

System will split the dataset into test data and Training data. System will use 30% the records to Testing and the rest of 70% will be used for Training the data. The More records System uses to Train the Model, System will learn more about the model. It will help the system to get a higher accuracy and it helps to predict the new reviews into Positive or Negative.

TF-IDF Vectorization

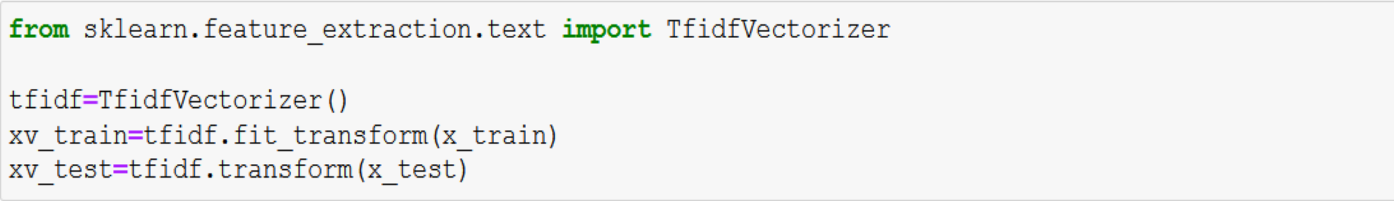
TF-IDF stands for “Term Frequency — Inverse Document Frequency”. Machine learning with natural language faces one big challenge. The algorithms typically work with numbers, But the Reviews are in Text format. So, the algorithm must translate the text into numbers, better known as text vectorization. TF-IDF allows one to equate each word in a document with a number that reflects the importance of each word in that document.

Term Frequency (tf): gives the frequency of the word in each text in the corpus. That is the ratio of the number of times the word appears in the text to the overall number of words in the document. It increases as the number of events inside the text increases. A Document has its own kind of document.

Inverse Data Frequency (idf): used to measure the weight of rare terms across all corpus records. Words that seldom appear in the corpus have a high IDF score.

System needs to understand the vocabulary of the training data. System will use **“fit_transform”** method to learn the vocabulary and IDF (Inverse Data Frequency) from training data.

The Below figure explains the process of vectorization.



```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf=TfidfVectorizer()
xv_train=tfidf.fit_transform(x_train)
xv_test=tfidf.transform(x_test)
```

Figure 13 Text Vectorization using TF-IDF

System applies the Training reviews and Testing Reviews to TF-IDF vectorizer.in this method, Training reviews will learn the Vocabulary and Inverse Data Frequency. Then it returns document-term matrix. After that Testing Reviews will learn the vocabulary from the Training reviews.

Machine Learning Algorithms

For the classification of Movie reviews, the system applies machine learning approaches to textual reviews.

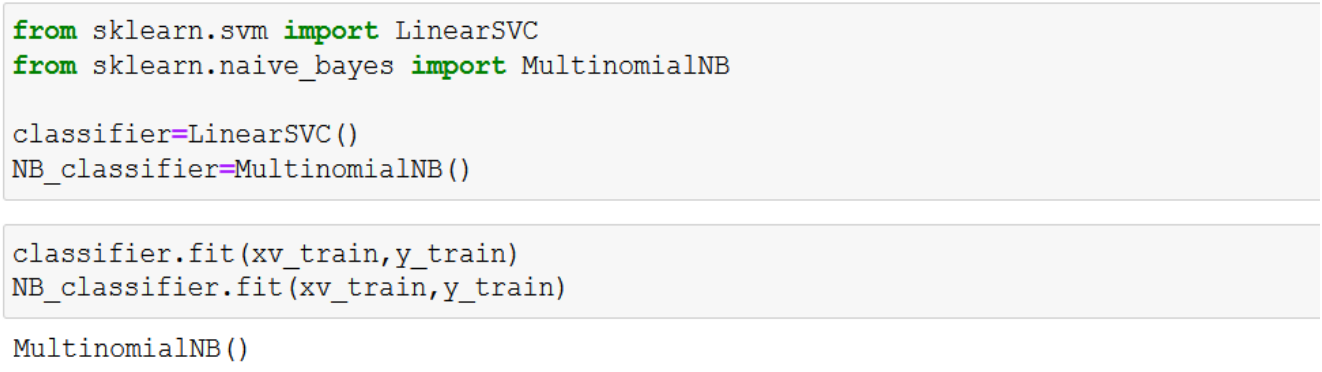
By Applying the Vectorized text Reviews to the Classifiers system will complete the Model Training Process. After Adding the classifiers, each classifier will give an Accuracy value of Vectorized Test Reviews and Testing values of Independent Variable. Proposed Machine Learning classifiers for this system are Multinomial Naïve Bayes and Linear SVC Classifier.

Multinomial Naïve Bayes Classifier

It is a classification technique based on Bayes' theorem with an assumption of independence between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes uses a related approach to estimate the probability of different class depending on different attributes. This algorithm is mainly used for text classification and has problems with several classes.

Linear SVC Classifier

As already discussed in the Literature Review part, Linear Support Vector Classifier uses a linear kernel function to perform classification and works well with many samples.



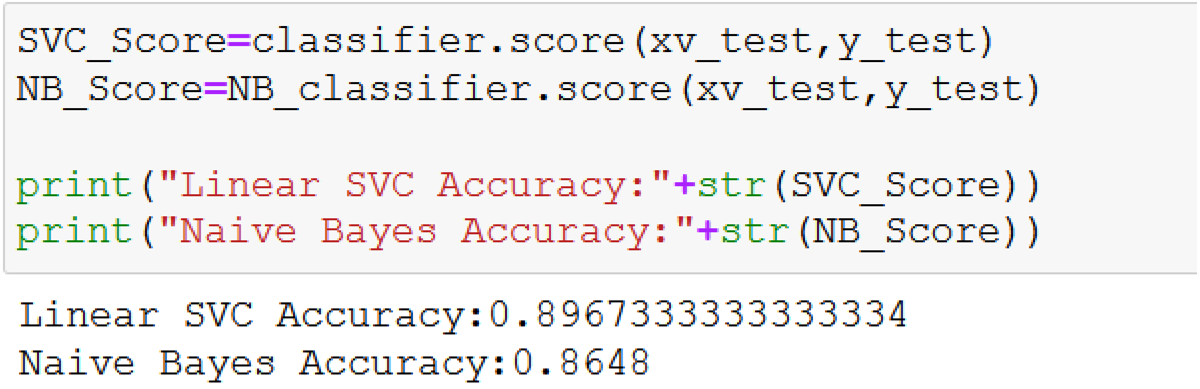
```
from sklearn.svm import LinearSVC
from sklearn.naive_bayes import MultinomialNB

classifier=LinearSVC()
NB_classifier=MultinomialNB()
```

```
classifier.fit(xv_train,y_train)
NB_classifier.fit(xv_train,y_train)
```

```
MultinomialNB()
```

Figure 14 Fitting the Vocabulary to Classifier Preview

Vectorized Training Reviews and Training Independent Variables will be fit in to each classifier. So now the Model has been trained Successfully. Now system must select only one Classifier. To select a suitable classifier, The Accuracy Level of Each Classifier will be compared. The Highest Accuracy value holding Classifier will be the Most suitable Classifier.



```
SVC_Score=classifier.score(xv_test,y_test)
NB_Score=NB_classifier.score(xv_test,y_test)

print("Linear SVC Accuracy:"+str(SVC_Score))
print("Naive Bayes Accuracy:"+str(NB_Score))
```

```
Linear SVC Accuracy:0.8967333333333334
Naive Bayes Accuracy:0.8648
```

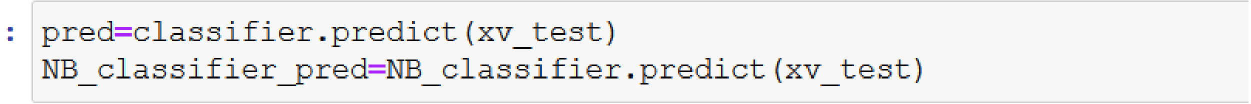
Figure 15 Accuracy Score of Classifiers

The Above figure explains the Accuracy of Each Classifier. Linear SVC Gives an Accuracy Value of **89%** and Multinomial Naïve Bayes Gives an Accuracy Value of **86%.** So, the Highest accuracy and the suitable classifier for the Model is Linear SVC Classifier.

Classification Report

Classification Report displays precision, recall, F1, and support scores of the Trained Model.

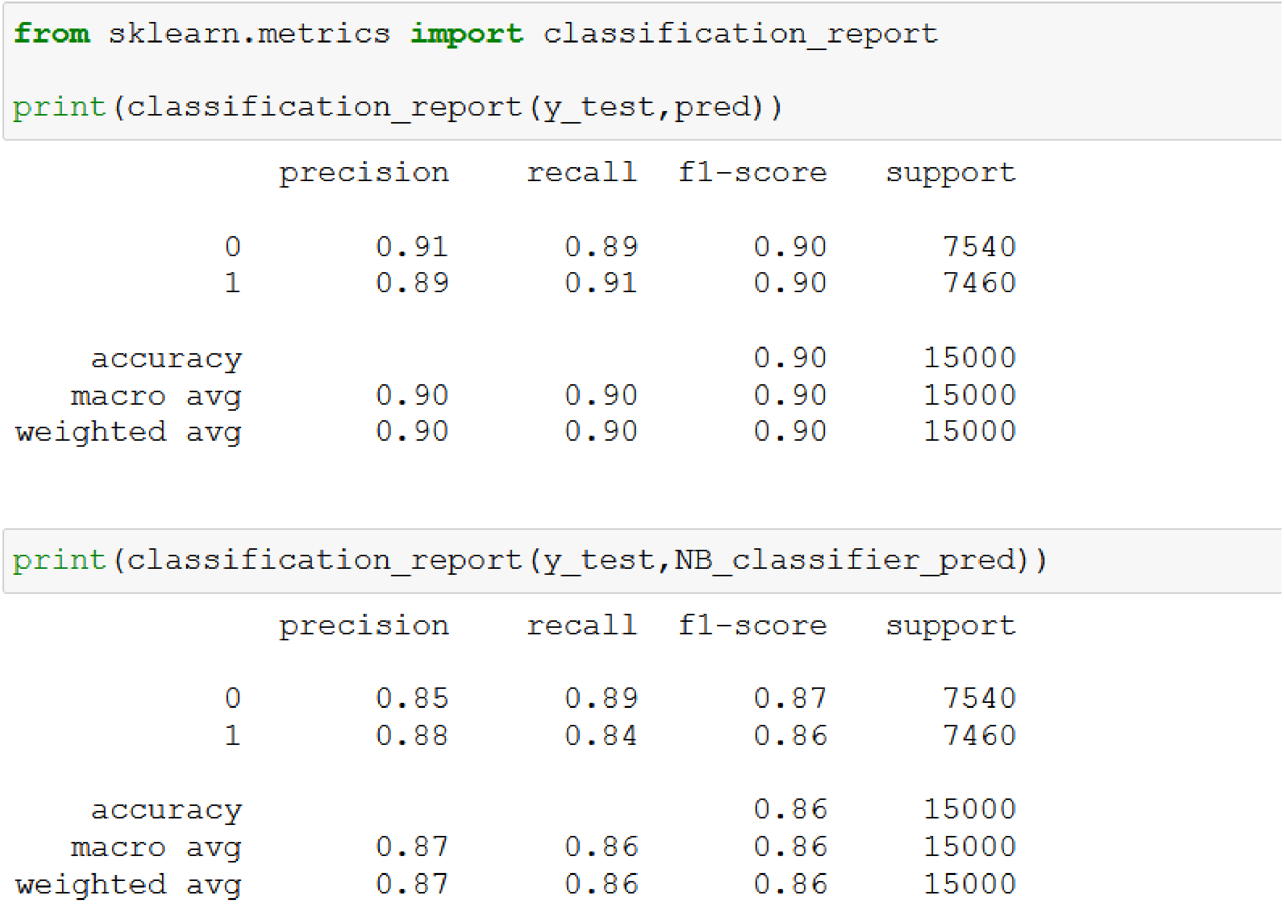
Below Figure Explains the prediction process of each classifier. To Display the Classification Report, each classifier needs to be predicted the Vectorized Testing Reviews. It will give an Array List of whether each review is 1 or 0.



```
: pred=classifier.predict(xv_test)
  NB_classifier_pred=NB_classifier.predict(xv_test)
```

Figure 16 Implementing Classification Report preview

Finally, System will display the Classification report using independent test variables and classifier predicted array list. The Below figure displays the classification report of each Classifier.



```
from sklearn.metrics import classification_report

print(classification_report(y_test,pred))
```

```
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7540
           1       0.89      0.91      0.90      7460

    accuracy                           0.90     15000
   macro avg       0.90      0.90      0.90     15000
weighted avg       0.90      0.90      0.90     15000
```

```
print(classification_report(y_test,NB_classifier_pred))
```

```
              precision    recall  f1-score   support

           0       0.85      0.89      0.87      7540
           1       0.88      0.84      0.86      7460

    accuracy                           0.86     15000
   macro avg       0.87      0.86      0.86     15000
weighted avg       0.87      0.86      0.86     15000
```

Figure 17 Classification Report Preview

Precision - Briefly, it is the Accuracy of Positive Predictions. Accuracy will be same as the Accuracy Score that the document has discussed above. Precision refers to a classifier's ability to avoid labeling a negative instance as positive. It is calculated as the ratio of true positives to the number of true positives and false positives for each class.

Recall - Correctly defined positives as a percentage of total positives. The capacity of a classifier to identify all positive instances is known as recall. It is calculated as the ratio of true positives to the number of true positives and false negatives for each class.

F1-Score - The F1 score is a weighted harmonic mean of precision and recall, with 1.0 being the highest and 0.0 being the lowest. F1 ratings are lower than accuracy tests because they factor in precision and recall. To evaluate classifier models, use the weighted average of F1 rather than global accuracy as a rule of thumb.

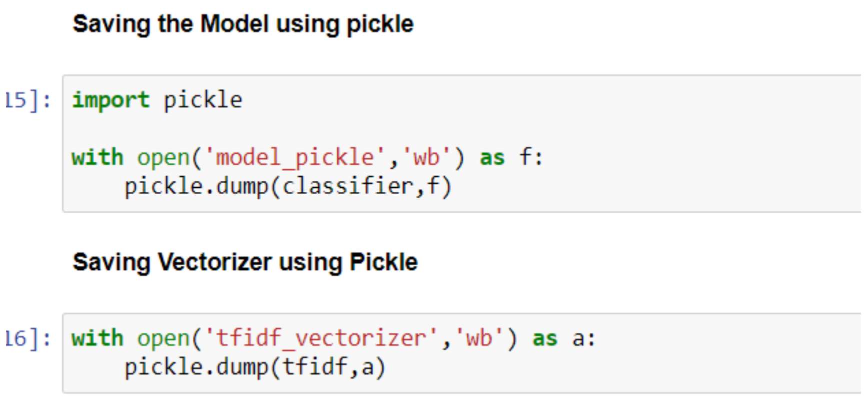
Support - The number of actual occurrences of the class in the listed dataset is known as support.

Finally, System have displayed the classification report and Now System successfully Trained the Model. Between the Two Machine Learning Classifiers, Linear SVC has Given the Highest Prediction Accuracy Level of **89%**. So, system has selected and Continue the Process using Linear SVC Classifier.

Saving the Trained Classifier and Text Vectorizer

We must store the classifier and the TfidfVectorizer for usage in production in order to move the model to production. Using the pickle Python package, this is possible in Python. The pickle module implements binary serialization and deserialization protocols for Python objects. Pickling is the process of converting a Python object into a byte stream.

Pickle's dump function stores the vectorizer and classifier to the current working directory.



Saving the Model using pickle

15]:
```
import pickle

with open('model_pickle','wb') as f:
    pickle.dump(classifier,f)
```

Saving Vectorizer using Pickle

16]:
```
with open('tfidf_vectorizer','wb') as a:
    pickle.dump(tfidf,a)
```

Figure 18 Saving Classifier and Vectorizer Code Preview

After saving the classifier and the Text Vectorizer, Web Application no longer need to train the model to predict new reviews. Using the saved classifier and Text Vectorizer, Web Application can predict the reviews easily and rapidly.

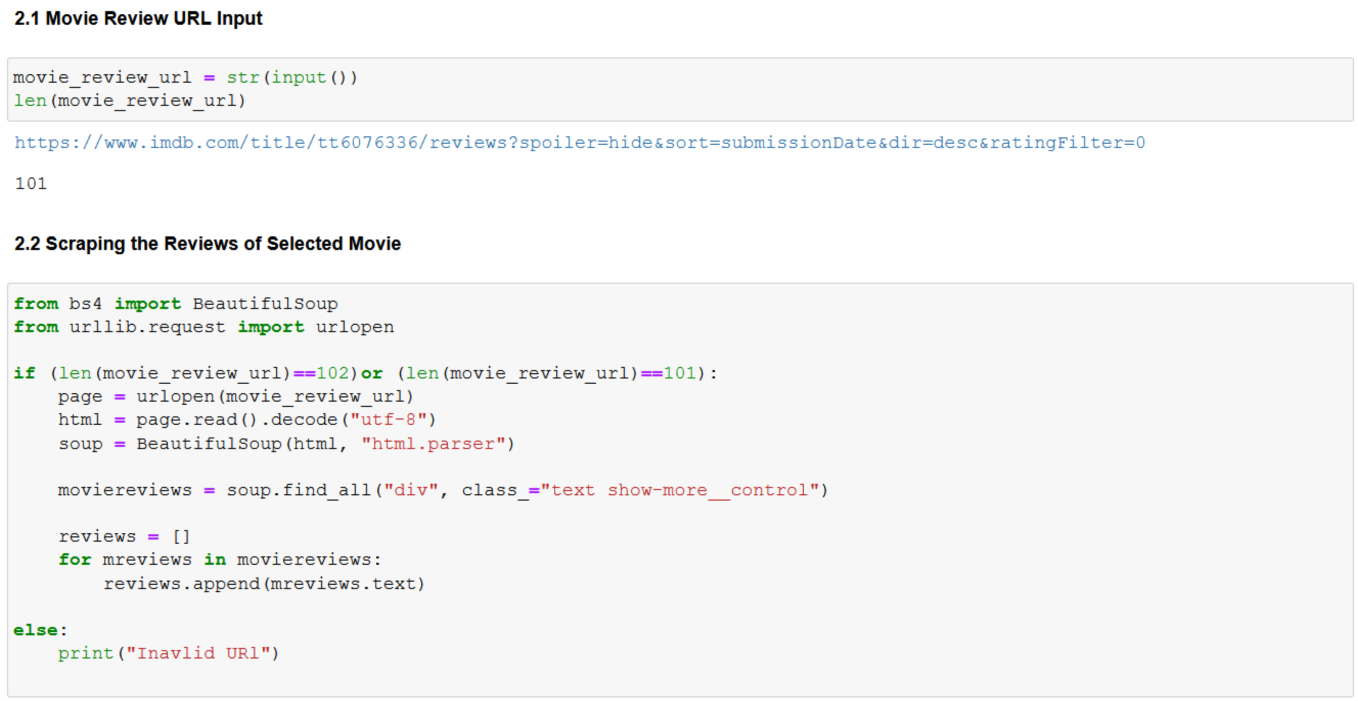
Web Scraping the Movie Reviews

Web Scraping is the practice of extracting information from the Internet. The terms "Web Scraping" typically apply to a method requiring automation. Some websites do not like it when automated scrapers extract their data, while others do not mind it.

