

**SENTIMENTAL ANALYSIS ON MOVIE REVIEWS USING NLP AND
MACHINE LEARNING APPROACH**

BY

**Zeba Fauzia Nisi
ID: 181-15-11121**

AND

**Nashit Farzana Noushin
ID: 181-15-10947**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Zerin Nasrin Tumpa
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Abdus Satter
Assistant Professor
Department of CSE
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
13 SEPTEMBER 2022**

APPROVAL

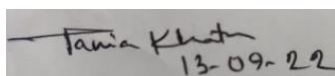
This Project titled “Sentimental Analysis on Movie Reviews using NLP And Machine Learning Approach”, submitted by Zeba Fauzia Nisi and Nashit Farzana Noushin to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13 September 2022.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Chairman



Tania Khatun (TK)
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Mohammad Monirul Islam (MMI)
Senior Lecture
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Dewan Md Farid
Professor
Department of Computer Science and Engineering
United International University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Zerin Nasrin Tumpa, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



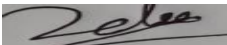
Zerin Nasrin Tumpa
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Zeba Fauzia Nisi
ID: 181-15-11121
Department of CSE
Daffodil International University



Nashit Farzana Noushin
ID: 181-15-10947
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Zerin Nasrin Tumpa, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Touhid Bhuiyan, Professor, and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Film is an art medium that consists of a composite collection of basic media such as literature, music, painting, photography, drama etc. It has surpassed the basic media. Film has a reputation as an influential medium and one of the best tools of education. Its popularity is steadily increasing among people and all over the world. A movie review introduces a movie to the audience. Film analysis can play a role in developing the audience's perception of the film. Movie review establishes an effective link between the audience and the director. Criticism establishes an effective link between the audience and the director. Similarly, movie review analysis is now an important topic all over the world. Movie analysis is important not only to understand movies but also to know people's interest or people's emotions. Sentiment analysis is the most commonly used method for predicting user evaluations. It is the art of analyzing data on what people, public truly thinks about your business, text, opinion, social media etc. It's an incredibly powerful tool in analytics toolkit. To analyze reviews, we must count the number of positive and negative words in a given text. Machine learning algorithms can help to understand whether a movie review is positive or negative. In this research, we discussed how to predict positive and negative reviews of movies using machine learning approaches.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
 CHAPTER	
 CHAPTER 1: INTRODUCTION	 1-5
1.1 Introduction	1-2
1.2 Motivation	2
1.3 Objective	2-3
1.4 Rational of study	3
1.5 Research Questions	3
1.6 Expected Outcome	3-4
1.7 Project Management and Finance	4
1.8 Layout of the Report	4-5
 CHAPTER 2: BACKGROUND	 6-10
2.1 Introduction	6
2.2 Related works	6-7
2.3 Research Summary	9

2.4 Scope of the problem	9-10
2.5 Challenges	10
CHAPTER 3: RESEARCH METHODOLOGY	11-29
3.1 Introduction	11
3.2 Research Subject and instrumentation	11-12
3.3 Data Collection Procedure	12
3.4 Proposed Methodology	13-27
3.5 Statistical Analysis	28
3.6 Implementation Requirement	29
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	30-35
4.1 Introduction	30
4.2 Experimental Setup	31-34
4.3 Experimental Result	34
4.4 Discussion	35
4.5 Summary	35
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	36
5.1 Impact on Society	36
5.2 Impact on Environment	36
5.3 Ethical Aspects	36
5.4 Sustainability Plan	36

CHAPTER 6:	SUMMARY, CONCLUSION,	37
RECOMMENDATION	AND IMPELICATION FOR	
FUTUTRE RESEARCH		
6.1 Summary of the Study		37
6.2 Conclusions		37
6.3 Recommendation		37
6.4 Implication for Further Study		37
REFERENCES		38-39

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Data Collections	12
Figure 3.2: Workflow for review	13
Figure 3.3: Used Library	14
Figure 3.4: Bar graph showing sentiment	15
Figure 3.5: Pie graph showing sentiment	15
Figure 3.6: Positive word cloud of reviews	16
Figure 3.7: Negative word cloud of reviews	17
Figure 3.8: Import NLTK	18
Figure 3.9: Import Stop words	19
Figure 3.10: Split train and test data	21
Figure 3.11: How the random forest classifier operates in steps	22
Figure 3.12: How the random forest classifier works	23
Figure 3.13: How the Decision Tree classifier operates in steps	24
Figure 3.14: How the Decision Tree classifier operates works	24
Figure 3.15: How the Logistic Regression classifier operates in steps	25
Figure 3.16: How the naive bayes operates	26
Figure 3.17: How the naive bayes works	26
Figure 3.18: How the kNN operates	27
Figure 3.19: How the KNN works	28
Figure 4.1: Work Flow	30
Figure 4.2: Results using Logistic Regression	31
Figure 4.3: Results using Naive Bayes	31
Figure 4.4: Results using Random Forest	32
Figure 4.5: Results using Decision Tree	32
Figure 4.6: Results using KNN	32
Figure 4.7: Results comparison bar chart	34

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Previous works on related domain	8
Table 3.1: Statistical Analysis	28
Table 4.1: Result all details of different algorithm	33
Table 4.2: Results of Different Algorithm	34

CHAPTER 1

INTRODUCTION

1.1 Introduction

Film is a wonderful discovery of human civilization. It is one of the best art medium and mass media of modern times. No other art media is capable of establishing such a connection with the general public. Cultural elements are closely related to films. Film has a reputation as an influential medium of art, a powerful medium of entertainment and one of the best tools of education. Its popularity among people in all countries of the world is increasing and it has already gained recognition as a powerful medium of mass communication. Films can best communicate with the common people because of the synergy of the visual world. The film represents the culture in which it is made. As a result of its outstanding ability to influence the lives of people of all walks of life, films are playing a wide and far-reaching role in social life in modern times. Films have a lot of role in social organization and development of people's sense of life.

Although the primary role of films is entertainment, films can also play an educational role like radio and television. Films have already established their place in the modern educational system as teaching aids in educational institutions. In the advanced countries of the world including America, Russia, Germany, Japan, educational institutions are showing informational films along with textbook-based education of students. That's why these days it is said that film is part of education.

Similarly, movie review analysis is now an important topic all over the world. Movie analysis is important not only to understand movies but also to know people's interest or people's emotions. Films have various elements that contribute to the final result and the audience's understanding of the story. Film criticism not only helps how we watch films, but also the film industry itself. Reading and watching these critics' reviews helps expand one's worldview and ability to empathize with other people. It also informs the filmmakers, enabling them to avoid the issues that afflict bad movies and place greater attention on

aspects that were successful in earlier movies. This raises the standard of the business as a whole and encourages the creation of more varied and intriguing films.

Machine learning algorithms can help to understand whether a movie review is positive or negative. To analyze reviews, we must count the number of Using machine learning methods, identify the both positive and negative terms in a document.. In this research, we intend to use sentiment analysis to discuss how to predict positive and negative reviews of movies using machine learning approaches.

1.2 Motivation

We were excited to try something different. So, we intend to conduct research in the area of computer science and artificial intelligence. As a result, we begin to look for new ideas from the internet and different websites. However, no strategy could satisfy us. Then we thought we will work with sentiment analysis using a machine learning approach.

Movie reviews are typically used to express an opinion, a complaint, or a compliment about something. It provides the possibility of a message that motivates the audience and public. This paper proposes a method for categorizing reviews from the movie site that is likely to motivate learning activity. This method is advantageous because it eliminates many inappropriate messages while avoiding messages in foreign languages.

1.3 Objective

Our primary goal is to classify different sentimental analysis from IMDb movie review. So, we can describe our goals like this-

- To check whether movie review is positive, negative or neutral.
- To know a user opinion on a target object by analyzing various sources.
- To develop a system that can identify different human behaviour from movie reviews.
- To visualize some of human emotions classified by classifier algorithm.
- To understand people opinion.