The Urllib module is a Python module for managing URLs. It is used to get URLs from the internet (Uniform Resource Locators). It makes use of the urlopen function and may retrieve URLs using a variety of protocols. By using urlopen Library, system will open the URL which is Submitted by the End User. After directing to the URL, System can Start the Web Scraping process.

Beautiful Soup is a Python library for collecting data from HTML, XML, and other markup languages. Beautiful Soup helps to delete specific material from the web page, remove the HTML markup, and save the details. It is a database scraping app that helps clean up and parse documents that have been taken down from the web.

When the User Copy the Link from IMDB Web and Paste it in the Web Applications Text Area which is provided to Paste the link, System will Scrape the latest 25 Movie Reviews.



2.1 Movie Review URL Input

```
movie_review_url = str(input())
len(movie_review_url)
```

```
https://www.imdb.com/title/tt6076336/reviews?spoiler=hide&sort=submissionDate&dir=desc&ratingFilter=0

101
```

2.2 Scraping the Reviews of Selected Movie

```
from bs4 import BeautifulSoup
from urllib.request import urlopen

if (len(movie_review_url)==102)or (len(movie_review_url)==101):
    page = urlopen(movie_review_url)
    html = page.read().decode("utf-8")
    soup = BeautifulSoup(html, "html.parser")

    moviereviews = soup.find_all("div", class_="text show-more__control")

    reviews = []
    for mreviews in moviereviews:
        reviews.append(mreviews.text)

else:
    print("Inavlid URl")
```

Figure 19 Web Scraping Method Preview

Above Figure explains the Process of Web Scraping using Beautiful Soup. Beautiful Soup will scrape the latest reviews from the given URL. If the URL is valid, System will Continue the Scraping process. If the Provided URL Not Valid, System will display an error message displaying “Invalid URL”.

Development Platform

- Programming Language

Proposed System is based on Python Programming language. Python is the preferred programming language of choice for machine learning for several IT giants like Google, Instagram, Twitter, Dropbox, Netflix, Walt Disney, YouTube, Uber, Amazon, and Reddit. Python is the undisputed king and by far the strongest machine-learning language of today, and here is why:

- Extensive selection of books and packages
- Readability of the Code
- Flexibility (Luashchuk,2019).

Python 3.9.1 version used to implement the proposed System.

HTML is Hyper Text Markup Language. It is used to build web pages in the markup language. HTML is a mixture of Hypertext and Markup. Hypertext describes a relation between a web page. Markup language is used to describe a text document inside a tag that describes a web page layout. HTML 5 is the sixth and latest versions of HTML. Since the Proposed Solution is a Web Application HTML5 will be for Display Data with Eye catching GUI.

- IDE

PyCharm is one of the commonly used Python IDEs generated by JetBrains. It is one of the better IDEs for Python. PyCharm is what the developer's required for the efficient development of Python. With PyCharm, developers can write a tidy and up-to-date file (Solution,2018).

- Framework

Python frameworks are plentiful because it is one of the most popular programming languages. Each framework has its own set of benefits and drawbacks. As a result, the decision must be made based on the project's requirements as well as the preferences of the developer.

Flask is a web-based programming platform written in Python. Flask is a lightweight framework that is commonly referred to as a micro framework. Flask comes with several basic features and lets developers to add as many libraries or plugins as they like to an extension. It is being developed by Armin Ronacher, who leads an international community of Python enthusiasts called Pocco. Flask is based on the WSGI toolkit toolkit and the Jinja2 template engine. They are both Pocco ventures (Singh,2021). The suggested System will be built using the work of the Flask Framework.

Chapter 05: Solution Concept

Component Diagram

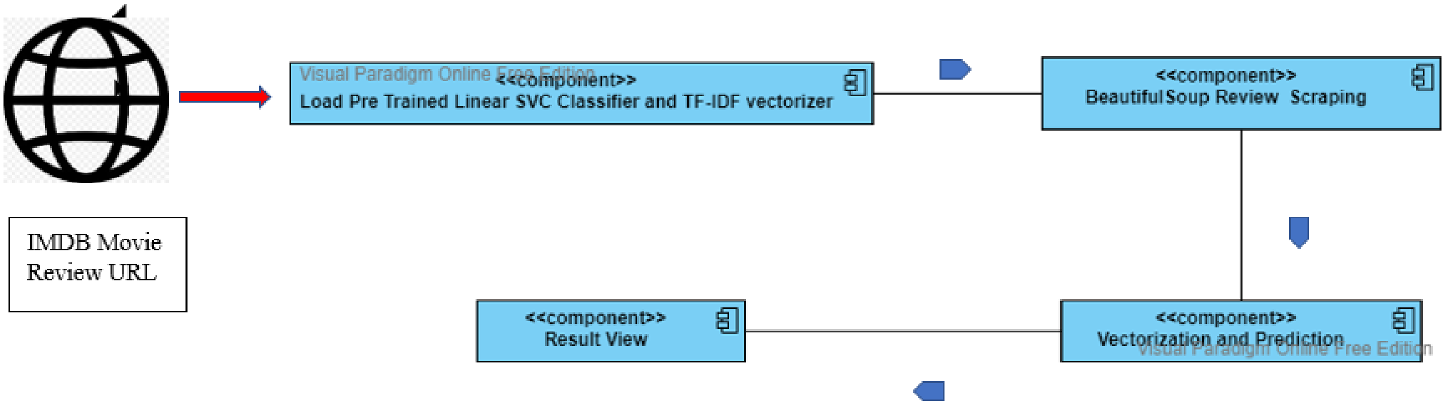

IMDB Movie Review URL
Visual Paradigm Online Free Edition
<<component>>
Load Pre Trained Linear SVC Classifier and TF-IDF vectorizer
<<component>>
Beautiful Soup Review Scraping
<<component>>
Vectorization and Prediction
Visual Paradigm Online Free Edition
<<component>>
Result View

Figure 20 Component Diagram

When user Submits the valid IMDB Movie Review URL system will Load the Pre trained Linear SVC Model and TF-IDF Vectorizer component. Then system will scrape the latest 25 reviews from the URL using BeautifulSoup Review Scarping Component. Then Vectorization and predictions component will take the reviews through Data cleaning method, vectorize them and predict the polarity. After that Result View component will display the information in the result page.

Chapter 06: Requirement Analysis

High Level Use Case Diagram

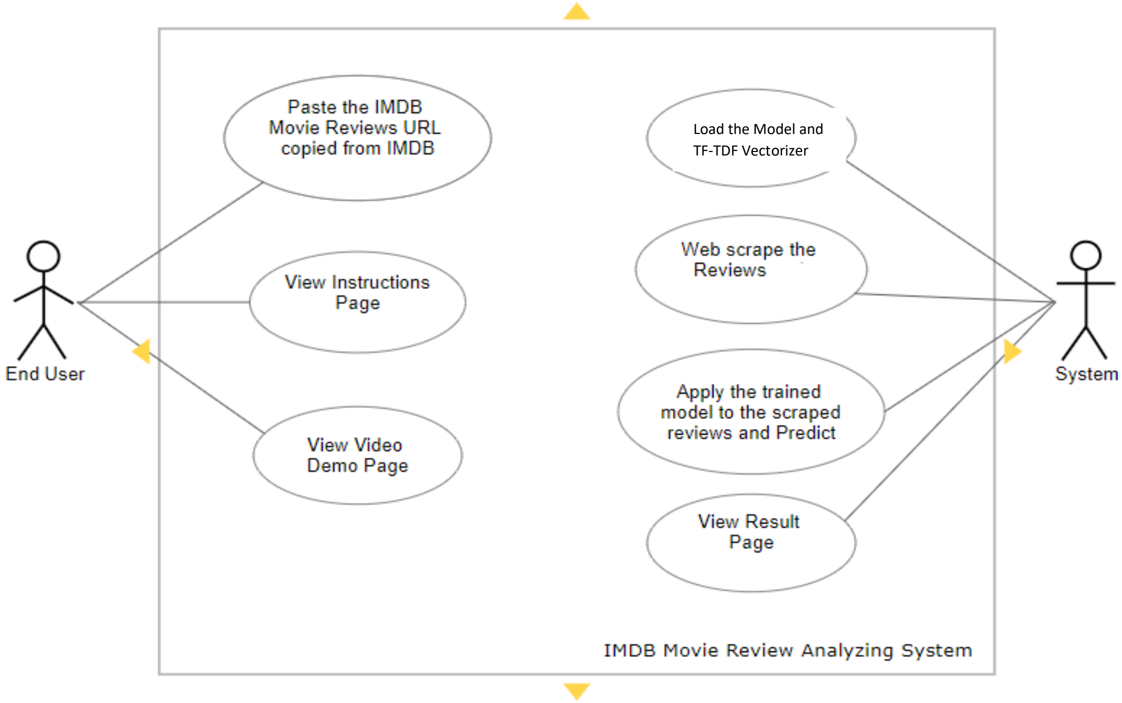

Paste the IMDB Movie Reviews URL copied from IMDB
View Instructions Page
View Video Demo Page
Load the Model and TF-TDF Vectorizer
Web scrape the Reviews
Apply the trained model to the scraped reviews and Predict
View Result Page
End User
System
IMDB Movie Review Analyzing System

Figure 21 High Level Use Case Diagram

Use case Scenarios

Use Case Name	Paste the IMDB Movie Reviews URL Copied from IMDB
Use Case ID	001
Description	Users must provide the IMDB Movie Review URL of selected Movie to Start the process.
Actor	User
Steps	01.User must select a movie from IMDB web and view the User Reviews of the Selected Movie. 02. User must Make sure that “Hide Spoilers” Checkbox option is checked. 03.User must copy the Web URL and Paste it in the System Home Page Text Area.
Precondition	None

Use Case Name	View Instructions Page
Use Case ID	002
Description	To view the instructions page, user must select View Instructions from Navigation bar.
Actor	User
Steps	01.User Selects View Instructions. 02. System displays instructions page.
Precondition	None

Use Case Name	View Video Demo Page
Use Case ID	003
Description	To view the Video demo page, user must select View Video Demo from Navigation bar.
Actor	User
Steps	01.User Selects Video Demo. 02. System displays Video Demo Page.
Precondition	None

Use Case Name	Load the Model and TF-IDF Vectorizer
Use Case ID	04
Description	Loading the model is the first step of the process. Model and TF-IDF Vectorizer will play an important role in the prediction process.
Actor	User
Steps	01. System will Load the pickle file which contains the pre trained Linear SVC model. 02. System will Load the pickle file which contains the TF-IDF vectorizer.
Precondition	The IMDB Movie review URL should be a valid one to start the process. If else system will display an error page.

Use Case Name	Web Scrape the Reviews
Use Case ID	05
Description	After loading the model, system needs to apply it to the User selected movie reviews. To collect the User selected movie reviews, system will web scrape the reviews.
Actor	User
Steps	01. System will go through the pasted URL and start to scrape the reviews. 02. System will store the Scraped reviews in a Array.
Precondition	The IMDB Movie review URL should be a valid one to start the process. If else system will display an error page.

Use Case Name	Apply the trained model to the scraped reviews and predict
Use Case ID	06
Description	System must predict the new reviews into Positive reviews or Negative reviews. For that system will apply the previously trained model to new reviews.
Actor	User
Steps	01. System will preprocess the new reviews. 02. System will vectorize the reviews. 03. System will apply the previously trained classifier to new reviews to predict. 04. System will predict and return each review with sentiment Type (Positive or Negative)
Precondition	The IMDB Movie review URL should be a valid one to start the process. If else system will display an error page.

Use Case Name	View Result Page
Use Case ID	07
Description	After predicting the new reviews into positive or negative reviews, system will calculate and display it in result page.
Actor	User
Steps	01. System will display a bar chart referring the amount of negative and positive reviews, Percentage of predicted positive and negative reviews and recommendation status on Result Page.
Precondition	The IMDB Movie review URL should be a valid one to start the process. If else system will display an error page.

Chapter 07: System Design

Activity Diagrams

<u>Review Analyzing Activity Diagram</u>

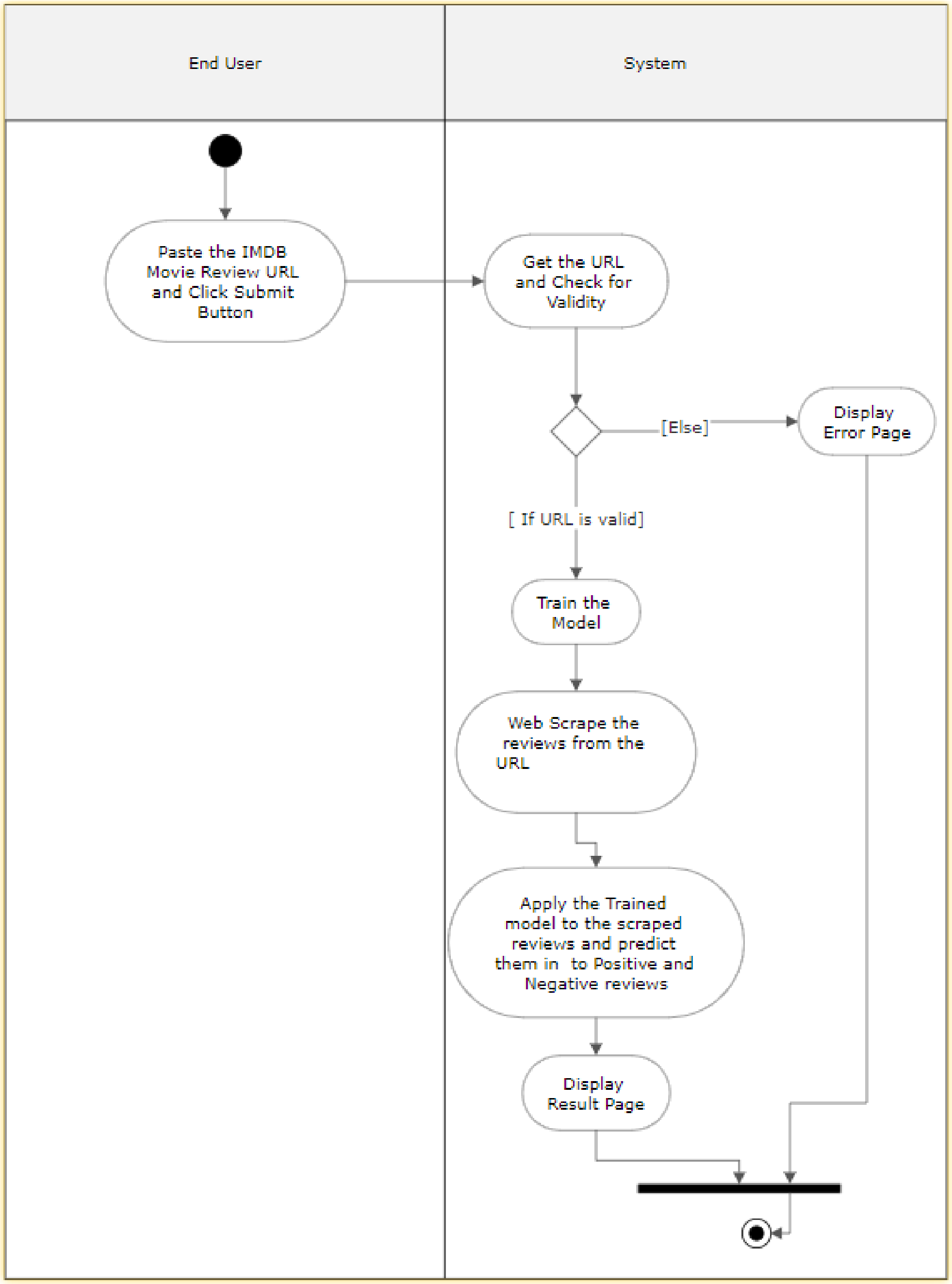

End User
System
Paste the IMDB Movie Review URL and Click Submit Button
Get the URL and Check for Validity
[Else]
Display Error Page
[If URL is valid]
Train the Model
Web Scrape the reviews from the URL
Apply the Trained model to the scraped reviews and predict them in to Positive and Negative reviews
Display Result Page

Figure 22 Review Analyzing Activity Diagram

View Instructions Page Activity Diagram

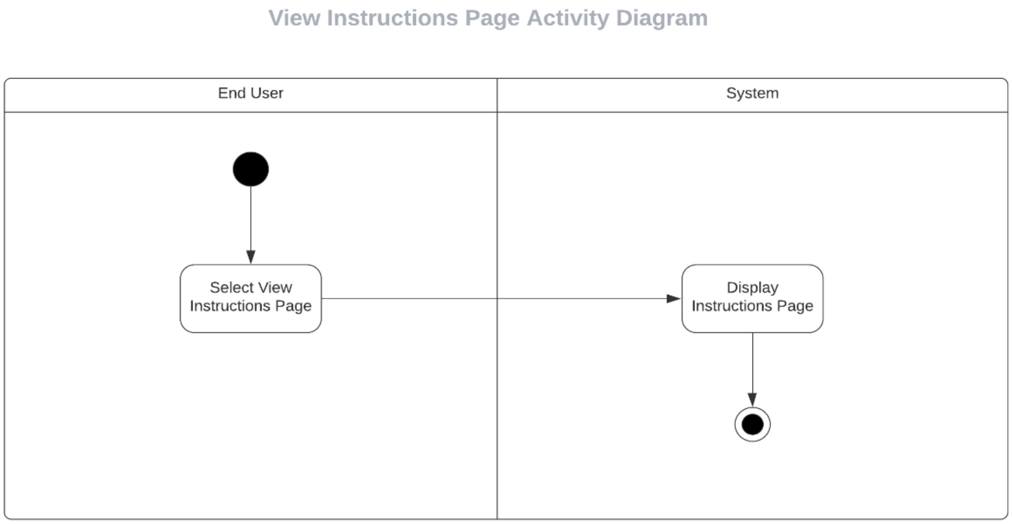

View Instructions Page Activity Diagram
End User
System
Select View Instructions Page
Display Instructions Page