- Helps to find good movies.

1.4 Rational of the study

There have been Numerous workshops have been held on movie reviews and others before. Many scholars are focusing on refining the sentiment analysis model to discern between good and negative reviews with clarity. Movie reviews can be used not only as a marketing tool but also as a predictor as to how a film will perform financially and technically. It helps you to understand your films reviewers.

1.5 Research Questions

It turned into hence hard to complete the study and fulfil our purpose for us. So as to conquer a practical, economical, and precise reaction to the matter, we want to endorse the subsequent queries to express the intuitions and consequences of this drawback:

- Can we collect data?
- Can we retrain my program with Machine Learning and NLP approaches?
- Can Machine learning be good for Sentiment analysis?
- Can the Machine Learning process accurately predict the sentiment analysis?
- Can this Machine Learning process show prediction of a person in real time?

1.6 Expected Outcome

There are many agendas that are the main anticipated affair of our work. The primary goal of sentiment analysis on social media is to obtain valuable business insights that will assist you in improving your overall business performance. The main ideal of this exploration-grounded design is to make a complete and effective procedure that can estimate the sentiment analyzing many movie reviews from the Internet Movie Database (IMDb). Some expected outcome points are:

- Classify different human emotions.
- Analysis Constructive criticism of people.
- Analysis Constructive criticism of movie review.
- Help in psychological treatment.
- It helps you understand your audience
- Reduce false news.
- Decrease negativity.

1.7 Project Management and Finance

As we collected our data from Kaggle, so it is free of cost. We applied five Machine learning algorithms in google colab which is also a free resource. So, we don't need any cost to complete this research.

1.8 Layout of the Report

The objective, motivation, research questions, rationality, and expected result are covered in the first chapter. In this section we will narrate the entire format of the report.

In the second chapter, we have discussed the previous research works and researches which have been done on this particular topic. The next part of this section shows their problem or limitations of their research and how we studied their limitations with our research idea. And at last, the limitations that have been faced by us are discussed.

In chapter three, we have discussed the theoretical discussion related to our research area. To do so, the statistical methods of our work are described in this section. And at the last part of this chapter, confusion matrix is displayed to show the precision name of the classifier.

In chapter four, we have discussed the experimental result, evaluated the performance of the model and discussed on the output of our result. The visualization of output is shown too here.

In chapter five, we have put the gist of our research, our future work on this field and concluded the research. It shows Impact on Society, Environment and Sustainability. The visualization of output is shown too.

At the last part of this section, we have discussed our limitations which have appeared during our research. These limitations may help others or give scope to others who will do further research on this domain.

CHAPTER 2

BACKGROUND

2.1 Introduction

Here, related works, the definition of the evaluation, and demanding situations concerning this evaluation will be discussed. We will discuss outstanding assessment articles and their research, tactics, and their studies that might be applied to our study in the section on related works. We'll provide a definition of our related works in the evaluation define part. We're going to talk about in the part on difficult circumstances, but we have a tendency to increase the accuracy level and try to be more accurate.

2.2 Related works

On this subject, several researchers have already worked. They used several algorithms, including Nave Bayes, K-Nearest Neighbour, and Random Forest [1] to assess the movie reviews in the paper written by Palak Baid, Apoorva Gupta, and Neelam Chaplot.

A study employing the improvised random forest classification method was published in the International Journal of Advanced Research in Computer Science in 2019 [2]. The study proposes a prediction model for the sentiment analysis of movie reviews.

Turkish Journal of Computer and Mathematics Education released a review article in 2021 that suggested sentiment analysis as a method of predicting data from a dataset, such as the IMDB database, for user reviews or viewer information [3].

To determine whether components of the movie were favored or despised by viewers, Saba Shireen Khan and Rajni Pamnani focused on aspect-based sentiment analysis of movie reviews [4].

Savita Harer and Sandeep Kadam claim in their article published in the (IJCSIT) International Journal of Computer Science and Information Technologies that they

designed and created the numerous methodologies needed for sentiment analysis of the movie domain in a mobile environment [5].

In addition to giving comparison findings of various deep learning networks, four researchers—Nehal Mohamed Ali, Marwa Mostafa Abd El Hamid, and Aliaa Youssif—introduced a new categorization sentiment analysis utilizing deep learning networks [6].

Oaindrila Das and Rakesh Chandra Balabantaray describe an unique method for classifying online movie reviews using machine learning algorithms and components of speech in their paper [7].

In order to conduct sentiment analysis on IMDb movie reviews, three researchers from the Department of Information Science and Engineering at JSS Science and Technology University in Mysuru, Karnataka (India) used the hybrid feature extraction method [8].

A universal approach to emotion categorization was proposed by S. Prathap and Sk. Moinuddin Ahmad [9].

Two researchers, Nanda Kumar AN and Mohan Kumar AV, proposed two distinct grouping and categorization methods. Positive and negative words are categorized using ROCK and CART [10].

Table 2.1: Previous works on related domain

Paper Name	Methodology Used	Accuracy
Machine Learning Methods for Movie Review Sentiment Analysis [1].	Machine learning algorithms.	Naive Bayes 81.45% , RFC 78.65% , KNN 55.30% .
Sentiment classification of movie reviews using feature-selected improvised random forests [2].	The work that is being suggested uses a hybrid approach to feature selection.	90.69%
A Sentiment Analysis Using Long Short Term Memory Networks for Movie Reviews [3].	Use LSTM for proposed technique.	Accuracy is found to be over 86%.
Mining and Summarizing Movie Reviews in Mobile Environment [5].	Frequency-based method and Latent Semantic Analysis (LSA) algorithm.	Rfc 100%
Deep learning models for sentiment analysis of movie reviews [6].	4 deep learning techniques were implemented.	MLP 86.74%, LSTM 86.64% , CNN 87.70% , 89.20% CNN-LSTM

2.3 Research Summary

On this subject, much more study has already been done. On this subject, every effort has produced various positive outcomes.. Some researcher applied Machine learning, some Deep learning, some doing some another new algorithm to done this work before and they achieved good accuracy and result by doing this work.

According to Humera Shaziya, G.Kavitha, and Raniah Zaheer utilized weka to assess the effectiveness of the chosen feature [11].

The optimal method for sentiment analysis is identified by K. Amulya, S. B. Swathi, P. Kamakshi, and Y. Bhavani, who also demonstrate that deep learning algorithms offer reliable results [12].

Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas Analysis study on more than 1000 Facebook posts made during newscasts, [13].

In a different publication, Addlight Mukwazvure and K.P. Supreethi develop a hybrid method for analyzing the sentiment in news comments that makes use of a sentiment lexicon to identify polarity [14].

2.4 Scope of the problem

Sentiment analysis has been a valuable tool for brands seeking to understand what their customers are thinking and feeling. The primary goal of sentiment analysis on social media is to obtain valuable business insights that will assist you in improving your overall business performance. It is a straightforward kind of analytics that helps businesses find their most important areas of weakness (negative attitudes) and strengths (positive sentiments).

2.5 Challenges

The principal demanding situations of this research are gathering and mannering the dataset. When preparing a research paper, it is common to face challenges. These difficulties cause errors to occur frequently, although they can be avoided. We were under covid situation when we started our research work. So, we faced a lot of problems while collecting our primary data. Though several works on this topic have done before, we faced a lot of challenges to overcome the limitations of the researchers who have done their work before. The challenges we have faced are given below:

Lack of data: It was one of the biggest problems for us that we have spent a huge time to collect the data from the trusted sites or sources. We have collected our data from Kaggle.

Problem finding: As for some related papers have already been created for prediction. We made every effort to identify any errors and features that the paper should have had..

Tools: We have used Machine learning methods. As a machine learning program, the whole program is written in python programming language. The tools we used is google colab, jupyter notebook, numpy.