Figure 23 View Instructions Page Activity Diagram

View Video Demo Page Activity Diagram

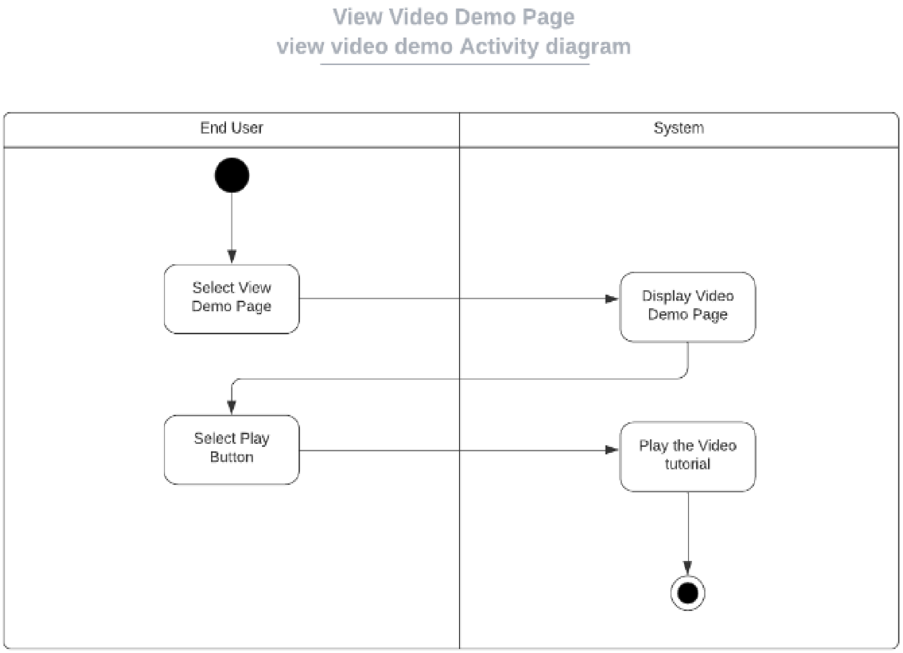


Figure 24 View Video Demo Activity Diagram

Sequence Diagram

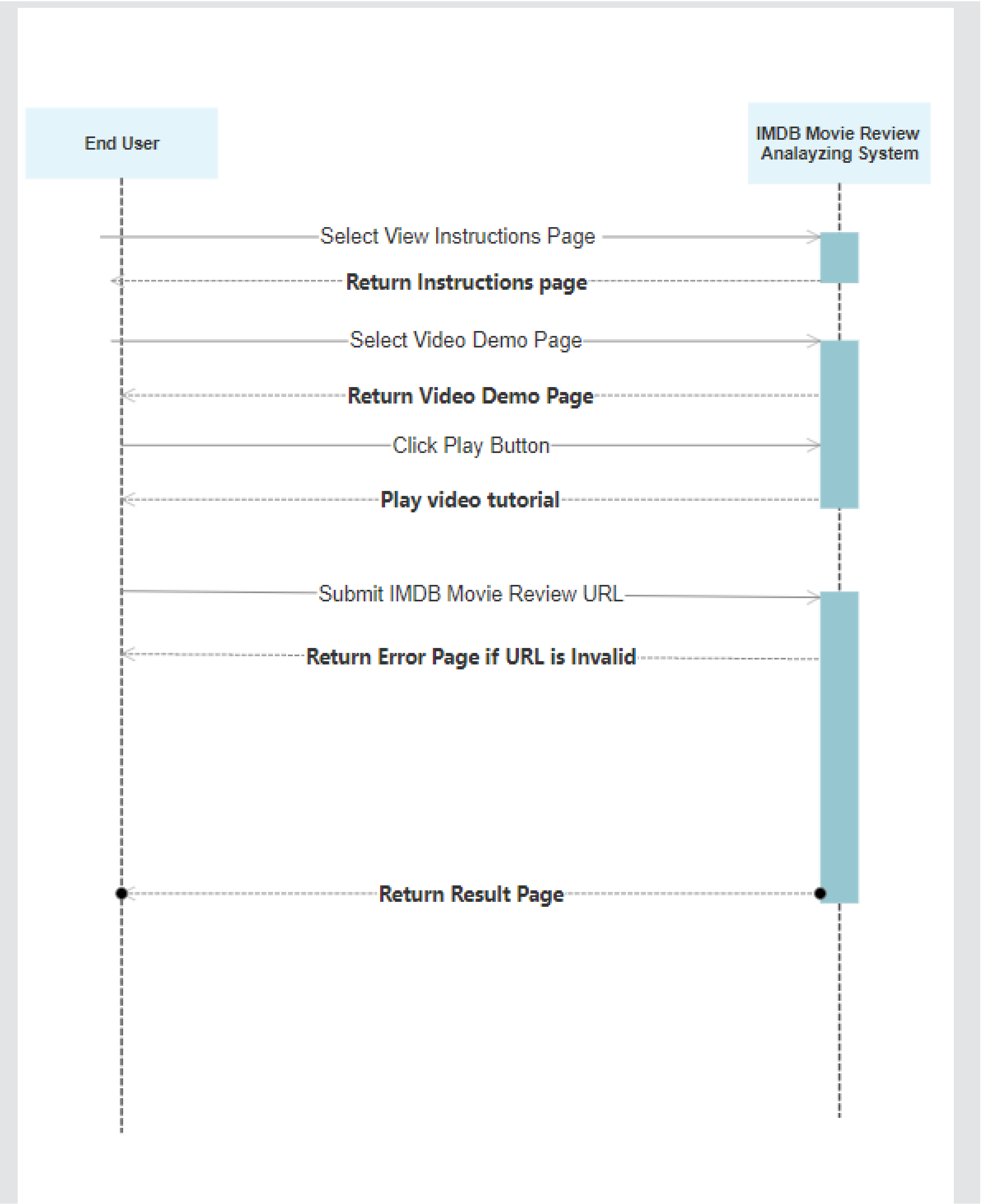


Figure 25 Sequence Diagram

Chapter 08: Implementation

Hardware and Software Requirements

Hardware Requirements

- Device Mechanism - MSI GF63 Thin 95c
- RAM – 8GB
- VGA- 4GB NVIDIA GEFORCE GTX 1650
- Core i5 9th Gen

Software Requirements

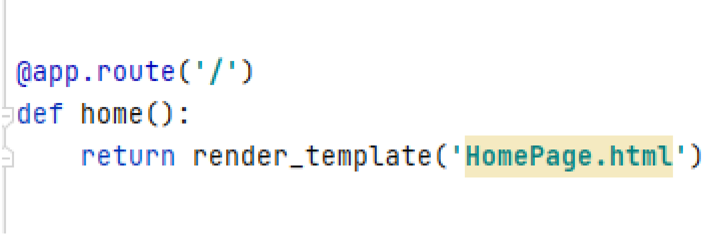
- Operating System – Windows 10
- Pycharm IDE
- Web Browser (Google Chrome, Mozilla Firefox etc.)

Core Features and Code Excerpts

As discussed in the Methodology, This Back end of this Web Application will be based Python programming Language using Flask Framework. Front end of this wen Application will be based on HTML5, Booostrap5 CSS and External CSS.

System has only one python class. it is named as “main.py” file. This python class contains the backend implementation of the web application. In this Part document will discuss about the code methods for each process and how each process works.

The flask.templating package's render template method is a Flask function. render template is used to generate output from a template file located in the templates folder of the application (Makai,2021). In this project, the template files will be HTML files.

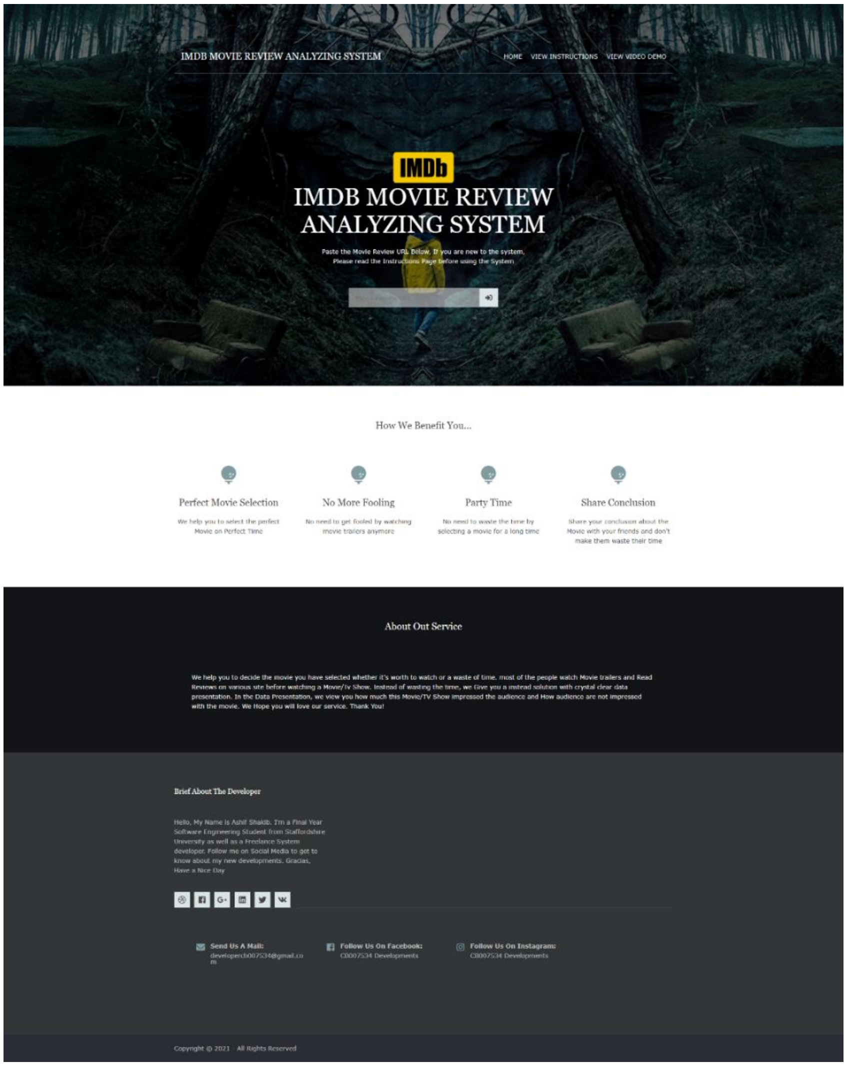


```
@app.route('/')
def home():
    return render_template('HomePage.html')
```

Figure 26 Home Page direction Method in Web Application

System uses this method to display the home page to the user. This method will direct the system to display the home page. “HomePage.html” will be the template file of Homepage.

Below figure displays the preview of the Home Page.



IMDB MOVIE REVIEW ANALYZING SYSTEM

Paste the Movie Review URL Below, If you are new to the system, Please read the Instructions Page before using the System

How We Benefit You...

Perfect Movie Selection

We help you to select the perfect Movie on Perfect Time

No More Fooling

No need to get fooled by watching movie trailers anymore

Party Time

No need to waste the time by selecting a movie for a long time

Share Conclusion

Share your conclusion about the Movie with your friends and don't make them waste their time

About Out Service

We help you to decide the movie you have selected whether it's worth to watch or a waste of time. most of the people watch Movie trailers and Read Reviews on various site before watching a Movie/Tv Show. Instead of wasting the time, we Give you a instead solution with crystal clear data presentation. In the Data Presentation, we view you how much this Movie/TV Show impressed the audience and How audience are not impressed with the movie. We Hope you will love our service. Thank You!

Brief About The Developer

Hello, My Name is Ashif Shakib. I'm a Final Year Software Engineering Student from Staffordshire University as well as a Freelance System developer. Follow me on Social Media to get to know about my new developments. Gracias, Have a Nice Day

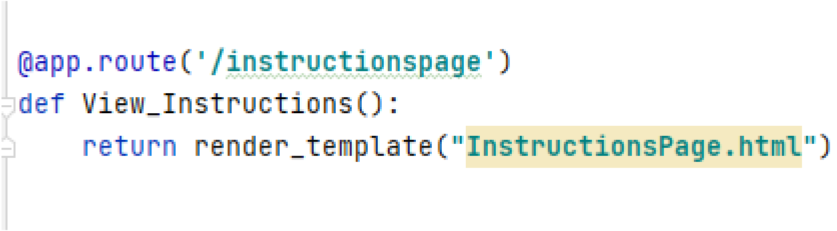
Send Us A Mail:
developercb007534@gmail.com

Follow Us On Facebook:
CB007534 Developments

Follow Us On Instagram:
CB007534 Developments

Copyright © 2021 · All Rights Reserved

Figure 27 Homepage of Web Application Preview

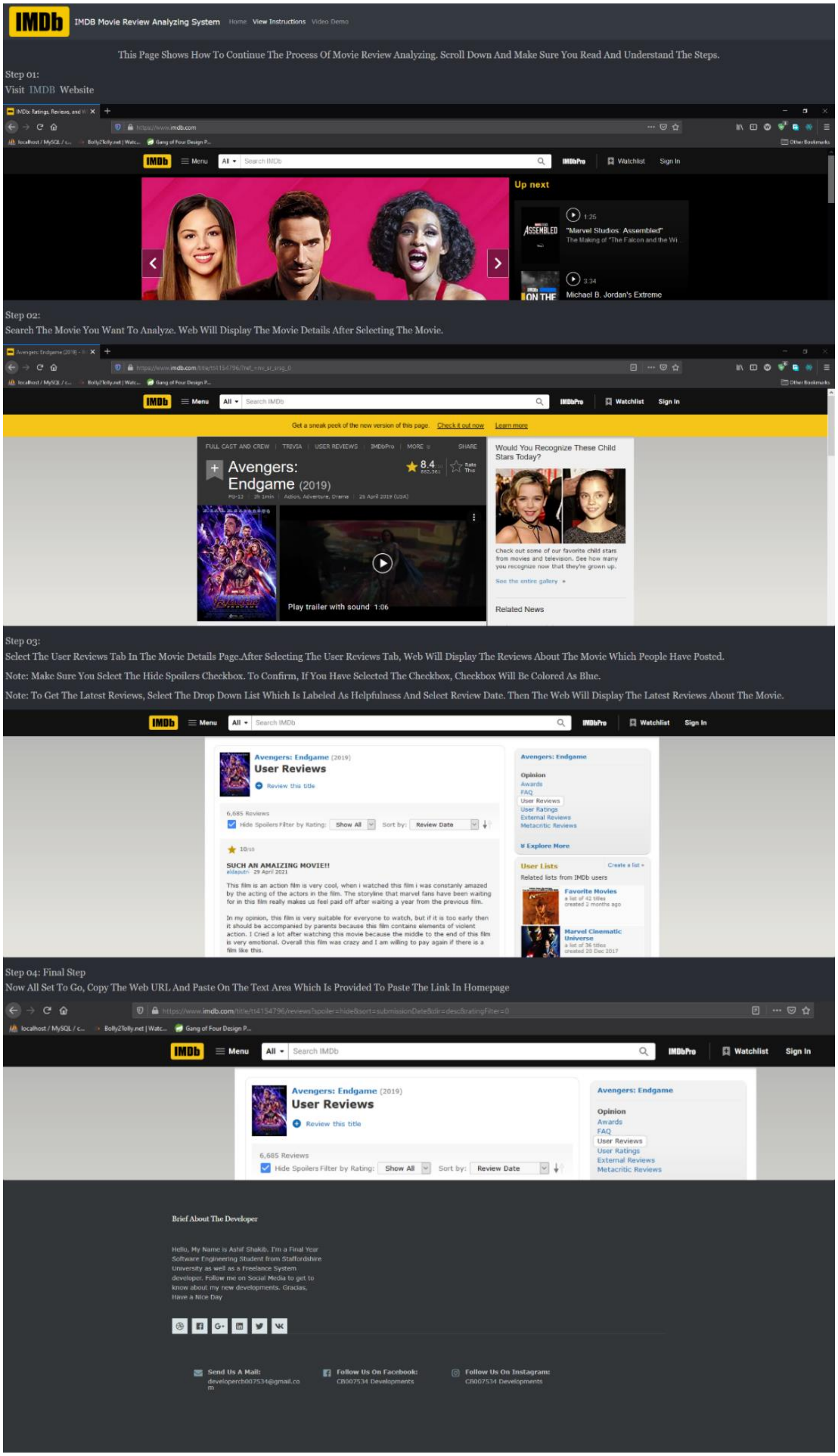


```
@app.route('/instructionspage')
def View_Instructions():
    return render_template("InstructionsPage.html")
```

Figure 28 Instructions Page direction method in Web Application

System uses this method to display Instruction page to the user. When user selects View Instructions from Navigation bar System will direct to Instruction page. "InstructionsPage.html" will be the template file of Instruction page.

This page will display user how to copy the link from the IMDB Web. If the link is not in the proper format, then system will display an error message. Below figure displays the preview of the Instruction Page.



This Page Shows How To Continue The Process Of Movie Review Analyzing. Scroll Down And Make Sure You Read And Understand The Steps.

Step 01:
Visit IMDB Website

Step 02:
Search The Movie You Want To Analyze. Web Will Display The Movie Details After Selecting The Movie.

Step 03:
Select The User Reviews Tab In The Movie Details Page.After Selecting The User Reviews Tab, Web Will Display The Reviews About The Movie Which People Have Posted.

Note: Make Sure You Select The Hide Spoilers Checkbox. To Confirm, If You Have Selected The Checkbox, Checkbox Will Be Colored As Blue.

Note: To Get The Latest Reviews, Select The Drop Down List Which Is Labeled As Helpfulness And Select Review Date. Then The Web Will Display The Latest Reviews About The Movie.

Step 04: Final Step
Now All Set To Go, Copy The Web URL And Paste On The Text Area Which Is Provided To Paste The Link In Homepage

Brief About The Developer

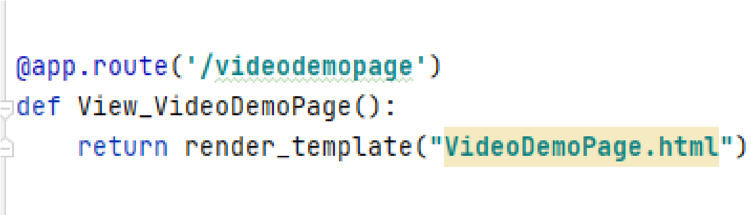
Hello, My Name is Ashif Shakib. I'm a Final Year Software Engineering Student from Staffordshire University as well as a Freelance System developer. Follow me on Social Media to get to know about my new developments. Gracias, Have a Nice Day

Send Us A Mail:
developercb007534@gmail.com

Follow Us On Facebook:
CB007534 Developments

Follow Us On Instagram:
CB007534 Developments

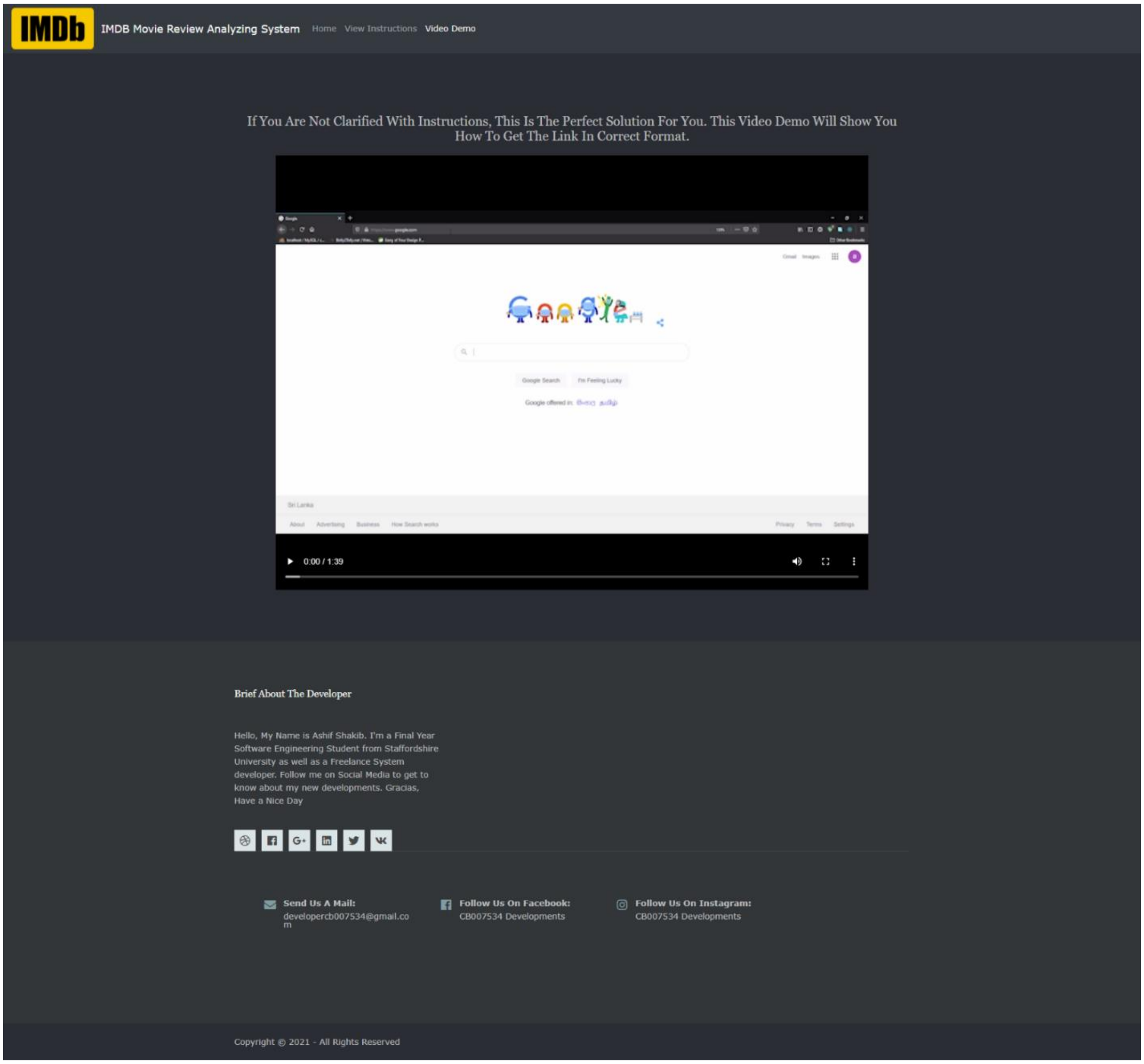
Figure 29 Instructions Page Preview



```
@app.route('/videodemopage')
def View_VideoDemoPage():
    return render_template("VideoDemoPage.html")
```

Figure 30 Video demo page direction method in Web Application

System uses this method to video page to the user. When user selects Video Demo from Navigation bar System will direct to Video demo page. “VideoDemoPage.html” will be the template file of Video Demo page. This page will display a quick demo video of how to copy the link of selected Movie/TV Show from IMDB web. Below figure displays the preview of the video demo page.



If You Are Not Clarified With Instructions, This Is The Perfect Solution For You. This Video Demo Will Show You How To Get The Link In Correct Format.

Brief About The Developer

Hello, My Name is Ashif Shakib. I'm a Final Year Software Engineering Student from Staffordshire University as well as a Freelance System developer. Follow me on Social Media to get to know about my new developments. Gracias, Have a Nice Day

Send Us A Mail: developercb007534@gmail.com

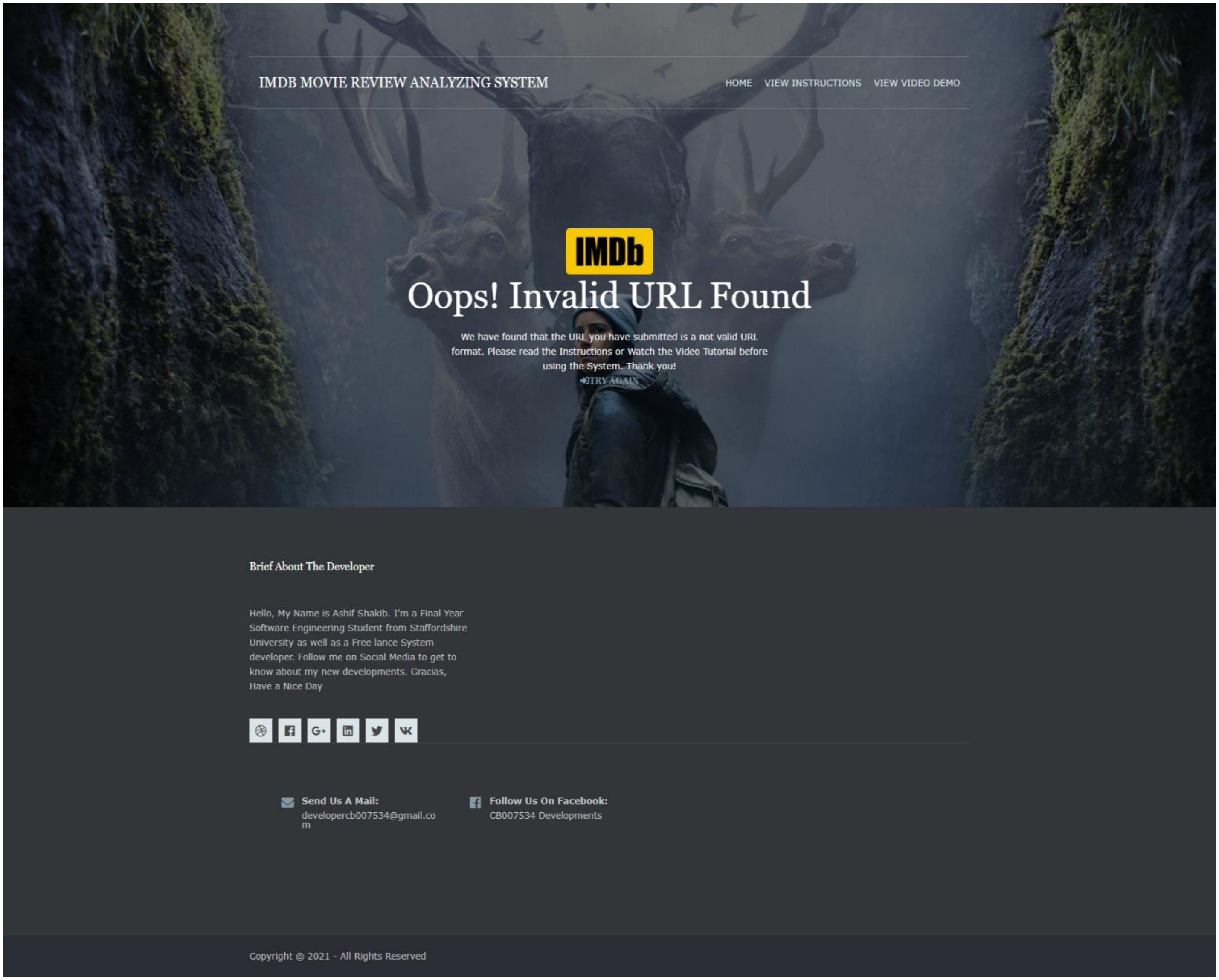
Follow Us On Facebook: CB007534 Developments

Follow Us On Instagram: CB007534 Developments

Copyright © 2021 - All Rights Reserved

Figure 31 Video Demo Page Preview

If the text area which is provided to paste the link in Home Page found empty, then system will display error message requesting user to paste the URL. In additional, System will direct the user to Error Page. The length of the expected URL format is 101 characters, or 102 characters and URL should start with **“https://www.imdb.com/title/”.** When user pasted the URL and submits, system will check the length of the pasted link and whether it starts with the valid format. If the link is not valid, system will direct the user to an Error Page. The below figure displays the preview of the scenario.



Oops! Invalid URL Found

We have found that the URL you have submitted is a not valid URL format. Please read the Instructions or Watch the Video Tutorial before using the System. Thank you!

Brief About The Developer

Hello, My Name is Ashif Shakib. I'm a Final Year Software Engineering Student from Staffordshire University as well as a Free lance System developer. Follow me on Social Media to get to know about my new developments. Gracias, Have a Nice Day

Send Us A Mail:
developercb007534@gmail.com

Follow Us On Facebook:
CB007534 Developments

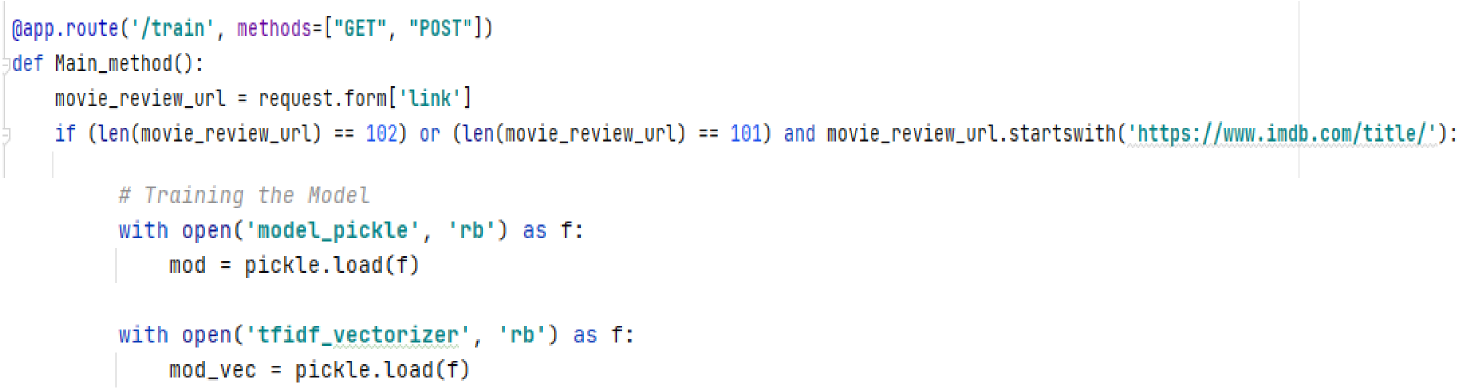
Copyright © 2021 - All Rights Reserved

Figure 32 Invalid URL page Preview

If it is a valid link, system will start the process. The process will take some time to complete.

Now onwards, document will discuss about the main objective of the project. If the given link is valid, then system will go through the main process.

First, system needs to trained Model to predict the Movie Reviews in to Positive or Negative.

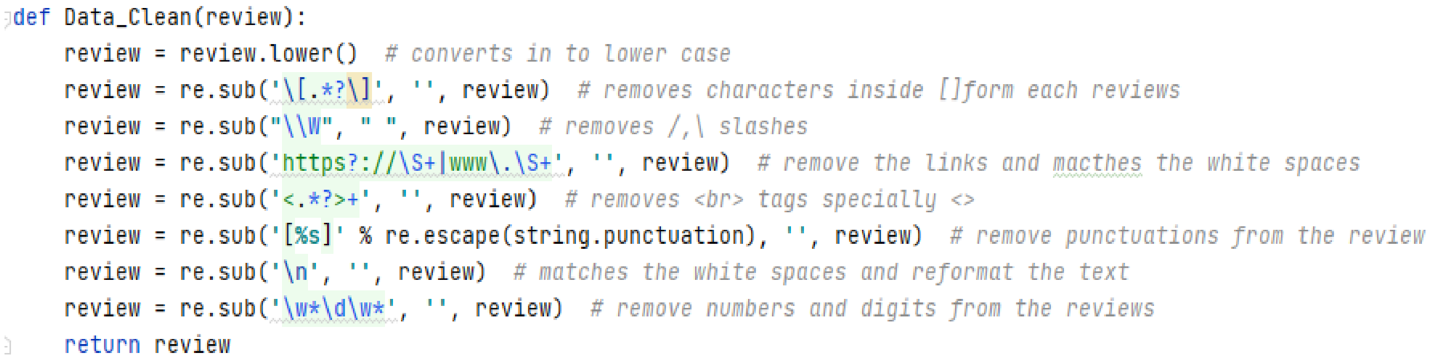


```
@app.route('/train', methods=["GET", "POST"])
def Main_method():
    movie_review_url = request.form['link']
    if (len(movie_review_url) == 102) or (len(movie_review_url) == 101) and movie_review_url.startswith('https://www.imdb.com/title/'):

        # Training the Model
        with open('model_pickle', 'rb') as f:
            mod = pickle.load(f)

        with open('tfidf_vectorizer', 'rb') as f:
            mod_vec = pickle.load(f)
```

Figure 33 Model Training Method in Web Application

As discussed in the methodology part, system will use the previously Trained and Saved Classifier and TF-IDF Vectorizer pickle file to predict the New Reviews.

Data preprocessing is a key stage in Machine Learning because the quality of data and the relevant knowledge that can be learnt from it has a direct impact on the model's capacity to learn. the method system uses to Pre-Process the raw reviews as shown in the below figure. It converts text into an easier-to-understand format, allowing machine learning algorithms to perform better (Singhal,2020).



```
def Data_Clean(review):
    review = review.lower()  # converts in to lower case
    review = re.sub('\[.*?\]', '', review)  # removes characters inside []form each reviews
    review = re.sub("\\W", " ", review)  # removes /,\ slashes
    review = re.sub('https?://\S+|www\.\S+', '', review)  # remove the links and macthes the white spaces
    review = re.sub('<.*?>+', '', review)  # removes <br> tags specially <>
    review = re.sub('[%s]' % re.escape(string.punctuation), '', review)  # remove punctuations from the review
    review = re.sub('\n', '', review)  # matches the white spaces and reformat the text
    review = re.sub('\w*\d\w*', '', review)  # remove numbers and digits from the reviews
    return review
```

Figure 34 Data Preprocessing Method in Web Application

This data preprocessing method works as follows.

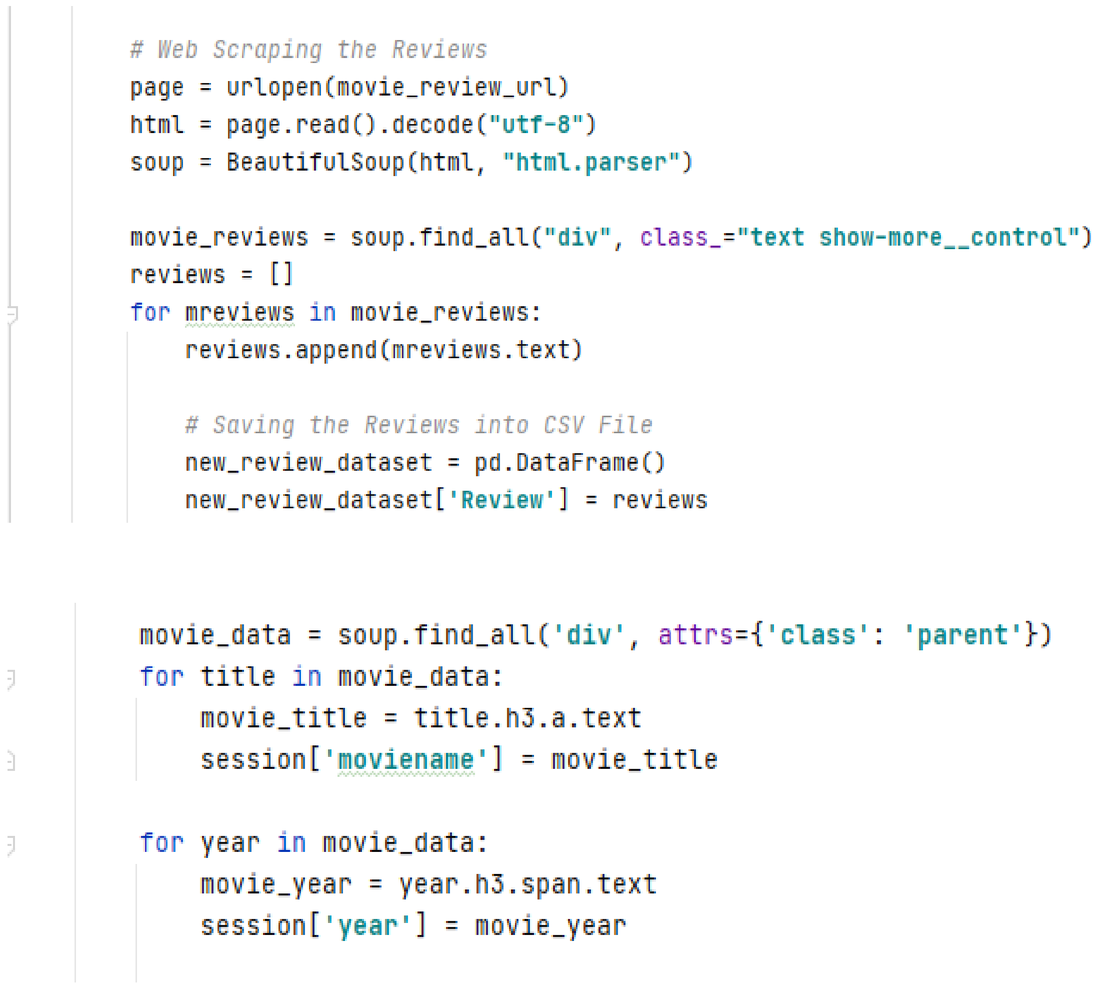
- First method will convert all the characters into lower case characters - During the Machine Learning Process, it aids in maintaining the consistent flow.
- Then Method will check for array lists (Square Brackets) in each review and remove characters inside array lists.
- Then method will remove the front slash and back slash characters from each review.

- Then methods will remove the web URL containing characters which are unwanted characters in machine learning.
- Then method will remove HTML tags such as **<br>** and <>.
- Then system will remove punctuation from each review.
- After that method will matches the white spaces and reformat each review.
- Finally, method will remove all numerical values which are not useful in machine learning modules.

After each review went through the data preprocessing method, now system will have the refined reviews which are not including unwanted characters.

Now system needs to get the reviews from the link which use have entered. To get the new reviews system uses Web Scraping method.