Limitations of the researchers: We faced a lot of challenges to overcome the limitations of the researchers who have done their work before.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This paper provides a framework for the research techniques used in the investigation. We depict the examination structure that was chosen with the end goal of this investigation in mind, as well as the reasons for this decision. We also discuss the strategies used to make our proposed model. This part by providing evidence of this project's measurable theories and giving a precise idea of the execution requirements. The implementation requirements are discussed on the last section finally.

3.2 Research Subject and instrumentation

First of all, we talked about the movie reviews and how much positive and negative reviews can be predicted by the machining learning model. In this process we use some instruments, like hardware and software. Here is a list of the instruments used.

Software and Hardware:

- Intel Core I5
- 4GB RAM
- Windows 10

Developments Tools:

- Python
- Numpy
- Pandas
- Matplotlib
- Seaborn

- NLTK
- Scikit Learn

3.3 Data Collection Procedure

Dataset was obtained via Kaggle. The data has to be cleaned up and rid of several symbols as well as various punctuation marks. The type of data in this set is comments. We eliminate several of these words because they are not necessary for our coding part. That term is eliminated using natural language processing. Finally, obtain a review prediction using a machine-learning method.

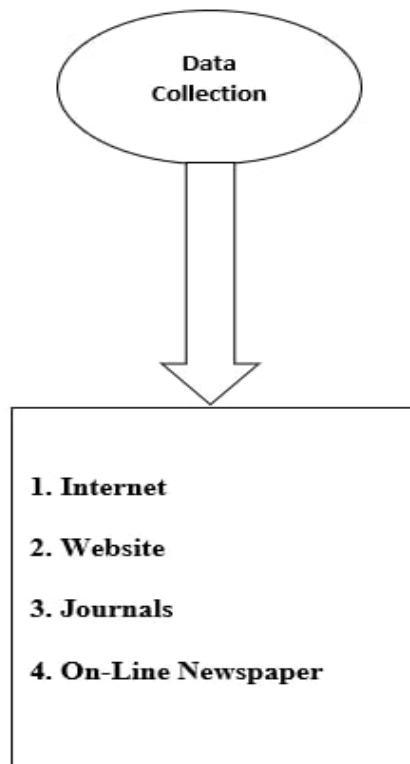


Figure 3.1: Data Collections

3.4 Proposed Methodology

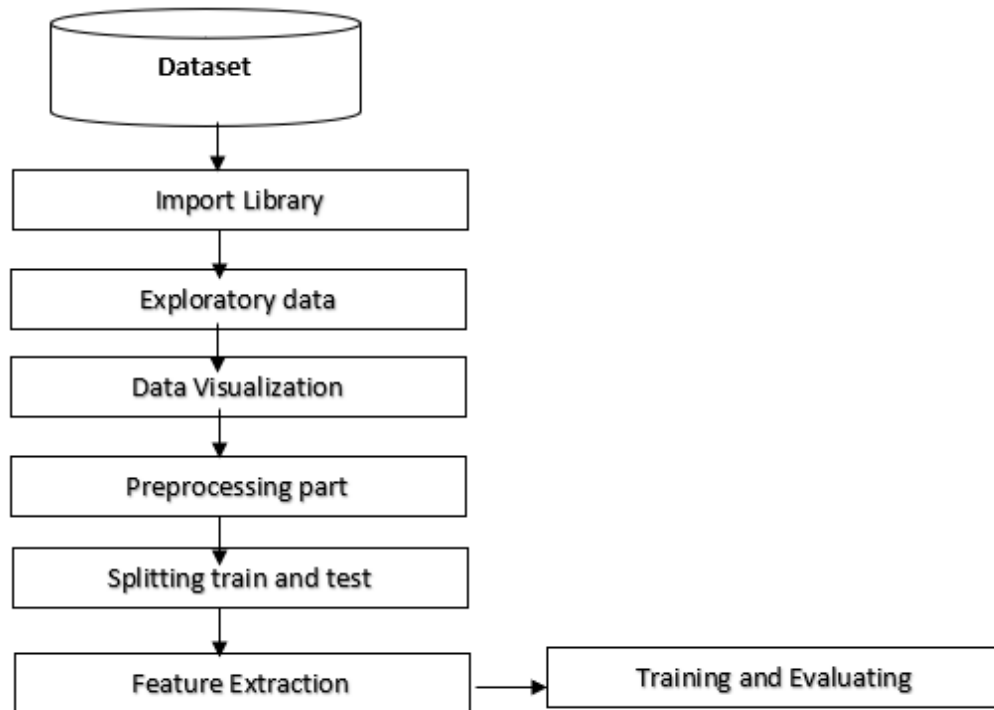


Figure 3.2: Workflow for review

3.4.1 Get Dataset

We used Kaggle to get our data. There are several datasets available on Kaggle. Due to our interest in working with movie reviews, we chose the Kaggle platform and the IMDB data collection. This dataset is used in several ways. We want to forecast positively and negatively using a variety of algorithms. Here, we've improved the preprocessing stage and incorporated a few more algorithms.

3.4.2 Import Library

Many libraries are being used in this project. like as

- Python Library

- Machine learning packages
- Data analysis library
- Data visualization library
- Text preprocessing
- Machine learning model

```
In [58]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB, MultinomialNB
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
import warnings
warnings.filterwarnings('ignore')
```

Figure 3.3: Used library

3.4.3 Data Analysis

First Data analysis is importing the dataset and employing several command lines. Two columns are review and sentiment. This dataset has 2 types and 50,000 reviews. There are no null values. this data set is imbalanced 25000 positive and 25000 negative types.

3.4.4 Data Visualization

Here, some visualization is being used.

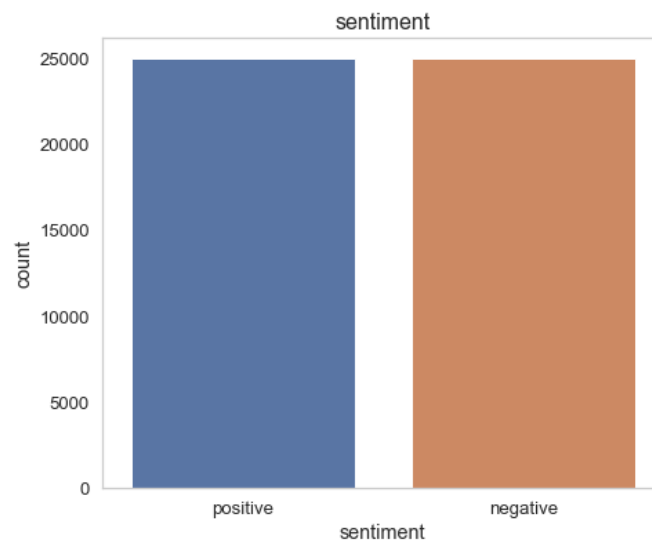


Figure 3.4: Bar graph showing sentiment

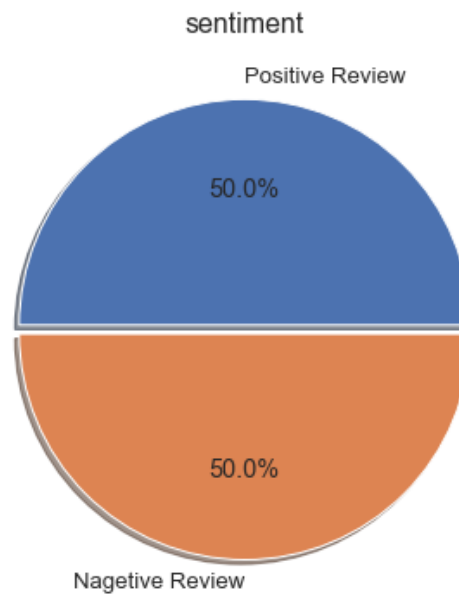


Figure 3.5: Pie graph showing sentiment

Word cloud:

A text visualization approach called a "word cloud" or "tag cloud". It is a process of finding crucial words. A word's bigger font size more appropriately conveys its importance in relation to other words in the cluster. Although there are other techniques to build word clouds, the most popular method uses the frequency of words in our corpus. Therefore, we will use the Frequency type to create our word cloud.

Our data set has two types of sentiment: negative and positive. There are 50000 rows and we get negative and positive sentiments by using reviews. So that we take the meaning of the full word and make a word cloud and we understand easily how many words are used for negative and positive word clouds.

```
In [25]: plt.figure(figsize = (15, 15), facecolor = None)
plt.imshow(word_cloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```

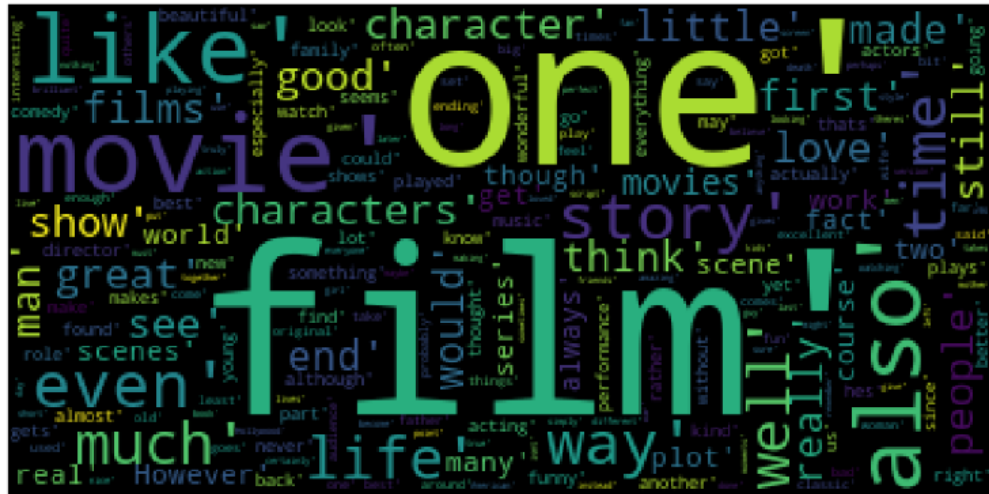


Figure 3.6: Positive word cloud of reviews

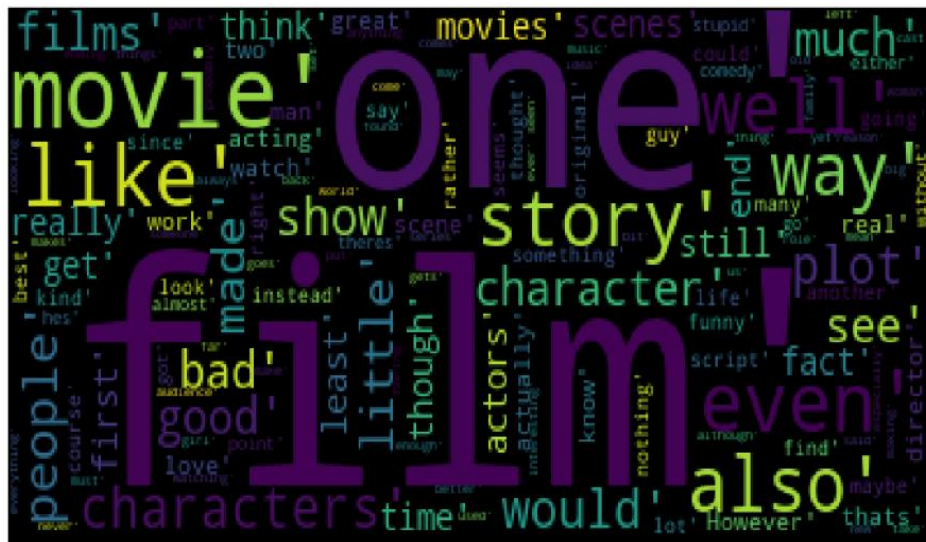


Figure 3.7: Negative word cloud of reviews

3.4.5 Data Pre-Processing

This data has two columns' reviews and sentiment. Sentiments have two types positive and negative. The data is balanced data so that using binary classification machine learning algorithm. This review data was text data so there a lot off word and punctuation. Using Natural language processing preprocessing the review data. Taking some step to preprocess the data like Import NLTK library, stop words, case normalization, remove punction and also used some step to remove white space and URLs.

Import NLTK it helps out to preprocessing

Symbolic and analytical natural language processing (NLP) for English is supported by the Natural Language Tool kit (NLTK), a collection of Python-based modules and applications. NLTK supports categorization, tokenization, stemming, tagging, parsing, and semantic reasoning.

NLTK is one of the most well-liked tools for using linguistic data. provides user-friendly APIs for a variety of text preparation methods. It has a sizable and active community that helps the library grow and flourish. It is open-source and free for Linux, Mac OS X, and Windows. Natural language processing has long considered text preparation to be an essential stage (NLP). It makes text simpler to grasp, which improves the performance of machine learning systems.

```
In [71]: import nltk

In [72]: nltk.download('stopwords')

[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\User\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!

Out[72]: True
```

Figure 3.8: Import NLTK

Stop word

Stop words are any words that make little sense when added to a sentence. Without endangering the sense of the statement, they may be safely disregarded. Among the most popular short function phrases for different search engines are the, is, at, which, and on.

Frequently used words like "the," "a," "an," or "in" are examples of stop words. Such expressions like "if," "but," "we," "he," "she," and "them" are considered stop words.

Typically, we can remove these words from texts without changing their meaning, and doing so improves the model's performance (but not always).

These keywords shouldn't take up unnecessary processing time or storage space in our database. By maintaining a list of words, you consider to be stop words, we may easily get rid of them.

```
In [73]: import nltk
         from nltk.corpus import stopwords
         list1=stopwords.words('english')
```

Figure 3.9: Import stop words

Case Normalization:

In this situation, we simply change the case of every letter in the text to uppercase or lowercase and nlp will be handled differently since Python treats case differently. You may rapidly change the string's case by using `str.lower ()` or `str.upper ()`.

Removing Punctuations

There is no benefit to the information from the punctuation in the text. Any term that has punctuation adds to it makes it more difficult to differentiate from other words.

3.4.6 Feature Extraction

In general, feature extraction is a technique where raw data is grouped into comprehensible units. The fact that these massive databases include a large large number of variables and the extensive computer resources these variables need Having resources to evaluate is a distinctive quality. Feature Extraction may thus be effective in this circumstance for choosing certain factors and combining some of the associated variables to reduce the data volume. The metrics for precision and recall would be utilized to assess the collected data. A linear dimensionality reduction method called PCA is often used. It is an unsupervised

learning algorithm. The Count Vectorizer tool from the Python scikit-learn toolbox is a wonderful tool. It is used to turn a text into a vector depending on how often (or how many times) each word occurs in the text. This is helpful when processing several of these texts and turning each word into a vector (for use in further text analysis).

A matrix generated by Count Vectorizer has a column for each unique word and a row is used to represent each text sample from the document. The worth of every column represents how many times the phrase appears in that specific text sample.

In this part, we also use a bag of words. Natural language processing uses the text modeling technique known as "bag of words." To explain it formally, It is a technique for extracting features from text data. This method for extracting characteristics from texts is straightforward and versatile.

A "bag of words" is a textual representation of word recurrence in a document. We only keep track of word counts; we don't worry about grammatical rules or word arrangement. Since any information on the organization or structure of the words inside the text is disregarded, it is referred to as a "bag" of words. The model is just concerned with whether recognized terms exist in the text, not with where in the text they do.

3.4.7 Splitting into X and Y variables

As a result, we've divided the characteristics into two categories:

X: review

Y: sentiment

70% of the data we used in our tarin model and 30% of the total data is used for testing

TRAIN DATA TEST DATA SPLIT

```
In [31]: x=df['review']
         y=df['sentiment']
         X_train, X_test, y_train, y_test = train_test_split(x,y, test_size=0.3, random_state=42)