Scraping data from the internet is known as web scraping. System uses BeautifulSoup python library to scrape the reviews. Beautiful Soup is a Python package for extracting data from markup languages such as HTML, XML, and others. Beautiful Soup aids in the extraction of specific material from a webpage, the removal of HTML markup, and the saving of the data (Wieringa, 2020).



```
# Web Scraping the Reviews
page = urlopen(movie_review_url)
html = page.read().decode("utf-8")
soup = BeautifulSoup(html, "html.parser")

movie_reviews = soup.find_all("div", class_="text show-more__control")
reviews = []
for mreviews in movie_reviews:
    reviews.append(mreviews.text)

    # Saving the Reviews into CSV File
    new_review_dataset = pd.DataFrame()
    new_review_dataset['Review'] = reviews
```

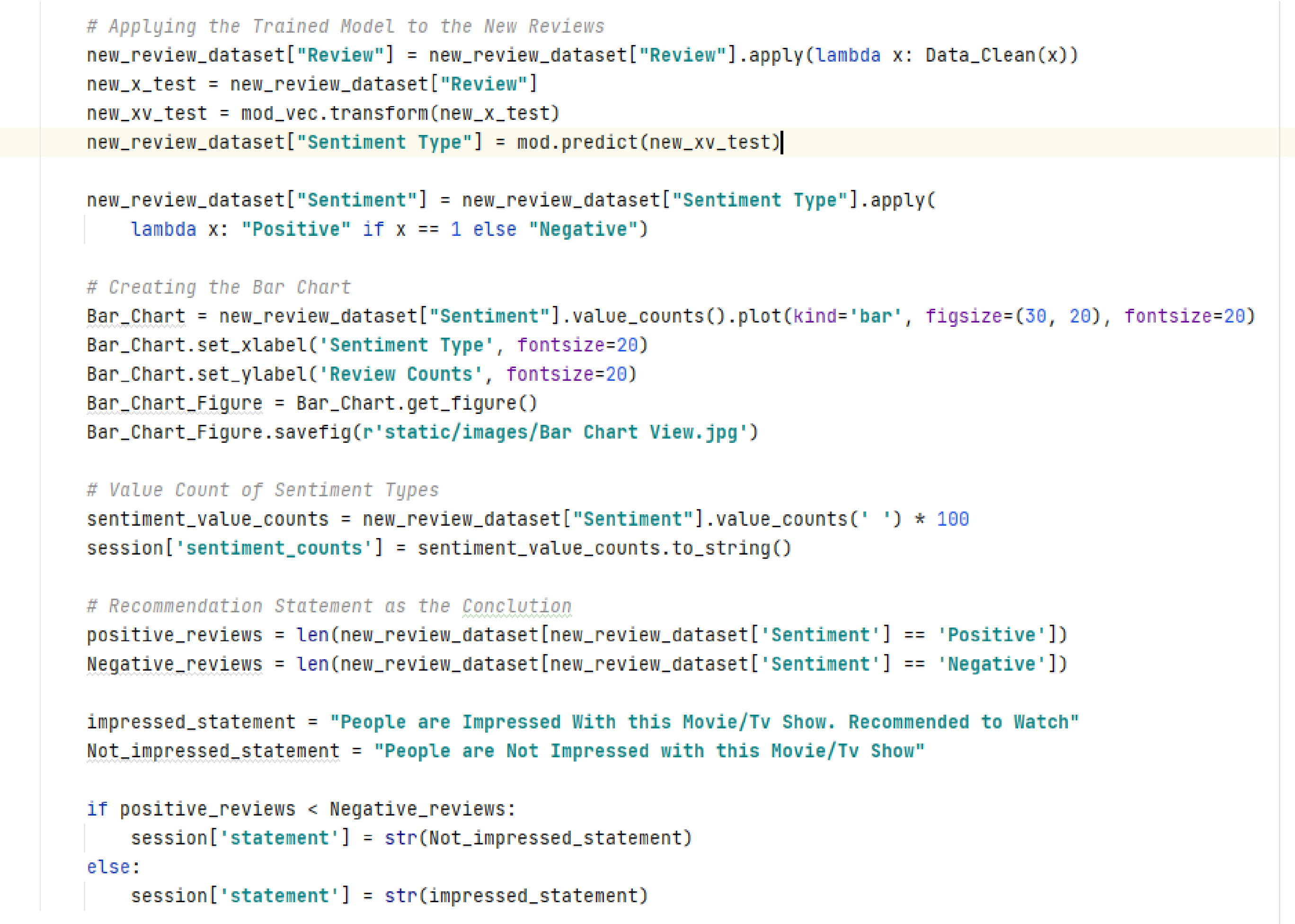
```
movie_data = soup.find_all('div', attrs={'class': 'parent'})
for title in movie_data:
    movie_title = title.h3.a.text
    session['moviename'] = movie_title

for year in movie_data:
    movie_year = year.h3.span.text
    session['year'] = movie_year
```

Figure 36 Web Scraping Method in Web Application

When the user enters the valid URL, system will go through that link and look for the **“div”** class which is named as **“text show more__control”** where all reviews are listed in IMDB web. System will scrape the latest 25 (depends on how many reviews are in the first page) reviews which appears in the first page of selected movie. System will store the scarped reviews into array list and convert the array into text format.

System will also look for **h3** tag which is under class **“parent”** which contains the selected movie and movie year. System will scrape the movie name and year, then converts into text format to parse it to the result page using session.

Now system must apply the trained model to the new scarped reviews. So that system can predict the new reviews into Positive and negative reviews.



```
# Applying the Trained Model to the New Reviews
new_review_dataset["Review"] = new_review_dataset["Review"].apply(lambda x: Data_Clean(x))
new_x_test = new_review_dataset["Review"]
new_xv_test = mod_vec.transform(new_x_test)
new_review_dataset["Sentiment Type"] = mod.predict(new_xv_test)

new_review_dataset["Sentiment"] = new_review_dataset["Sentiment Type"].apply(
    lambda x: "Positive" if x == 1 else "Negative")

# Creating the Bar Chart
Bar_Chart = new_review_dataset["Sentiment"].value_counts().plot(kind='bar', figsize=(30, 20), fontsize=20)
Bar_Chart.set_xlabel('Sentiment Type', fontsize=20)
Bar_Chart.set_ylabel('Review Counts', fontsize=20)
Bar_Chart_Figure = Bar_Chart.get_figure()
Bar_Chart_Figure.savefig(r'static/images/Bar Chart View.jpg')

# Value Count of Sentiment Types
sentiment_value_counts = new_review_dataset["Sentiment"].value_counts(' ') * 100
session['sentiment_counts'] = sentiment_value_counts.to_string()

# Recommendation Statement as the Conclution
positive_reviews = len(new_review_dataset[new_review_dataset['Sentiment'] == 'Positive'])
Negative_reviews = len(new_review_dataset[new_review_dataset['Sentiment'] == 'Negative'])

impressed_statement = "People are Impressed With this Movie/Tv Show. Recommended to Watch"
Not_impressed_statement = "People are Not Impressed with this Movie/Tv Show"

if positive_reviews < Negative_reviews:
    session['statement'] = str(Not_impressed_statement)
else:
    session['statement'] = str(impressed_statement)
```

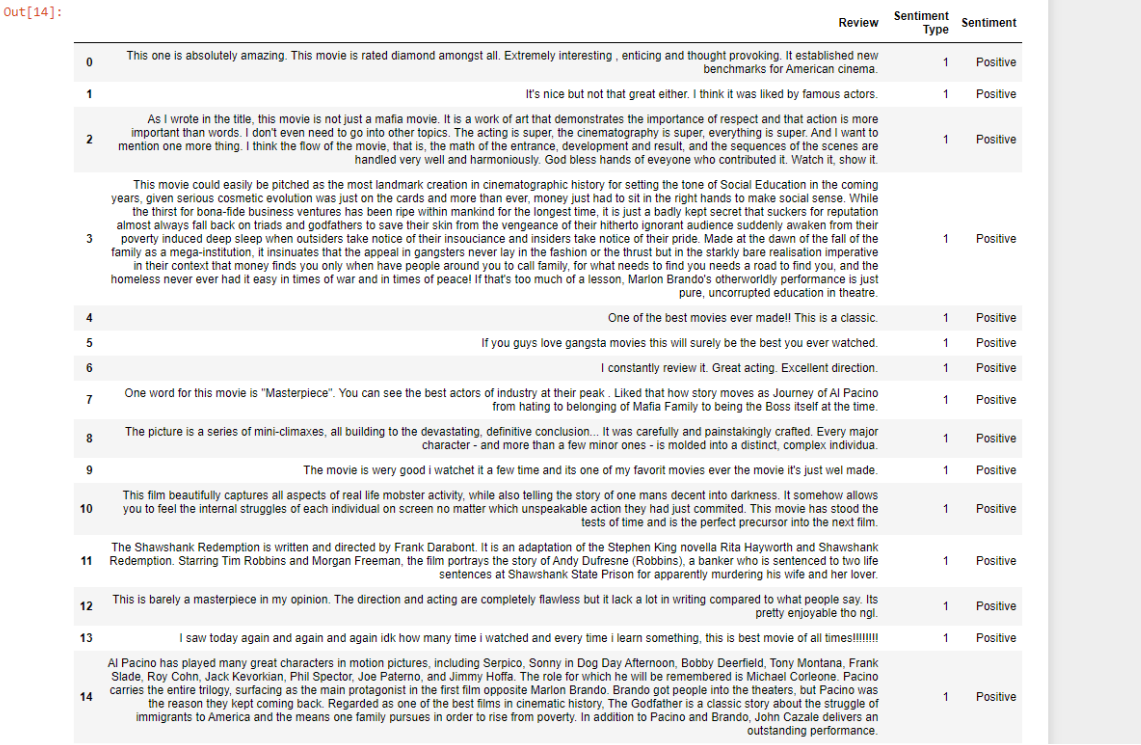
Figure 36 Prediction Method, Bar Chart Creation Method, Percentage Method and Recommendation Status Method in Web Application

First system needs to apply the data preprocessing method to new reviews to get rid of the unwanted characters. After going through the preprocessing method, system will convert raw reviews to preprocessed reviews.

After preprocessing system must vectorize the reviews using TF_IDF vectorizer. To load the saved classifier and vectorizer, system uses the **"pickle.load()"** tool. The review is transformed into a **tf-idf** representation after the vectorizer has been loaded. To make the prediction, the system uses the pickled classifier.

“tfidf.transform()” method will converted the reviews into document-term matrix. For this process system uses the loaded **“mod_vec”** file which contains the TD_IDF Vectorizer.

Finally, system need to predict the vectorized reviews using trained Linear SVC classifier. For this process system uses the loaded **“mod”** file which contains the Previously Trained Machine Learning Model. **“mod”** is the trained model’s classifier Variable. So, **“mod.predict()”** will predict the vectorized reviews in to 1 or 0 with the help of vocabulary which classifier learnt from the trained model. System needs to convert the numerical sentiment types to **“Positive”** or **“Negative”**. If the predicted value is 1, sentiment type will be positive. if the predicted value is 0, sentiment type will be negative. Below figure will explain the outcome of the prediction process in the new reviews.



Out[14]:

	Review	Sentiment Type	Sentiment
0	This one is absolutely amazing. This movie is rated diamond amongst all. Extremely interesting , enticing and thought provoking. It established new benchmarks for American cinema.	1	Positive
1	It's nice but not that great either. I think it was liked by famous actors.	1	Positive
2	As I wrote in the title, this movie is not just a mafia movie. It is a work of art that demonstrates the importance of respect and that action is more important than words. I don't even need to go into other topics. The acting is super, the cinematography is super, everything is super. And I want to mention one more thing. I think the flow of the movie, that is, the math of the entrance, development and result, and the sequences of the scenes are handled very well and harmoniously. God bless hands of eveyone who contributed it. Watch it, show it.	1	Positive
3	This movie could easily be pitched as the most landmark creation in cinematographic history for setting the tone of Social Education in the coming years, given serious cosmetic evolution was just on the cards and more than ever, money just had to sit in the right hands to make social sense. While the thirst for bona-fide business ventures has been ripe within mankind for the longest time, it is just a badly kept secret that suckers for reputation almost always fall back on triads and godfathers to save their skin from the vengeance of their hitherto ignorant audience suddenly awaken from their poverty induced deep sleep when outsiders take notice of their insouciance and insiders take notice of their pride. Made at the dawn of the fall of the family as a mega-institution, it insinuates that the appeal in gangsters never lay in the fashion or the thrust but in the starkly bare realisation imperative in their context that money finds you only when have people around you to call family, for what needs to find you needs a road to find you, and the homeless never ever had it easy in times of war and in times of peace! If that's too much of a lesson, Marlon Brando's otherworldly performance is just pure, uncorrupted education in theatre.	1	Positive
4	One of the best movies ever made!! This is a classic.	1	Positive
5	If you guys love gangsta movies this will surely be the best you ever watched.	1	Positive
6	I constantly review it. Great acting. Excellent direction.	1	Positive
7	One word for this movie is "Masterpiece". You can see the best actors of industry at their peak . Liked that how story moves as Journey of Al Pacino from hating to belonging of Mafia Family to being the Boss itself at the time.	1	Positive
8	The picture is a series of mini-climaxes, all building to the devastating, definitive conclusion... It was carefully and painstakingly crafted. Every major character - and more than a few minor ones - is molded into a distinct, complex individua.	1	Positive
9	The movie is wery good i watchet it a few time and its one of my favorit movies ever the movie it's just wel made.	1	Positive
10	This film beautifully captures all aspects of real life mobster activity, while also telling the story of one mans decent into darkness. It somehow allows you to feel the internal struggles of each individual on screen no matter which unspeakable action they had just commited. This movie has stood the tests of time and is the perfect precursor into the next film.	1	Positive
11	The Shawshank Redemption is written and directed by Frank Darabont. It is an adaptation of the Stephen King novella Rita Hayworth and Shawshank Redemption. Starring Tim Robbins and Morgan Freeman, the film portrays the story of Andy Dufresne (Robbins), a banker who is sentenced to two life sentences at Shawshank State Prison for apparently murdering his wife and her lover.	1	Positive
12	This is barely a masterpiece in my opinion. The direction and acting are completely flawless but it lack a lot in writing compared to what people say. Its pretty enjoyable tho ngl.	1	Positive
13	I saw today again and again and again idk how many time i watched and every time i learn something, this is best movie of all times!!!!!!!	1	Positive
14	Al Pacino has played many great characters in motion pictures, including Serpico, Sonny in Dog Day Afternoon, Bobby Deerfield, Tony Montana, Frank Slade, Roy Cohn, Jack Kevorkian, Phil Spector, Joe Paterno, and Jimmy Hoffa. The role for which he will be remembered is Michael Corleone. Pacino carries the entire trilogy, surfacing as the main protagonist in the first film opposite Marlon Brando. Brando got people into the theaters, but Pacino was the reason they kept coming back. Regarded as one of the best films in cinematic history, The Godfather is a classic story about the struggle of immigrants to America and the means one family pursues in order to rise from poverty. In addition to Pacino and Brando, John Cazale delivers an outstanding performance.	1	Positive

Figure 37 Scraped reviews Prediction outcome Preview

After successfully predicting the new reviews, system must calculate the sentiment types to present to the End user.

System will represent bar chart view to end use by calculating the sentiment type counts. Labels and values are used in a bar graph, where label is the name of the bar and value is the height of the bar. In data analytics, a bar graph is widely used to compare data and extract the most common or highest categories. System will create a bar chart image (.jpg) named **“Bar Chart View”** in **“static/images/”** location. X label of the Bar chart will be the sentiment count which are Positive and Negative. Y label of the Bar Chart will be the number of reviews.

Below figure display a preview of the bar chart view.

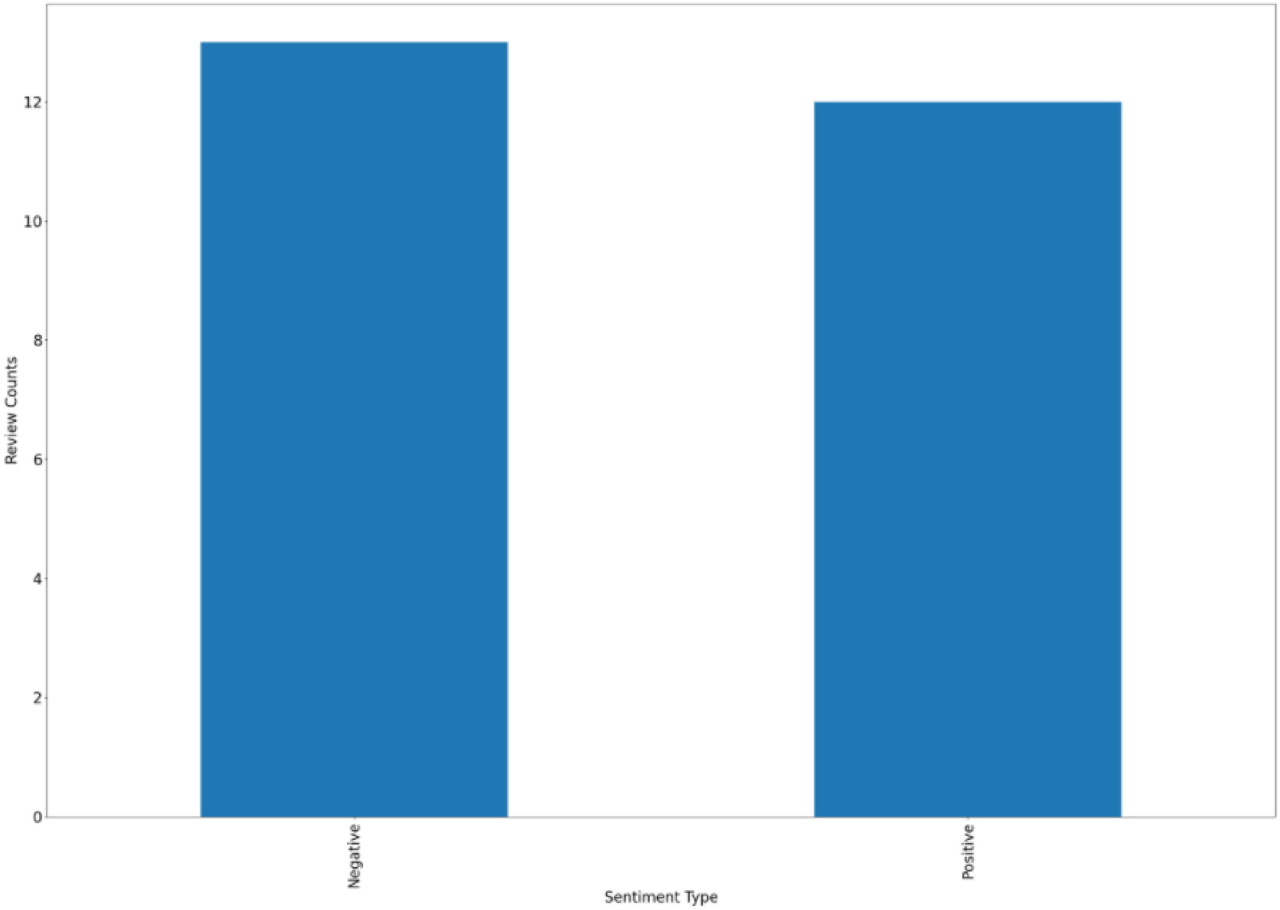
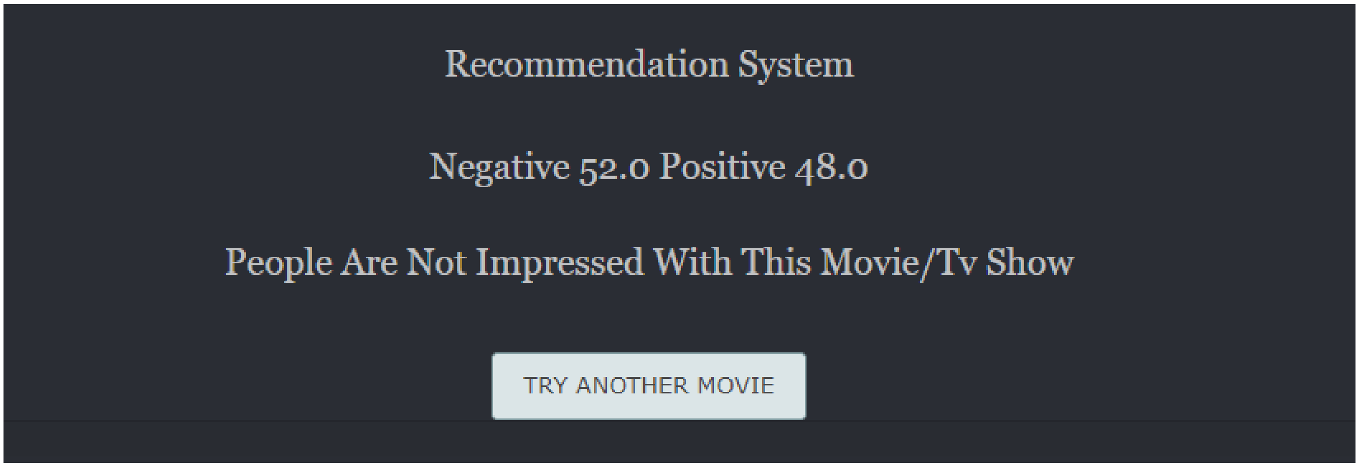


Figure 38 Sentiment Polarity Count Bar Chart Preview

This figure explains that system will give idea about what other people think about this movie/Tv series. End user can clearly get an idea by looking at the bar chart which is display how many Negative reviews and how many positive reviews posted by audience about the selected Movie/Tv Show.

For furthermore clarification of the End user, system will display the percentage amount of each Sentiment Type with a recommendation status. To calculate the percentage value system will use pandas **value_count** function. The Index.value counts() function in Pandas returns an object with counts of unique values. The resulting object will be sorted in descending order, with the first member being the most common. After getting the percentages, system will also count how many positive reviews and Negative reviews. If the positive reviews are higher than negative reviews, system will pass a status using session mentioning **"People are interested with this Movie/Tv Show. Recommended to watch"**. If Negative reviews are higher than positive reviews, system will pass a status using session mentioning **"People are Not Interested with this Movie/Tv Show".**

Below figure displays the above scenario.



Recommendation System

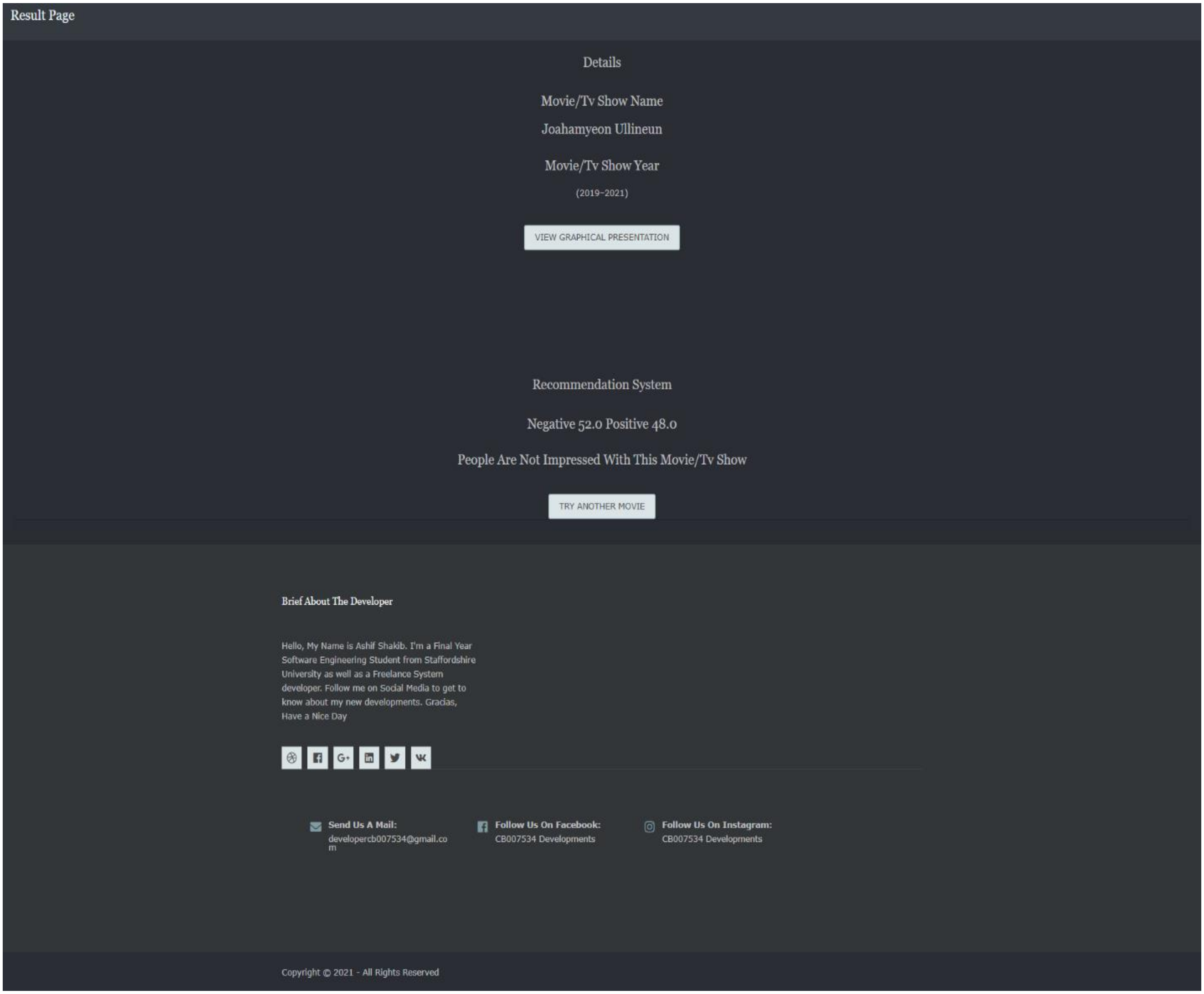
Negative 52.0 Positive 48.0

People Are Not Impressed With This Movie/Tv Show

Figure 39 Percentage and Recommendation Status in Results Page Preview

So, end user can get an even more clear idea about Movie/Tv show. System have now completed the scope and the main purpose of the project. Below figure displays the full view of the result page. As discussed above, system will scrape the movie/Tv show Name and Year and pass it to the HTML like in the below image top container.

As shown in the below figure, system will display do people enjoyed the movie or not. So, end user can come to a conclusion about the movie and decide to watch it or not.



Result Page

Details

Movie/Tv Show Name

Joahamyeon Ullineun

Movie/Tv Show Year

(2019–2021)

Recommendation System

Negative 52.0 Positive 48.0

People Are Not Impressed With This Movie/Tv Show

Brief About The Developer

Hello, My Name is Ashif Shakib. I'm a Final Year Software Engineering Student from Staffordshire University as well as a Freelance System developer. Follow me on Social Media to get to know about my new developments. Gracias, Have a Nice Day

Send Us A Mail:
developercb007534@gmail.com

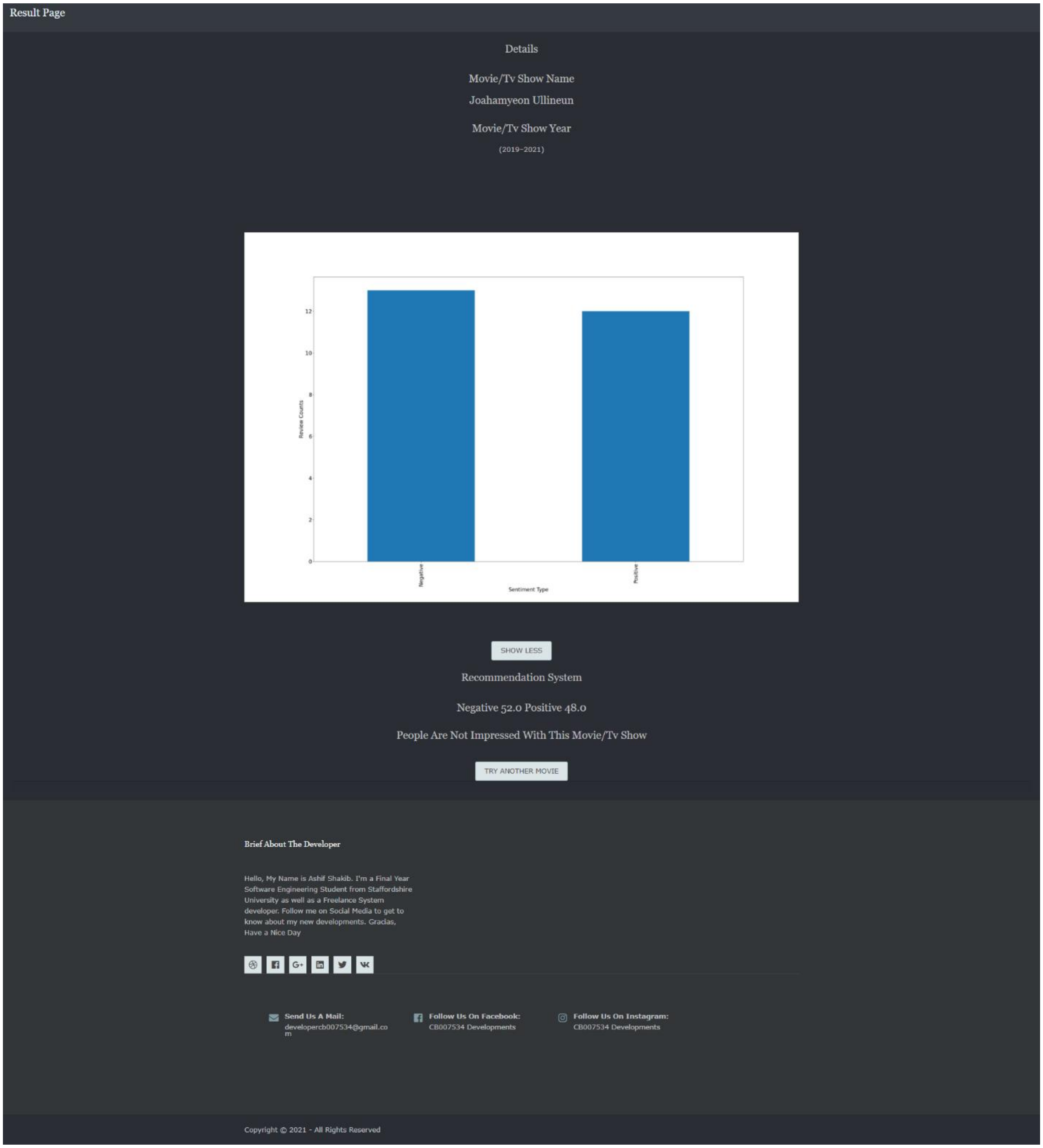
Follow Us On Facebook:
CB007534 Developments

Follow Us On Instagram:
CB007534 Developments

Copyright © 2021 - All Rights Reserved

Figure 40 Results Page in Web Application Preview

When user selects **"View Graphical Presentation"** button, system will display the Bar Chart. Reason for this is bar chart image takes a bit of space in the result page.so to avoid this system will display the bar chart using JavaScript. Above scenario will be displayed from the below figure.



Result Page

Details

Movie/Tv Show Name

Joahamyeon Ullineun

Movie/Tv Show Year

(2019–2021)

Recommendation System

Negative 52.0 Positive 48.0

People Are Not Impressed With This Movie/Tv Show

Brief About The Developer

Hello, My Name is Ashif Shakib. I'm a Final Year Software Engineering Student from Staffordshire University as well as a Freelance System developer. Follow me on Social Media to get to know about my new developments. Gracias, Have a Nice Day

Send Us A Mail: developercb007534@gmail.com

Follow Us On Facebook: CB007534 Developments

Follow Us On Instagram: CB007534 Developments

Copyright © 2021 - All Rights Reserved

Figure 41 Result Page with Bar Chart View

When user selects **“Show Less”** button system will hide the bar chart view and will display the textual data presentation. When user selects **“Try Another Movie”** button, system will direct the user to home page. After discussing about every method of implementation, document will end the discussion about System Implementation from here.

Chapter 09: Testing and Evaluation

Testing Strategy

Software testing is a technique for determining whether the actual software product meets the specified requirements and ensuring that it is defect-free. It entails the use of manual or automated methods to evaluate one or more properties of interest by executing software/system components. In contrast to actual requirements, software testing's goal is to find mistakes, gaps, and missing requirements. Arriving to the testing stage of the IMDB Movie Review Analyzing System online development process, this stage is all about certifying the software's quality.

- Software testing is essential for identifying problems and errors that occurred throughout the development process.
- It is critical since it ensures that the customer trusts the company and that their contentment with the application is maintained.
- It is critical to ensure the product's quality. Delivering a high-quality goods to customers aids in acquiring their trust.

Testing is usually divided into three areas.

1. Functional Testing
2. Non-Functional Testing (Performance Testing)
3. Maintenance

Functional Testing

The testing of the functional features of a software application is known as functional testing. When executing functional tests, the testing team must test each and every feature. They must determine whether or not they are achieving the desired outcomes (Hossain,2021). Functional testing can take several forms, including:

- Unit testing
- Integration testing
- End-to-end testing
- Smoke testing
- Sanity testing
- Regression testing
- Acceptance testing
- White box testing
- Black box testing
- Interface testing (Hossain,2021)

Functional testing for the IMDB movie review analyzing system online is carried out by doing tests on each of the web application's functions. To test the results of each function, multiple inputs are provided. The percentage calculated based on the number of passes and fails is used to determine the success rate. All the test suites with results are included under Testing and results part of the document.

Non – Functional Testing (Performance Testing)

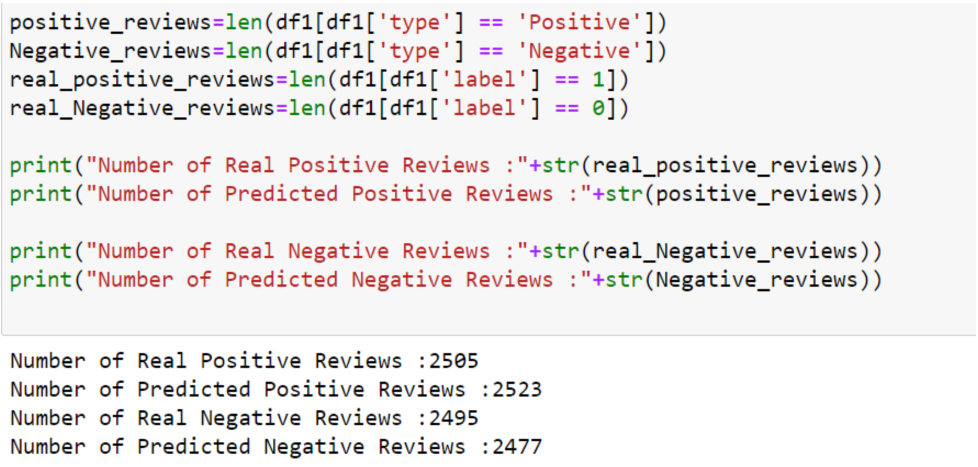
Non-functional testing involves evaluating features of an application that are not functional, such as performance, reliability, usability, security, and so on. After the functional testing, non-functional tests are conducted. Developers can greatly increase the quality of software by using non-functional testing. Although functional tests increase quality, non-functional tests allow developers to improve the product even further. Non-functional testing allows programmers to fine-tune their work. This type of testing is not concerned with whether or not the software is functional. It is more about how well the software performs and a variety of other factors (Hossain,2021).

System will use performance testing to test the performance task which happens in the Web Application. All the test suites with results are included under Testing and results part of the document.

Accuracy Testing

Accuracy testing comes under Performance testing. This part of the document will discuss about the Accuracy Testing of the Trained Machine Learning Model. Main Scope of this Testing is to check How well the Trained Machine Learning Model will perform in Sentiment Detection process. This testing will explain how accurate the sentiment detection process is using the trained Model.

For Accuracy testing part, Trained Machine Learning model has been tested with Testing Dataset which is collected from Kaggle. This training dataset has 5000 of records and 2 Columns which contains the reviews and the Polarity. This dataset will be more suitable because after testing, developer can check the predicted polarity with the original labeled polarity. Sentiment Detection testing results are explained below.

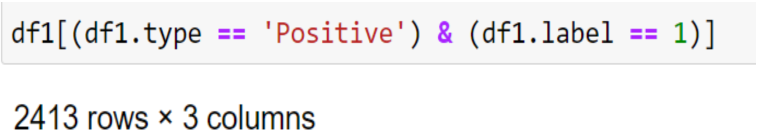


```
positive_reviews=len(df1[df1['type'] == 'Positive'])
Negative_reviews=len(df1[df1['type'] == 'Negative'])
real_positive_reviews=len(df1[df1['label'] == 1])
real_Negative_reviews=len(df1[df1['label'] == 0])

print("Number of Real Positive Reviews :"+str(real_positive_reviews))
print("Number of Predicted Positive Reviews :"+str(positive_reviews))

print("Number of Real Negative Reviews :"+str(real_Negative_reviews))
print("Number of Predicted Negative Reviews :"+str(Negative_reviews))
```

```
Number of Real Positive Reviews :2505
Number of Predicted Positive Reviews :2523
Number of Real Negative Reviews :2495
Number of Predicted Negative Reviews :2477
```

Figure 42 Accuracy Testing Preview

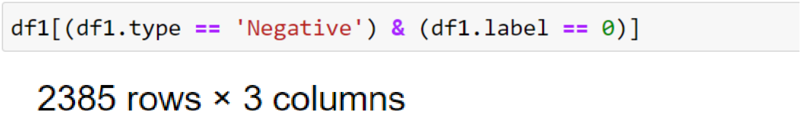
In the Training Dataset, it has 2505 records of Real Positive Reviews and 2495 records of Negative Reviews. Out of the 2505 of Positive reviews Model has predicted 2523 reviews as Positive and out of 2495 of Negative reviews Model has predicted 2477 as Negative Reviews.



```
df1[(df1.type == 'Positive') & (df1.label == 1)]
```

2413 rows × 3 columns

Figure 43 Number of Correctly Predicted Positive Reviews



```
df1[(df1.type == 'Negative') & (df1.label == 0)]
```

2385 rows × 3 columns

Figure 44 Number of Correctly Predicted Negative Reviews

Scenario	**Fraction**
Correctly predicted Positive Reviews	2413/5000
Correctly predicted Negative Reviews	2385/5000
Total Correctly Predicted Reviews	4798/5000

Conclusion of the Accuracy testing will contain in the conclusion part of the document.

Testing and Results

In order to assess if a feature of an application is performing correctly, a test case comprises components that define input, action, and an expected result. A test case is a series of instructions on "HOW" to validate a specific test objective/target, which when followed will inform us whether the system's expected behavior is met or not. This part of the document will contain the All the test cases of Functional and Non-Functional Testing.