In [32]: X_train.shape, y_train.shape
Out[32]: ((35000,), (35000,))

In [33]: X_test.shape,y_test.shape
Out[33]: ((15000,), (15000,))

In [34]: X_train.head()
Out[34]: 38094    much love trains couldnt stomach movie premise...
         40624    good PPV like Wrestlemania XX 14 years later W...
         49425    finding right words everybodys problem vaudevi...
         35734    Im really suprised movie didnt get higher rati...
         41708    Ill start confessing tend really enjoy action ...
         Name: review, dtype: object
```

Figure 3.10: Split train and test data

3.4.8 Training and Evaluating models

To implement our work, we need work with some machine learning algorithm. These algorithm are help us to do our work accurately.

- Decision Tree
- Random Forest Classifier
- Logistic Regression
- Naïve Bayes Model
- KNN Model

3.3.9 Implementation and Algorithms

Random Forest Classifier

A supervised learning technique is random forests. Both classification and regression may be done with its help. Random forests may be used for a range of applications, including recommendation engines, image classification, and feature selection.. It may be used to spot fraudulent activity, classify dependable loan candidates, and foresee diseases. It is the foundation of the Boruta approach, which selects relevant attributes from a dataset.

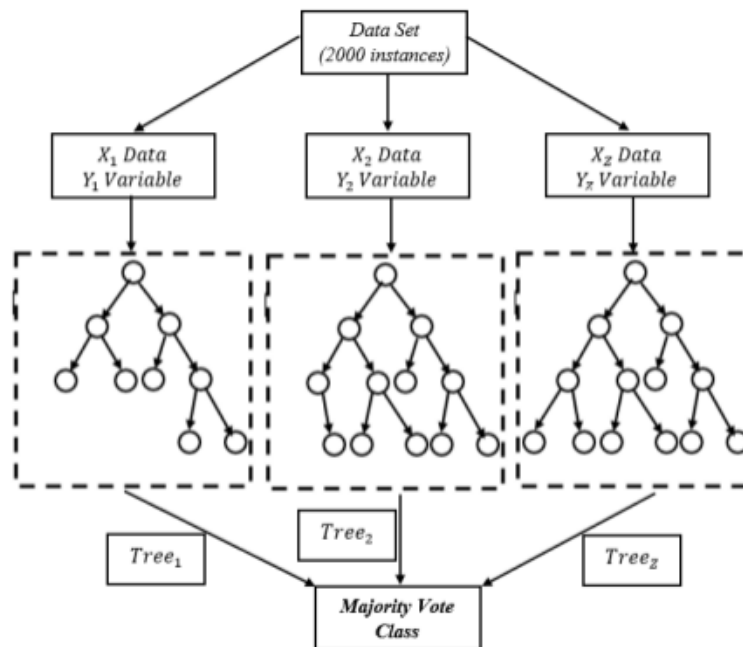


Figure 3.11: How the random forest classifier operates in steps

It can also be used to manage missing data. There are two ways to handle missing data: determining the proximity-weighted average or substituting continuous variables with their median values.

Using Gini significance or mean reduction in impurity (MDI) in a random forest, the relevance of each feature is determined. The Gini significance is the overall reduction in

node impurity. This is how much the model's fit or accuracy worsens when a variable is removed. The larger the decline, the more important the variable. In this situation, the mean decrease is a crucial parameter for variable selection. The entire explanatory power of the variables may be expressed using the Gini index.

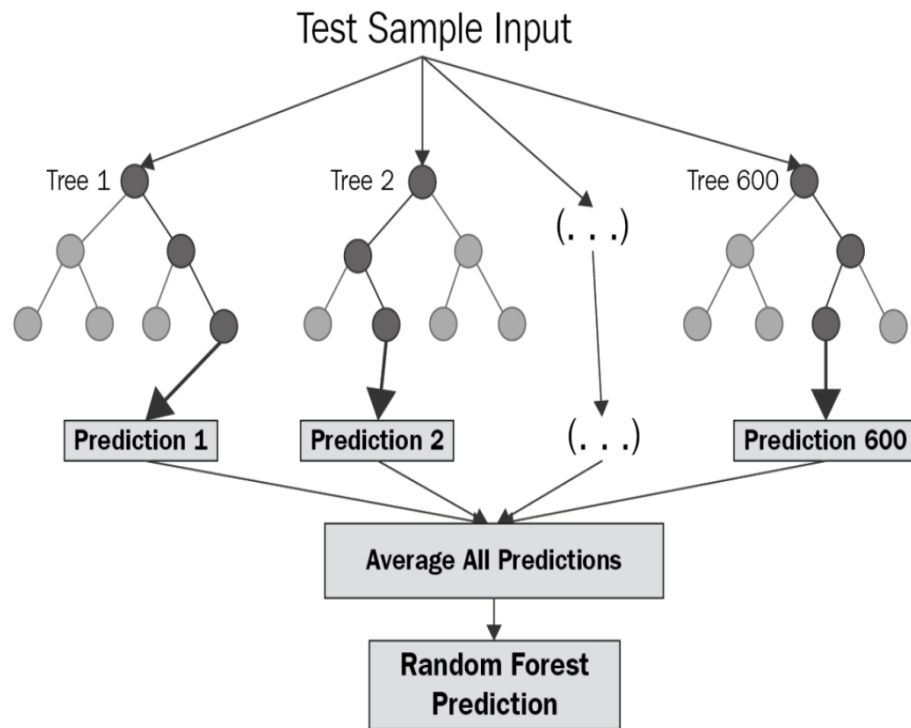


Figure 3.12: How the random forest classifier works

Decision Tree

A strategy for supervised learning is the decision tree. Problems involving classification are solved using it. Decision nodes and leaf nodes are the two nodes in a decision tree. Any choice is made using decision nodes, and the result of those decisions is represented by leaf nodes.

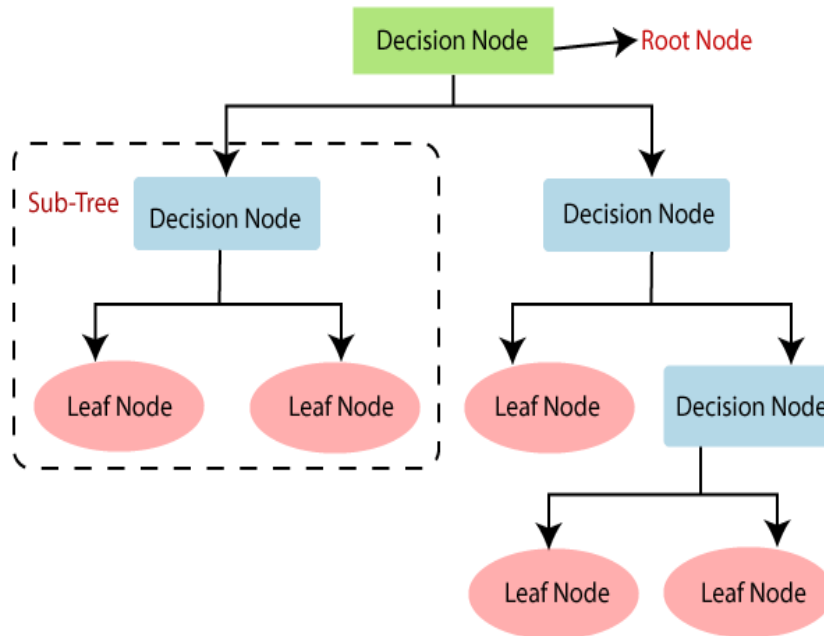


Figure 3.13: How the Decision Tree classifier operates in steps

Multiple algorithms were used in decision trees to decide whether to split a node into two or more sub-nodes. The process of creating sub-nodes improves the homogeneity of the result sub-nodes.

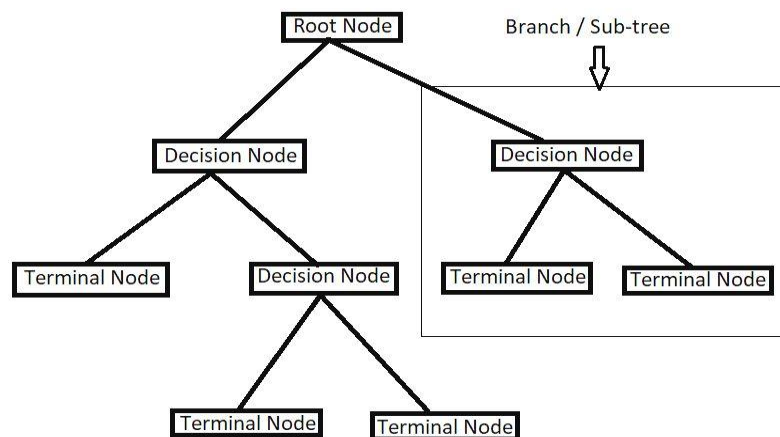


Figure 3.14: How the Decision Tree classifier operates works

Logistic Regression

A supervised classification technique is logistic regression. A key component of logistic regression is determining the threshold value, which is dictated by the classification problem. The threshold value decision is significantly influenced by the accuracy and memory levels. Accuracy and recall should always equal one in an ideal environment, but this is seldom the case.

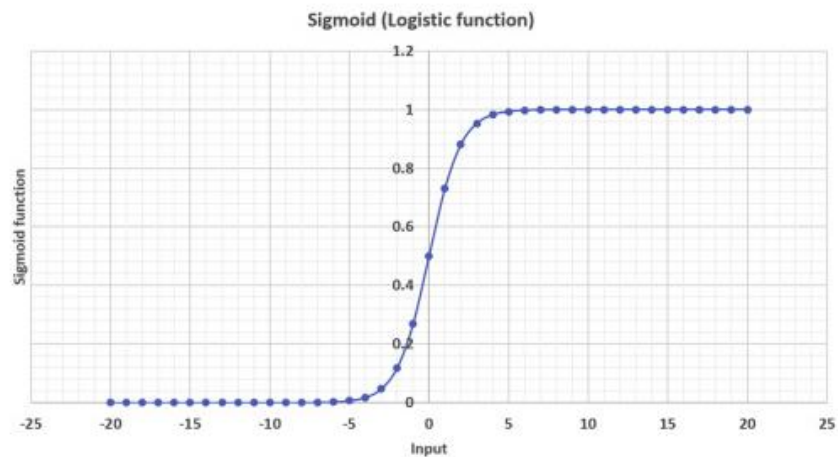


Figure 3.15: How the Logistic Regression classifier operates in steps

Naive Bayes

Naive Bayes algorithm is a supervised algorithm. A sizable training dataset is included, and it is utilized for text categorization. It functions as a probabilistic classifier.

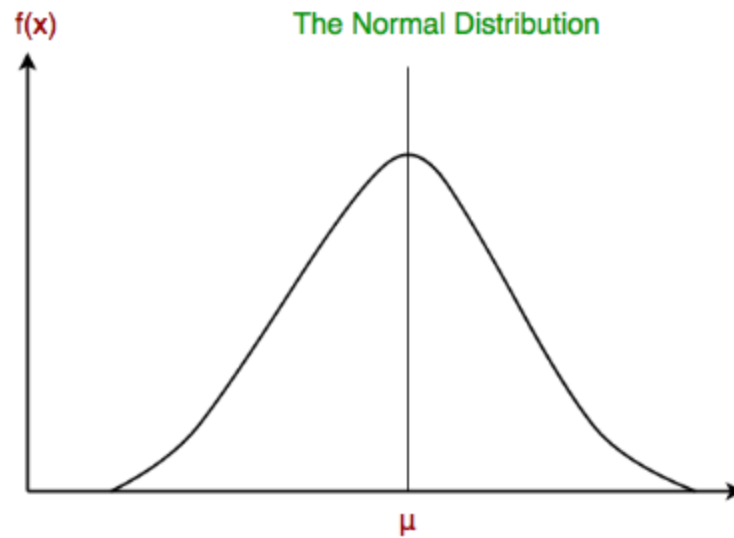


Figure 3.16: How the naive bayes operates

Sentiment analysis, spam filtering, recommendation systems, and other applications make up the majority of uses for naive Bayes algorithms.

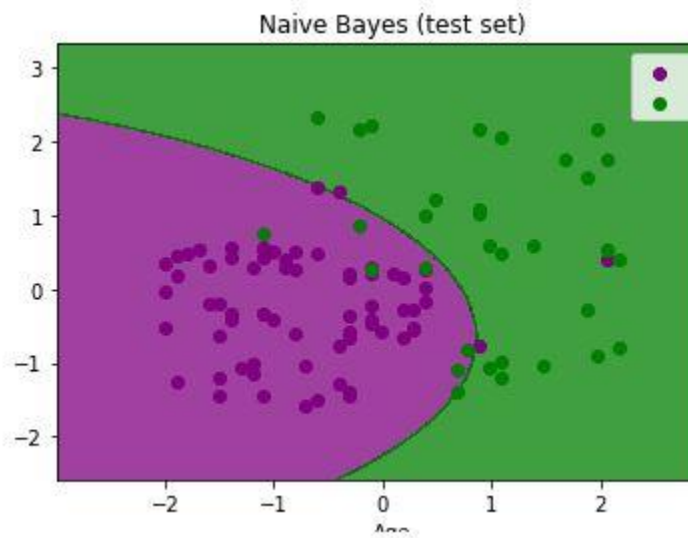


Figure 3.17: How the naive bayes works

KNN Model

The K-Nearest Neighbor method is a popular Supervised Learning technique for solving challenges with categorization and regression. It is a basic algorithm for machine learning. The K-NN method keeps all existing data and categorizes incoming data points based on how similar they are to existing data.

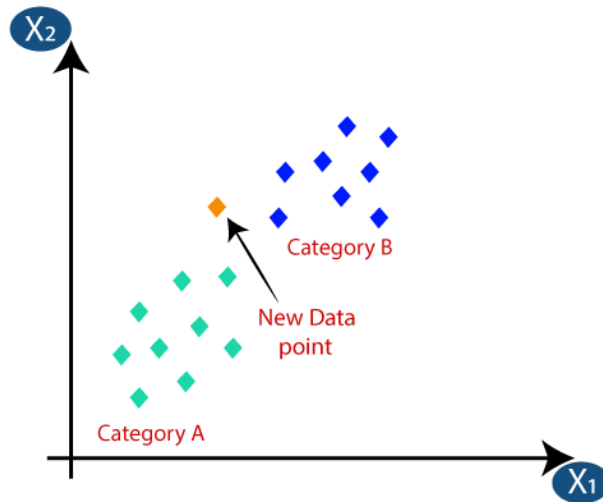


Figure 3.18: How the KNN operates

The K-NN algorithm equation-

$$\begin{aligned}d(p, q) &= d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}\end{aligned}$$

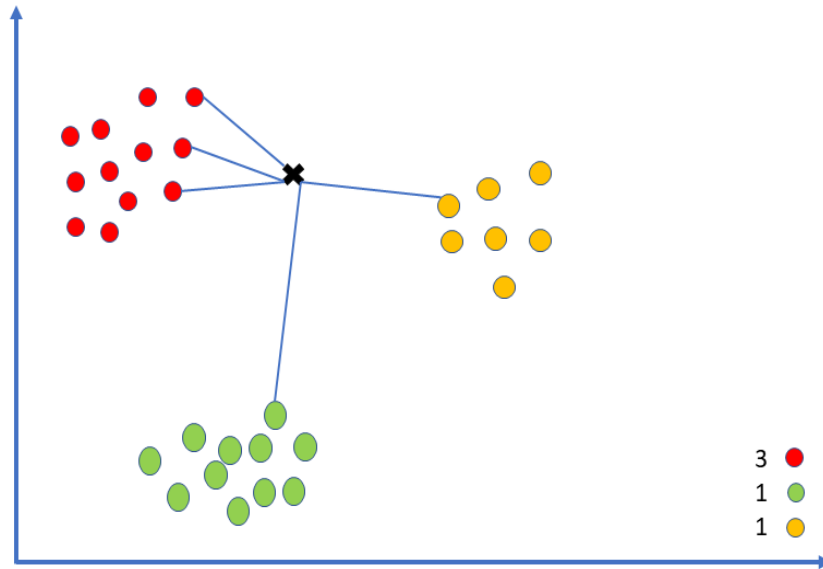


Figure 3.19: How the KNN works

3.5 Statistical Analysis

Table 3.1: Statistical Analysis

1.	The total number of the 2 columns
2.	Total number of rows 50000
3.	70% of the data we used in our tarin model
4.	30% for testing out of the overall amount of data
5.	Data is saved in csv files

3.6 Implementation Requirement

A list of necessary and compulsory tools has been identified after a long and effective analysis over all of the statistical and theoretical concepts and methods that were required for this study. These tools that are listed down are the basic requirement for our kind of image processing and classification. These necessary tools-

Software requirements:

- Operating System: Windows 10
- Hard Drive: 500 Giga Bytes
- RAM: 4 GB

Developing Tools:

- Python Environment
- Anaconda Navigator
- JupyterLab

CHAPTER 4

EXPERIMENTAL RESULTS, DISCUSSION AND CONCLUSION

4.1 Introduction

In this section, we go through the experiment's results.. Only the first five of the forthcoming algorithms have been covered thus far. We will now ascertain the accuracy of those algorithms. The accuracy of each of the five approaches will also be compared.

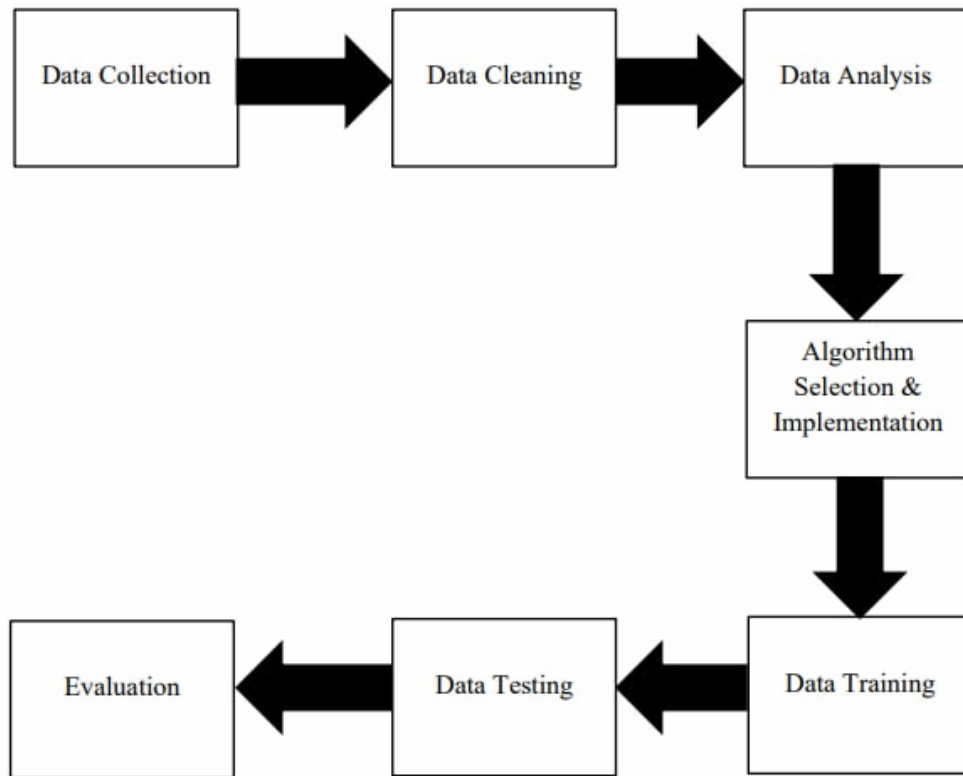


Figure 4.1: Work Flow

4.2 Experimental Result

No machine can produce 100% of what we need, as we all know. To increase accuracy and make sure the model is trained properly, we could additionally train it and make some parameter adjustments.

Output of Logistic Regression:

	precision	recall	f1-score	support
Positive Review	0.88	0.86	0.87	7411
Nagetive Review	0.87	0.89	0.88	7589
accuracy			0.87	15000
macro avg	0.88	0.87	0.87	15000
weighted avg	0.87	0.87	0.87	15000

Figure 4.2 : Result using Logistic Regression

Output of Naive Bayes :

	precision	recall	f1-score	support
Positive Review	0.84	0.84	0.84	7411
Nagetive Review	0.84	0.84	0.84	7589
accuracy			0.84	15000
macro avg	0.84	0.84	0.84	15000
weighted avg	0.84	0.84	0.84	15000

Figure 4.3 : Results using Navie Bayes

Output of Random Forest:

	precision	recall	f1-score	support
Positive Review	0.73	0.84	0.78	7411
Nagetive Review	0.82	0.70	0.75	7589
accuracy			0.77	15000
macro avg	0.77	0.77	0.77	15000
weighted avg	0.77	0.77	0.77	15000

Figure 4.4 : Results using Random Forest

Output of Decision Tree:

	precision	recall	f1-score	support
Positive Review	0.73	0.47	0.57	7411
Nagetive Review	0.62	0.83	0.71	7589
accuracy			0.65	15000
macro avg	0.67	0.65	0.64	15000
weighted avg	0.67	0.65	0.64	15000

Figure 4.5 : Results using Decision Tree

Output of KNN:

	precision	recall	f1-score	support
Positive Review	0.60	0.67	0.63	7411
Nagetive Review	0.64	0.55	0.59	7589
accuracy			0.61	15000
macro avg	0.62	0.61	0.61	15000
weighted avg	0.62	0.61	0.61	15000

Figure 4.6 : Results using KNN

The different algorithms' accuracy, nevertheless, is rather good. Here are a few images that clearly show how much work we put into our research. These tables show precision, recall, the f1 score, support, and accuracy.

Table 4.1: Result all details of different algorithm

Algorithm Name	sentiment	Precision	Recall	F1-score	Accuracy
Logistic Regression	Positive	0.88	0.86	0.87	0.87
	Negative	0.87	0.89	0.88	
Naive Bayes	Positive	0.84	0.84	0.84	0.84