Functional Testing Test Cases

ID	Description	Input	Expected Output	Actual Output	Result
001	To view the instructions page from homepage, User must select View instructions from Navigation Bar	Select View Instructions	System will Display Instructions page.	System Displayed Instructions page.	Pass
002	To view the Video demo page from homepage, User must select View Video Demo from Navigation Bar	Select View Video Demo	System will Display Video demo page.	System Displayed Video demo page.	Pass
003	To view the home page from Instructions Page, User must select Home from Navigation Bar	Select Home	System will Display homepage.	System Displayed homepage.	Pass
004	To view the video demo page from Instructions Page, User must select View Video demo from Navigation Bar	Select View Video Demo	System will Display video demo page.	System Displayed video demo page.	Pass
005	To view the home page from video demo page, User must select home from Navigation Bar	Select Home	System will Display Homepage.	System Displayed Homepage.	Pass
006	To view the instructions page from video demo page, User must select View instructions from Navigation Bar	Select View Instructions	System will Display Instructions page.	System Displayed Instructions page.	Pass

007	To view the instructions page from Invalid URL page, User must select View instructions from Navigation Bar	Select View Instructions	System will Display Instructions page.	System Displayed Instructions page.	Pass
008	To view the homepage from Invalid URL Page, User must select Home from Navigation Bar	Select Home	System will Display homepage.	System Displayed homepage.	Pass
009	To view the Video Demo page from Invalid URL Page, User must select Video Demo from Navigation Bar	Select View Video Demo	System will Display Video demo page.	System Displayed video demo page.	Pass
010	To return to home page from result page, user must select Try Another Movie button	Select Try Another Movie	System will Display homepage.	System Displayed homepage.	Pass
011	To play the video in video demo page, user must select play button	Select play button	System will play the video	system played the video	Pass
012	If the submitted URL is valid, then system will start the process and will display the result page.	Enter Valid URL and submit	System will start the process and display the result page	System started the process and displayed result page	Pass
013	If the submitted URL is invalid, then system will display the invalid URL page.	Enter invalid URL and submit	System will display invalid URL page	System displayed invalid URL page	Pass
014	System will check the validity of the URL by checking the length of it. If the length of URL is 101 or 102, system will valid it as valid URL.	Enter random string value which has a length of 5	System will display invalid URL page	System displayed invalid URL page	Pass

015	System will check the validity of the URL by checking the length of it. If the length of URL is 101 or 102 and the URL starts with ‘https://www.imdb.com/title/’ format, then system will valid it as valid URL.	Enter random string value which has a length of 101	System will display invalid URL page	System will display invalid URL page	Pass
016	System will check the validity of the URL by checking the length of it. If the length of URL is 101 or 102 and the URL starts with ‘https://www.imdb.com/title/’ format, then system will valid it as valid URL.	Enter random string value which has a length of 102	System will display invalid URL page	System will display invalid URL page	Pass
017	System will check the validity of the URL by checking the length of it. If the length of URL is 101 or 102 and the URL starts with ‘https://www.imdb.com/title/’ format, then system will valid it as valid URL.	Enter URL starts with ‘https://www.imdb.com/title/’ and rest 75 random characters	System will display HTTP error message.	System displayed “urllib.error.HTTPError” and system crashed	Fail
018	System will check the validity of the URL by checking the length of it. If the length of URL is 101 or 102 and the URL starts with ‘https://www.imdb.com/title/’ format, then system will valid it as valid URL.	Enter URL which is not starts with ‘https://www.imdb.com/title/’ and rest 75 random characters	System will display invalid URL pages	system displayed invalid URL page	Pass
019	When user selects “view graphical presentation” button, system will display the bar chart.	Select “view graphical presentation” button.	System will display the bar chart.	System displayed the Bar chart.	Pass

020	When user selects “Show Less” button, system will hide the bar chart.	Select “Show Less” button.	System will hide the bar chart.	System hided the bar chart.	Pass

<u>Non – Functional Testing Test cases</u>

ID	Description	Input	Expected Output	Actual Output	Result
001	Training dataset has 50000 records. So, training the model will take some time.	Enter the Valid URL and Submit	System will take more than 60 seconds to train the model.	Took 58 seconds to train the model.	Pass
002	Scraping the Reviews from the Submitted URL.	Enter the Valid URL and Submit	System will extract the available reviews in the first page.	System extracted the available reviews in the first.	Pass
003	System must scrape the reviews from the provided URL. So, scraping the reviews (25 Reviews) might take some time.	Enter the Valid URL and Submit	System will take 8 seconds to Scrape the reviews	Took 6 Seconds and 80 Milliseconds to scrape the reviews.	Pass
004	After submitting the URL, system have to some process to complete. So, to present the results page, System will take some time.	Enter the Valid URL and Submit	System will take 2 minute to display the result page.	System took 9 seconds 63 Milliseconds to display the result page.	Pass

005	To display the Bar chart view in the results page, System must create a bar charts image using the number of predicted Positive and Negative reviews.	Enter the Valid URL and Submit	System will Create a bar chart image.	System created a Bar Chart Image	Pass

Conclusions

<u>Function Testing Conclusion</u>

All the functional testing are done in every aspect and system responded to each testing very well. System has been tested with 20 testing environments and out of 20 testing system has failed only one testing environment. In terms of percentages, System has passed **95%** of tests while failing **5%** (Only one).

Only testing environment that system failed is testing the URL with an invalid format. The required URL starts with **‘https://www.imdb.com/title/’** and length of the URL will be 101 characters or 102 characters. So, these are the validations that system will check for the URL validation. This testing environment tested with randomly generated string characters which starts with ‘**https://www.imdb.com/title/**’. This starting format contains 27 characters. Another 75 auto generated string characters combined with the format and tested the system. Unfortunately, system crashed because even though the system validated as a valid URL, **“urlopen”** library which is using to open the submitted URL does not recognize the URL. But this issue will not happen in the real-world usage because main purpose of using this system is to get a conclusion about a movie or a TV show. So, end user will properly submit the valid format. Even the URL is invalid beside the format that discussed above, system will direct the end user to the Invalid URL page.

In conclusion, Beside the above scenario systems Functional aspects are functioning properly.

Non-Functional Testing Conclusion

Performance testing of this system tested with 5 testing environments. Overall performance of the system is performing well but the only drawback is each process take some time and it will cost an amount of time to display the result page. When End user submits the URL and it will take 10 seconds to display the results page.

After that System have to scrape the reviews from the provided URL. Scraping process will take around 7 seconds to complete. This fact proves that **BeautifulSoup** python library does a good job in web scaping processes. Even though the process takes some time to complete and present the results page, each non-functional process performances perfectly without any errors.

In conclusion, system has passed each non-functional testing. This fact proves that Training Model process, Web Scarping process, prediction process and data presentation process of predicted reviews performs properly without any errors.

Trained model is also performing very well in the prediction process. As discussed in the Performance training part of the document, model has tested with training dataset which have 5000 records. Out of the 5000 records model has successfully predicted 2413 positive reviews and 2385 Negative reviews correctly. It has **47.7%** Negative reviews prediction percentage and **48.26%** Positive reviews prediction percentage. So overall Model has **95.96%** successful review prediction percentage. This fact proves that Linear SVC Classifier does a great job in Sentiment Analysis process.

Chapter 10: Critical Evaluation

Discussion

The “IMDB Movie Review Analyzing System” is a program that analyzes movie or television show reviews on IMDB and predicts whether the film or show will wow the audience. The majority of people watch YouTube trailers before watching a movie or a TV show. Even if people watch the video, the trailer does not offer them a good picture of what the movie/TV show is about. Some people read reviews in various movie reviews containing platform such as IMDB, Rotten Tomatoes and various social media platforms. Reading that much of reviews will be a time-wasting process.

Occasionally, the trailer is intriguing, but the film or television show is not. This system will be the greatest answer for this problem. This system will read the latest IMDB reviews for the movie or TV show that the user wishes to watch and will advise the user whether or not to watch it. this system is very helpful for get a decision about movies and shows as well as Process of this system works faster than reading reviews and watching trailers. imagine User read bunch of reviews to come to a conclusion and then decides to not to watch the movie. it will be a waste of time. instead of that, user just have to paste the movie review link and system will does the rest for the end user.

Not only that, but system will also display graphical and textual representations about the selected movie/Tv show. Graphical presentation (Bar Chart) will be a huge advantage for the user, because Bar graphs are a useful tool for comparing items in different groupings. In presentations and reports, bar graphs are a very effective graphic. They are popular because they make it much easier for the reader to spot patterns or trends than a table of numerical data.

Next big advantage is Textual representations. System will calculate the reviews categories and offers user the percentage of Positive Reviews and Negative reviews. system not going to finish the textual representation from percentages. system will also present a Recommendation Status about the selected movie mentioning people are impressed with this movie/Tv show or not. So, this will be easy and best solution for the above discussed situations.

The Knowledge Gained

This project has been beneficial in terms of identifying some completely new areas of research, forcing me to break free from the limited scope of recurring implementations of the same type of software during the course of my academic career. This part of the document contains what are the new techniques I've gained knowledge about.

- Machine Learning

  All businesses rely on data to function. Data-driven decisions are increasingly determining whether a company keeps up with the competition or falls further behind. Machine learning has the potential to unlock the value of corporate and consumer data and enable companies to make decisions that keep them ahead of the competition. Machine Learning (or Deep Learning, AI, Data Science, and Computer Science in general) is the finest field to be in today. This is true in a variety of subjects, including mathematics and computer science. There are unlimited methods to tackle issues and infinite concepts to try and research in Machine Learning. Today, we have systems that can translate our language into something that can assist us with a variety of activities such as translation, question answering, classification, and more! Many of us take what machines do for us for granted. In many other fields, machines solve many of our problems. Being at the center of this change presents us with some of the most intriguing difficulties.

- Python

  Python has quickly risen to prominence among programmers and technologists. Python is quickly becoming the most popular language of choice in high-income countries, according to Stack Overflow question views (Kan,2018). In domains as diverse as bioinformatics, data science, machine learning, and even astronomy, popular Python libraries are widely integrated and used. I was able to teach myself Python within weeks of researching it after learning C++, C#, and Java as my initial programming languages.

- Data science

  I have gained a lot of knowledge about Data science field with the research for this project. Data Science is a collection of tools, algorithms, and machine learning techniques aimed at uncovering hidden patterns in large amounts of data. Data Science and Machine Learning are enabling Big Data to impact our world in amazing ways, at least as much as the industrial revolution did. Data science, in my opinion, is the future of data.

- NLP – Natural Language processing

  Despite the fact that this falls within the above machine learning part, NLP should be highlighted separately because the prediction of reviews into Positive and Negative reviews led to the realization of the NLP aspect of Machine Learning. Natural Language Processing (NLP) is an artificial intelligence subfield that helps computers interpret human speech. This was also a completely new technology, and the sentiment analysis technique, which falls under the NLP techniques, was to be used to meet the project's requirements. The sentiment analysis research aided in gaining a more balanced understanding of how machine learning is involved with sentiment analysis and how human languages are taught to computers, resulting in amazing services like Apple's Siri and Google Assistant, which assist people in the modern world with their daily tasks.

Future Work

Future of this project will eb focused on two things. Two things will be discussed below.

Display Poster of the Selected Movie in the Result Page

Future of this application will be focusing on Improving the Result Page. To make the result page more user friendly, developer will focus on display a poster of the Selected Movie/TV show in the Result Page. Currently developer have tested a above function and the drawback is the Poster which IMDB are using is very small in size. Because of the small size of the image, the developer is unable to resize it. The image will become blurry if the developer makes the poster larger in the HTML page. So, future of the web application will be focusing on accomplishing the scope somehow.

Scarping More Reviews

Current process will web scrape only the reviews which are in the first page. Reason for that is IMDB Movie review page is based on java script and rest of the reviews are contain in next pages There will be only 25 reviews in the first page. A powerful model which is trained with 50000 records applying for just 25 reviews will be a disadvantage. So, in the future system will focus on scraping all the reviews using java script scarping techniques. If the system scrapes all the reviews, then system can calculate more reviews for the final data presentation.

Reference

Luashchuk, A (2019). *Why I Think Python is Perfect for Machine Learning and Artificial Intelligence*.[Online] Available from: https://towardsdatascience.com/8-reasons-why-python-is-good-for-artificial-intelligence-and-machine-learning-4a23f6bed2e6.[Accessed: 29th May 2019].

Solutions, M (2018). *What is PyCharm IDE?* [Online] Available from: https://medium.com/@mindfiresolutions.usa/what-is-pycharm-ide-cc0735784f64 [Accessed: 17th April 2018].

Lewis, R (2017). *IMDB Web site* [Online] Available from: https://www.britannica.com/topic/IMDb [Accessed: 31st October 2017].

Brownlee, J (2020). *Train-Test Split for Evaluating Machine Learning Algorithms*. [Online] Available from: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/ [Accessed :24th July 2020].

Donna (2015). *The Influence of Online Film reviews*. [Online] Available from: https://www.franklymydearuk.co.uk/the-influence-of-online-film-reviews/ [Accessed: 25th June 2015].

Shalini (2016). *Social Media Marketing Case Study*. [Online] Available from: https://www.digitalvidya.com/blog/how-facebook-influences-movie-promotion-and-its-success/ [Accessed: 29th November 2016].

Algorithmia (2018). *Introduction to sentiment analysis: What is sentiment analysis?* [Online] Available from: https://algorithmia.com/blog/introduction-sentiment-analysis [Accessed: 26th March 2018].

IBM Clod Education (2020). *Machine Learning* [Online] Available from: https://www.ibm.com/cloud/learn/machine-learning [Accessed: 15th July 2020].

Gibbs, F and Elwert, F (2020). *Intro to Beautiful Soup* [Online] Available from: https://programminghistorian.org/en/lessons/intro-to-beautiful-soup [Accessed: 12th May 2020].

Breuss, M (2021). *Python Web Applications: Deploy Your Script as a Flask App*. [Online] Available from: https://realpython.com/python-web-applications/ [Accessed: 1st February 2021].

Gonfalonieri, A (2019). *How to Build A Data Set For Your Machine Learning Project*. [Online] Available from: https://towardsdatascience.com/how-to-build-a-data-set-for-your-machine-learning-project-5b3b871881ac [Accessed: 14th February 2019]

Gupta, K (2019). *An Introduction to TF-IDF*. [Online] Available from: https://medium.com/analytics-vidhya/an-introduction-to-tf-idf-using-python-5f9d1a343f77 [Accessed: 9th November 2019].

Rajkumar (2021). *What Is Software Testing | Everything You Should Know*. [Online] Available from: https://www.softwaretestingmaterial.com/software-testing/ [Accessed: 9th January 2021].

Cassidy, K (N/A). *Film history: the evolution of film and television*. [Online] Available from: https://www.videomaker.com/how-to/shooting/film-history-the-evolution-of-film-and-television/ [Accessed: N/A]

Nambiar, M (2019). *Many People Are Spending Their Free Time Watching Movies | Band 8 IELTS Essay Sample*. [Online] Available from: https://www.ielts-practice.org/many-people-are-spending-their-free-time-watching-movies-band-8-ielts-essay-sample/ [Accessed in: 27th March 2019].

Wolff, R (2020). *Sentiment Analysis with Machine Learning: Process & Tutorial* [Online] Available from: https://monkeylearn.com/blog/sentiment-analysis-machine-learning/ [Accessed in :2020].

RealnReel Team (2018). *Strategies For Movie Promotion And Marketing Using YouTube* [Online] Available from: https://www.reelnreel.com/strategies-for-movie-promotion-and-marketing-using-youtube/ [Accessed in: 28th May 2018]

Stegner, B (2020). *IMDb vs. Rotten Tomatoes vs. Metacritic: Which Movie Ratings Site Is Best?* [Online] Available from: https://www.makeuseof.com/tag/best-movie-ratings-sites/ [Accessed in: 24th August 2020]

Asiri, S (2018). *Machine Learning Classifiers.* [Online] Available from: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623 [Accessed in: 11th June 2018]

Chen, S (2020). *Getting Started with Text Vectorization*. [Online] Available from: https://towardsdatascience.com/getting-started-with-text-vectorization-2f2efbec6685 [Accessed in: 2nd April 2020]

Perez, M (2019). *What is Web Scraping and What is it Used For?* [Online] Available from: https://www.parsehub.com/blog/what-is-web-scraping/ [Accessed in: 6th August 2019]

Pandey, P (2019). *Data Preprocessing: Concepts* [Online] Available from: https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825 [Accessed in: 25th November 2019]

Singh, V (2021). *What is a Framework?* [Definition] Types of Frameworks [Online] Available from: https://hackr.io/blog/what-is-frameworks [Accessed in: 15th May 2021]

Hossain, A (2021). *Types of Software Testing* [Online] Available from: https://hackr.io/blog/types-of-software-testing [Accessed in: 13th May 2021]

Kan, E (2018). *Five Python Tricks You Need to Know Today* [Online] Available from: https://towardsdatascience.com/five-python-tricks-you-need-to-learn-today-9dbe03c790ab [Accessed in: 16th August 2018]

Singh, V., Mahajan, A., and Chaudhary, D (2020*). Sentimental Analysis of Hotel Reviews from TripAdvisor*. [Online] Volume:07 Issue: 06. ABES Engineering College, Ghaziabad, India. Available from: https://www.irjet.net/archives/V7/i6/IRJET-V7I6365.pdf [Accessed Date: 6th June 2020]

Baid, p., Chaplot, N., and Gupta, A (2017). *Sentiment Analysis of Movie Reviews Using Machine Learning Techniques*. [Online] Volume:179 No: 07. Jaipur Engineering College and Research Center Jaipur, Rajasthan, India. Available from: https://www.researchgate.net/publication/321843804_Sentiment_Analysis_of_Movie_Reviews_using_Machine_Learning_Techniques [Accessed Date: Dec 2017]

Amolik, A. et al (2020). *Twitter Sentiment Analysis of Movie Reviews Using Machine Learning Techniques*. [Online] Volume:07 No: 06. School of Computer Science and Engineering,VIT University, Vellore-632014, Tamilnadu, India Available from: https://www.researchgate.net/publication/291837156_Twitter_Sentiment_Analysis_of_Movie_Reviews_using_Machine_Learning_Techniques [Accessed Date: January 2016]

Mamtesh. And Mehla, S. (2020). *Sentiment Analysis of Movie Reviews Using Machine Learning Classifiers*. [Online] Volume:182 No: 50. National Institute of Technology Kurukshetra, India. Available from: https://www.ijcaonline.org/archives/volume182/number50/mamtesh-2019-ijca-918756.pdf [Accessed Date: April 2019]

GeeksForGeeks (2018). *Python | Lemmatization with NLTK* [Online] Available from: https://www.geeksforgeeks.org/python-lemmatization-with-nltk/ [Accessed Date: 6th November 2018]

GeeksForGeeks (2019). *ML | Using SVM to perform classification on a non-linear dataset* [Online] Available from: https://www.geeksforgeeks.org/ml-using-svm-to-perform-classification-on-a-non-linear-dataset/ [Accessed Date: 15th January 2019]

GeeksForGeeks (2020). *Splitting Data for Machine Learning Models* [Online] Available from: https://www.geeksforgeeks.org/splitting-data-for-machine-learning-models/ [Accessed Date: 20th August 2019]

Mesevage, T (2020). *Top Machine Learning Algorithms Explained: How Do They Work?* [Online] Available from: https://monkeylearn.com/blog/machine-learning-algorithms/ [Accessed Date: 29th August 2020]

Appendix

This part of the document discusses about the research that done to find out what are the things that people consider before watching or purchasing a Movie/TV show.

Research 01 (Online Research)

This research is done using online research papers which are completed using Online surveys on social medias platforms.

The excitement of watching a movie begins with the selection of the film. Because there are so many movies available on the market today, it might be tough to make a decision. This is due to the fact that the websites have a large selection of movies to choose from. There are so many movies to pick from that making a decision gets tough.

According to the survey, things that people mostly consider before watching a movie/TV show are listed and explained below.

- Read Online Reviews about the movie

  Reviews are crucial, especially when deciding whether or not a film is worth watching. This is where people find all of the reviews for a certain film. Individuals' critics are definitely from those who have already seen the film. People will learn about the plot of the film through the reviews, and then read different people's opinions on the film.

- Learn from Social media community such as Facebook Movie Groups

  The internet is a formidable tool. People are more likely to debate and voice their opinions about a certain movie or film in online movie forums.

  Through these forums, you will learn about the most popular films as well as the differences between modern and classic films. You will eventually learn about popular films as a result of these exchanges.