	Negative	0.84	0.84	0.84	
Random Forest	Positive	0.73	0.85	0.78	0.77
	Negative	0.82	0.69	0.75	
Decision Tree	Positive	0.73	0.47	0.57	0.65
	Negative	0.62	0.83	0.71	
KNN	Positive	0.60	0.67	0.63	0.61
	Negative	0.64	0.55	0.59	

Table 4.2: Results of Different Algorithm

Model	Score
Logistic Regression	0.87
Naïve Bayes	0.84
Random Forest	0.76
Decision Tree	0.65
Knn	0.61

4.3 Descriptive Analysis

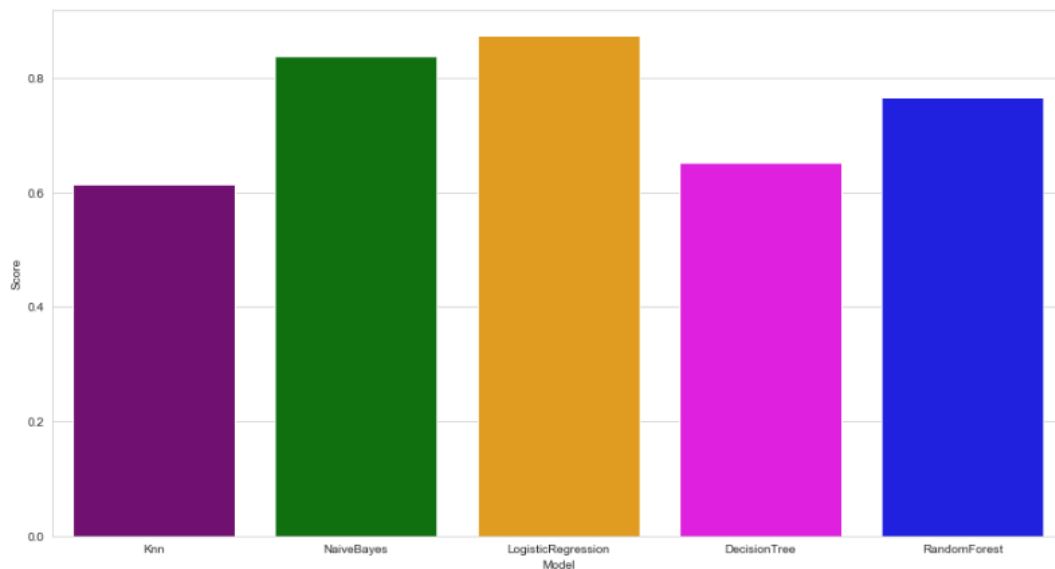


Figure 4.7 : Results comparison bar chart

4.4 Discussion

At the end we tried our best to find the best accuracy. Logistic regression gives us the most accurate results. In this study, a few currently used classification techniques have been discussed in terms of their accuracy for sentiment analysis.

4.5 Summary

Logistic Regression produced the best results. Logistic Regression achieved 87% accuracy, Random Forest achieved 76% accuracy, and K-Nearest Neighbor achieved 61% accuracy. We determined which algorithm produced the best sentiment analysis results on IMDb movie reviews.

CHAPTER-5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Social and social media are significantly impacted by sentiment analysis. Sentiment analysis has become an essential tool for mining social networks for information. It makes it evident how people feel or perceive your viewpoint, text, review, etc.

5.2 Impact on Environment

Environment and climate change are significantly impacted by sentiment analysis. Decision-makers and policymakers may build effective adaptation and mitigation plans to enhance the state of the planet by taking into account public opinion.

5.3 Ethical Aspects

The study of the feelings, viewpoints, and attitudes represented in written language is known as sentiment analysis. By employing algorithms to separate the emotions underlying the words, sentiment analysis may evaluate whether a communication implies a good, negative, or neutral attitude. Despite the fact that the technology is not new, headlines concerning sentiment analysis have led to legitimate questions about its ethics and veracity.

5.4 Sustainability Plan

The project's long-term sustainability is described in the sustainability plan. We may use sentiment analysis to ascertain what the general agreement among reviewers is about a certain movie or television program. This guarantees that the project's resources are not squandered. We have a great chance to use this study to discover how Netflix and other entertainment firms might profit from sentiment analysis.

CHAPTER-6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the Study

In our analysis, we use the Kaggle dataset. Our dataset was either integer or floating. So, first and foremost, we preprocess our dataset. To achieve the highest level of accuracy, we employ some machine learning models. And after five algorithms, we get the desired result. There was a desired model from which we obtained the most, and this is what we used as the desired maximum.

5.2 Conclusions

Our paperwork performs accurately. Every algorithm collaborates closely, and the results are usually accurate. We used five algorithms, and their performance is excellent. We discovered which algorithm produced the best results for sentiment analysis on movie reviews. we discover our expected outcome in a completely new domain with new parameters, and the accuracy is far superior to the rest of the project-related work.

5.3 Implication for Further Study

Long-term objectives include classify more human emotions using machine learning. We encountered some problems that are still unresolved. We will attempt to correct our proposed model's inconsistent accuracy problem. We can also try to make the algorithm more efficient so that training and compilation take less time. We hope it will bring another period the period of advanced technologies.

References

- [1] P. Baid, A. Gupta and N. Chaplot, "Sentiment Analysis of Movie Reviews using Machine Learning Techniques", *International Journal of Computer Applications*, vol. 179, no. 7, pp. 45-49, 2017. Available: 10.5120/ijca2017916005.
- [2] E. Kaur, "SENTIMENT ANALYSIS OF MOVIES REVIEWS USING IMPROVED RANDOM FOREST WITH FEATURE SELECTION", *International Journal of Advanced Research in Computer Science*, vol. 10, no. 3, , pp. p60-65. 6p, 2019. [Accessed 30 August 2022].
- [3] P. Muhammad, R. Kusumaningrum and A. Wibowo, "Sentiment Analysis Using Word2vec And Long Short-Term Memory (LSTM) For Indonesian Hotel Reviews", *Procedia Computer Science*, vol. 179, pp. 728-735, 2021. Available: 10.1016/j.procs.2021.01.061.
- [4] S. Shireen Khan and R. Pamnani, "International Journal of Innovative Research in Computer and Communication Engineering", *Analysis of Movie Reviews with Prediction of Movie Characters*, vol. 5, , no. 4, 2017. [Accessed 29 August 2022].
- [5] S. Harer and S. Kadam, "Sentiment Classification and Feature based Summarization of Movie Reviews in Mobile Environment", *International Journal of Computer Applications*, vol. 100, no. 1, 2018. [Accessed 29 August 2022].
- [6] N. Mohamed Ali, M. Abd El Hamid and A. Youssif, "SENTIMENT ANALYSIS FOR MOVIES REVIEWS DATASET USING DEEP LEARNING MODELS", *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 9, , no. 23, , 2019. [Accessed 29 August 2022].
- [7] O. Das and R. Balabantaray, "Sentiment Analysis of Movie Reviews using POS tags and Term Frequencies", *International Journal of Computer Applications*, vol. 9625, , 2014. [Accessed 29 August 2022].
- [8] K. Kumar, B. Harish and H. Darshan, "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, p. 109, 2019. Available: 10.9781/ijimai.2018.12.005 [Accessed 29 August 2022].
- [9] S. Prathap and S. Ahmad, "Sentiment Analysis on Movie Reviews", *IJSRSET*, vol. 4, no. 8, 2018. [Accessed 29 August 2022].

- [10] M. Kumar AV and N. Kumar AN, "Sentiment Analysis Using Robust Hierarchical Clustering Algorithm for Opinion Mining On Movie Reviews-Based Applications", *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. -8, no. -8, 2019. [Accessed 29 August 2022].
- [11] H. Shaziya, G. Kavitha and R. Zaheer, "Text Categorization of Movie Reviews for Sentiment Analysis", vol. 4, , no. 11, 2015. [Accessed 29 August 2022].
- [12] K. Amulya, S. Swathi, P. Kamakshi and Y. Bhavani, "Sentiment Analysis on IMDB Movie Reviews using Machine Learning and Deep Learning Algorithms", *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2022. Available: 10.1109/icssit53264.2022.9716550 [Accessed 29 August 2022].
- [13] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros and T. By, "Sentiment Analysis on Social Media", *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012. Available: 10.1109/asonam.2012.164 [Accessed 29 August 2022].
- [14] A. Mukwazvure and K. Supreethi, "A hybrid approach to sentiment analysis of news comments", *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 2015. Available: 10.1109/icrito.2015.7359282 [Accessed 29 August 2022].

Final Version Final Defense

ORIGINALITY REPORT

22%

SIMILARITY INDEX

19%

INTERNET SOURCES

10%

PUBLICATIONS

17%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Daffodil International University

Student Paper

7%

2

dspace.daffodilvarsity.edu.bd:8080

Internet Source

3%

3

Submitted to Manchester Metropolitan University

Student Paper

3%

4

www.hindawi.com

Internet Source

1%

5

Megha Raizada. "Anatomizing Text Without Training Dataset", 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), 2022

Publication

1%

6

Rifqi Ramadhani Almassar, Abba Suganda Girsang. "Detection of traffic congestion based on twitter using convolutional neural network model", IAES International Journal of Artificial Intelligence (IJ-AI), 2022

Publication

1%