- The Movie Actors

  Some judgments about whether or not to watch a film are influenced by the popularity of the actors in the film. If you like a film because of a certain actor, you will seek other films that include that actor. When an artist is well-known, the urge to see the film is even stronger.

  Good artists are frequently associated with excellent films in the minds of moviegoers. As a result, when looking for a movie to watch, the actors in the film should also be considered.

- <u>The story's beginnings</u>

  The type of movies that people watch is also determined by the film's origin. There are those who enjoy viewing Mexican films, others who enjoy American films, and still others who enjoy films from the Philippines. These are just a few of the countries that are well-known for creating excellent films. People often begin their hunt for movies by going to the source of the film.

- <u>Recommendations</u>

  Many people are introduced to movies through word of mouth. This occurs after someone has become enamored with a film. They propose watching the same movies to others in order to pass on their experience. If someone recommends a film to you because they thought it was intriguing, you should consider seeing it.

- <u>Film genre</u>

  Every type of film does not appeal to everyone. At the end of the day, everyone has a favorite type of film. We have, for example, scientific films and investigative films, among other genres. When it comes to watching movies, a person's taste speaks volumes. People often link themselves with the film genres that they enjoy the most.

- <u>The prevailing atmosphere</u>

  The state of mind is equally critical. Everyone has a specific mood with which they would like to attach themselves with a film. When you want to get into an exciting mood, you can hunt for an interesting movie to watch. Reading other people's reviews can help you reach the mood factor. That way, you will be able to discern right away what kind of mood a film is in.

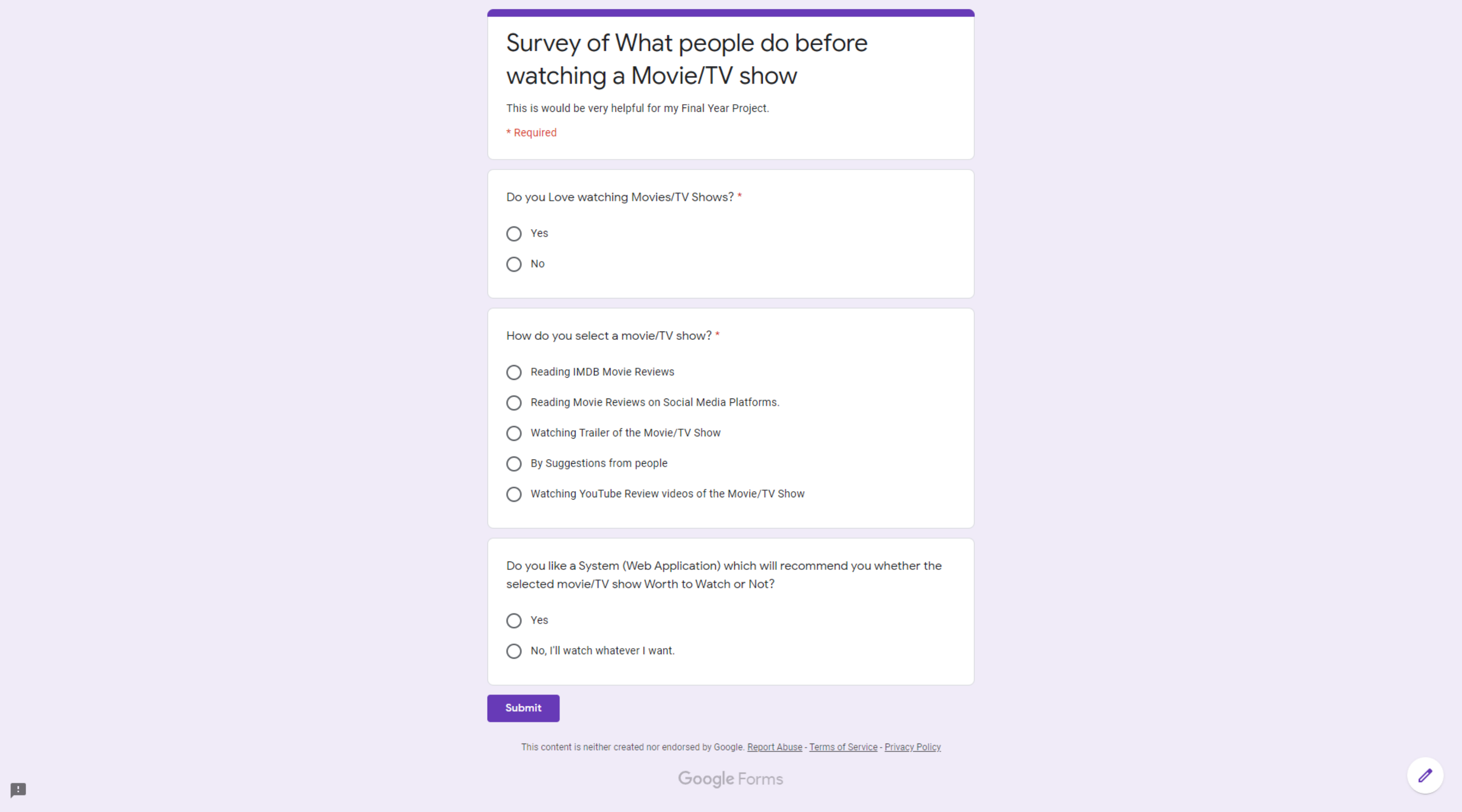
- <u>the duration of the film</u>

  Time may appear to be an inconsequential aspect, yet it is one that must be considered. We have a variety of films to choose from, all with varying lengths. There are films that take longer to see than others. Depending on the amount of time available, some people enjoy extended movies while others prefer short films. Some people are unable to sit for lengthy periods of time and watch movies. As a result, some people prefer series films while others prefer single films.

When it comes to watching movies, everyone has their own preferences. People's movie preferences are influenced by a variety of things. People do not just go around watching random movies. Many of them look for the genre of film they like, the country of origin, the length of the film, the actors in the film, and a variety of other things. According to above research maximum people are fixed with Reading Online movie reviews in Movie review platforms. Because in online movie reviews platforms there will so many reviews regarding the selected movie, and it will be easy to read the reviews. In conclusion, solution must focus on the Online movie reviews platforms and must do more research to select which is the better online movie review platform to get data source to develop the system.

<u>Research 02 (Online Survey)</u>

This research is done using online Survey using Google Forums. To get the peoples opinion about What they consider before watching a movie and how they select a movie, I created a Survey forum and shared the link to the local people. Below figure explains the Structure of the provided Survey.



Survey of What people do before watching a Movie/TV show

This is would be very helpful for my Final Year Project.

* Required

Do you Love watching Movies/TV Shows? *

○ Yes

○ No

How do you select a movie/TV show? *

○ Reading IMDB Movie Reviews

○ Reading Movie Reviews on Social Media Platforms.

○ Watching Trailer of the Movie/TV Show

○ By Suggestions from people

○ Watching YouTube Review videos of the Movie/TV Show

Do you like a System (Web Application) which will recommend you whether the selected movie/TV show Worth to Watch or Not?

○ Yes

○ No, I'll watch whatever I want.

Submit
This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy
Google Forms

Figure 45 Survey forum preview

This survey basically contains 3 logical questions.

- Logical Question 01: Do you Love watching Movies/TV shows?

  This question basically checks whether the people love to watch movies or not. This question only have 2 options and they are **“Yes”** and **“No”**. So, if the person loves to watch movies or TV shows, He / She can select **“Yes”** if else they do not watch movies, they can select **“No”.**

- Logical Question 02: How do you select a movie/TV show?

  This question basically checks the opinions of the person, what He/ She will do select a movie to watch. This question has 5 options to select. These are the available options where people normally get to know about movies/TV Shows. So, if the person loves to watch movie, they must have an option to how they select or get to know about the movie.

  1. Reading IMDB Movie Reviews
  2. Reading Movie Reviews on Social Media Platforms.
  3. Watching Trailer of the Movie/TV Show
  4. By Suggestions from people
  5. Watching YouTube Review videos of the Movie/TV Show

- Logical Question 03: Do you like a System (Web Application) which will recommend you whether the selected movie/TV show Worth to Watch or Not?

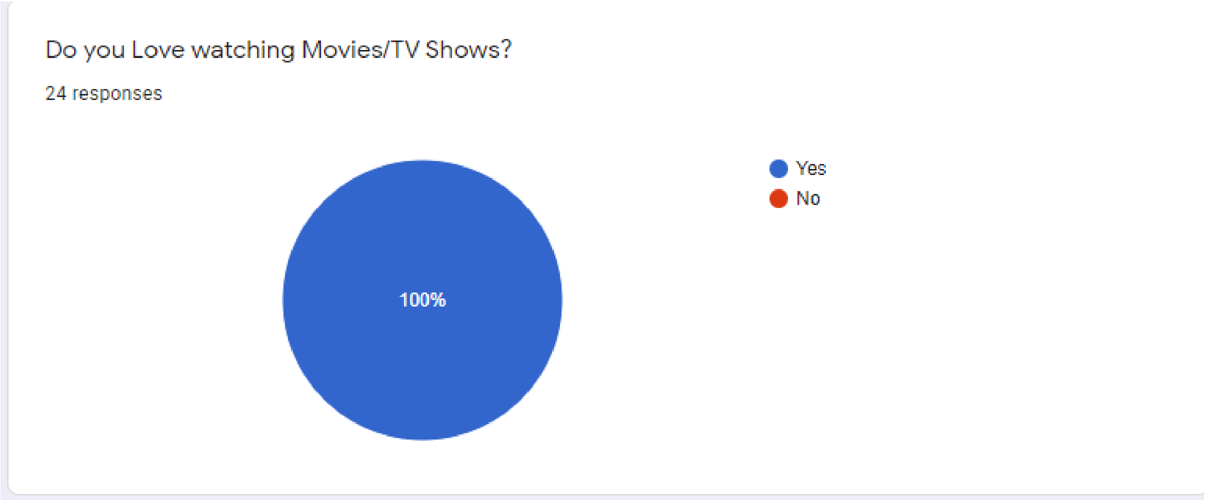
  This is the last question, and it basically checks whether the person would really like to try a Web Application which will help them to get to know more about and get a recommendation about the selected movie. This question only have 2 options and they are **“Yes”** and **“No”**. So, if the person would like to try the Proposed Solution, He / She can select **“Yes”** if else they do not care about the proposed solution and they will watch if the movie is not good, they can select **“No”.**

Survey Result

As a result of this survey, I only got just 24 responses from the local people. Even though the number of responses is little lower, the opinions are honest. So that is what matters because I can know what they would consider before purchasing or watching a movie.

Response for Question 01

So basically, I got the expected results for this question. Out of the 24 response each person loves to watch movies/TV shows and the response rate is **100%**. Below Chart explains the responses from the survey.

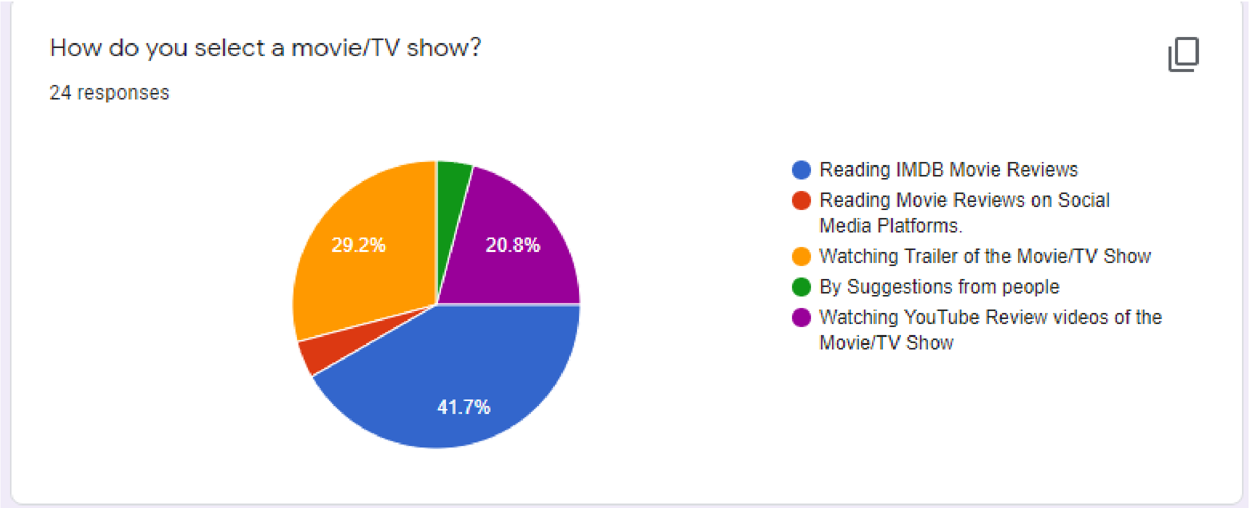


Do you Love watching Movies/TV Shows?

24 responses

Figure 46 Response for Question 01 Preview

Response for Question 02



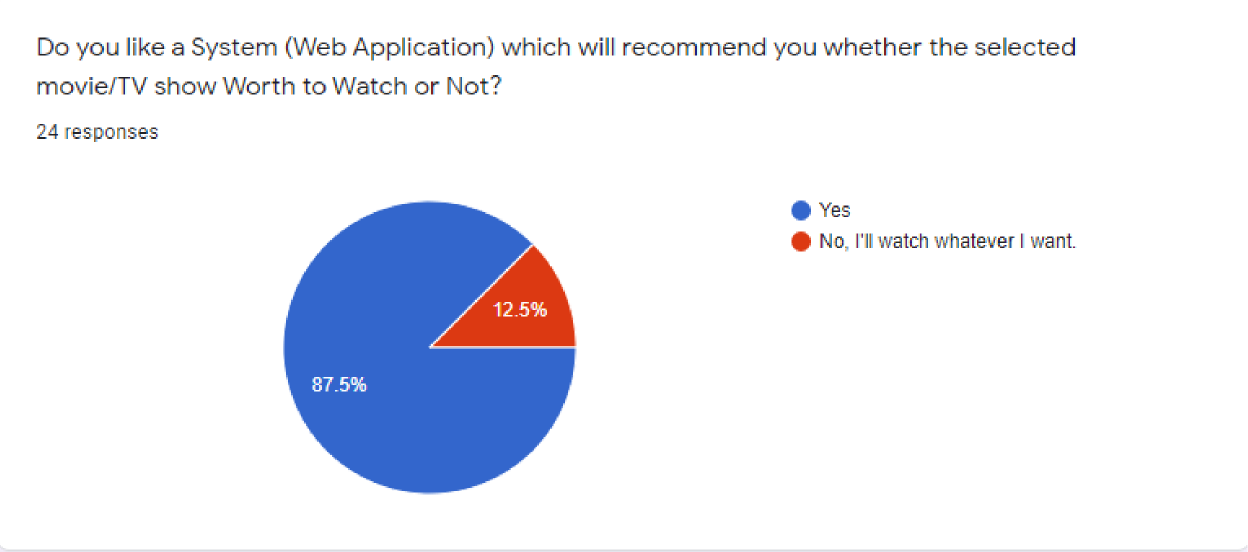
How do you select a movie/TV show?

24 responses

Figure 47 Response for Question 02 Preview

So basically, did not expect this result for the question. My expectations were percentage for the **"Watching Trailer of the Movie/TV Show"** would go higher than other options. But out the 24 people, in percentage **41.7%** selected **"Reading IMDB Movie Reviews"** option. This percentage views explains that most of the people use IMDB Reviews to select or purchase a Movie.

Response for Question 03



Do you like a System (Web Application) which will recommend you whether the selected movie/TV show Worth to Watch or Not?

24 responses

Figure 48 Response for Question 03 Preview

As I expected response for the “Yes” option had higher percentage than “NO” which means people are willing to have a solution which will assist them to decide about the movie. So, the proposed Web Application will be a better solution for people because this system will easily recommend the people about the selected movie by reading the latest reviews on IMDB.

In conclusion, most of the people love to watch movies and most of them use IMDB movie reviews as their movie deciding tool. Hence, they are willing to try a Web Application which will help them to decide about the movie, proposed solution will be very familiar with them and will be useful for them for daily usage.

Project Plan (Gantt Chart)

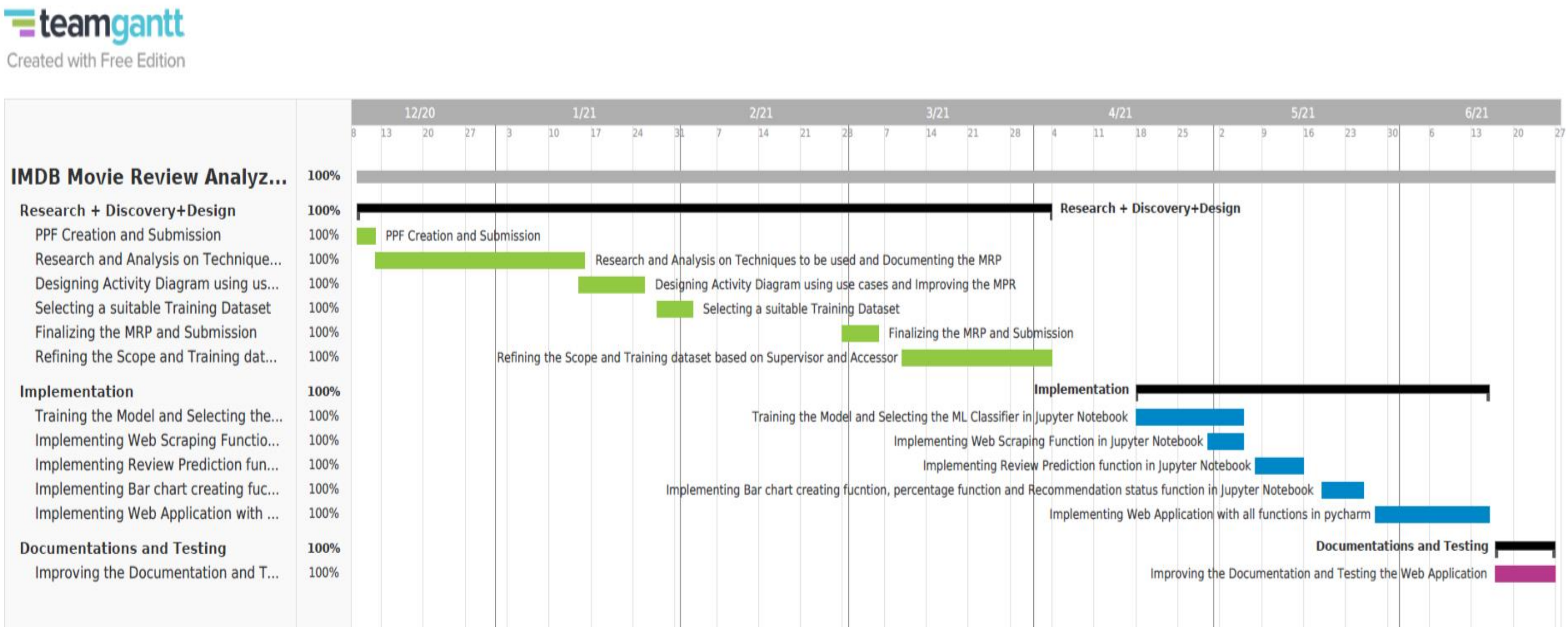


Figure 49 Project Plan (Gantt Chart)