SENTIMENT ANALYSIS OF MOVIE REVIEWS USING MACHINE LEARNING TECHNIQUES

BY

ABIR HASAN PIASH
ID: 191-15-2761
AND

AFIA JANNAT ANTORA
ID: 191-15-2512

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

Supervised By

Md. Sabab Zulfiker
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project titled "**SENTIMENT ANALYSIS OF MOVIE REVIEWS USING MACHINE LEARNING TECHNIQUES**", submitted by **Abir Hasan Piash**, ID No: **191-15-2761** and **Afia Jannat Antora**, ID No: **191-15-2512** 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 25-01-2023.

BOARD OF EXAMINERS

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

Dr. Md. Atiqur Rahman **Internal Examiner**
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Shayla Sharmin

Shayla Sharmin **Internal Examiner**
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

25-01-23

Dr. Dewan Md. Farid **External Examiner**
Professor
Department of CSE
United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Name, Designation, 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:

24.01.23

Md. Sabab Zulfiker
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

For, 26.1.23

Mohammad Jahangir Alam
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Abir

Abir Hasan Piash
ID: 191-15-2761
Department of CSE
Daffodil International University

Afia Jannat

Afia Jannat Antora
ID: 191-15-2512
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 Md. Sabab Zulfiker, Lecturer (Senior Scale),** 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 **Professor Dr. Touhid Bhuiyan**, 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

Movie reviews assist viewers in determining whether a film is worth their time. Sentiment analysis is the procedure of investigating digital text to calculate whether the emotional tone of a word is favorable, unfavorable, or neutral. In the proposed study we Used IMDb Dataset., Because IMDb one of the most well-known internet databases for movies and people. This gives users access to a huge and varied dataset for sentiment analysis. and make the data overwhelming numerous measures such as word clouds and text stemming methods. Natural language processing (NLP) takes employed toward develop the suggested prototype because movie comments lack grammatical structures, and experiments have been conducted to come up to the current investigation with already-existing learning model. We also applied some machine learning classifiers such as Logistics Regression (LR), Multinomial Naïve Bayes (MNB), Support Vector Classifiers (SVC), Decision Tree (DT), and Random Forest (RF). In addition, the proposed approaches are 5-fold cross-validation to obtain the accuracy rate as well as hyperparameter tuning in separate classifiers to allocate the finest parameters. The applied approaches presentation was assessed to regulate “Accuracy”, “Precision”, “Recall” and “f-score”. at what time all methods were likened, “Support Vector Classifier” gives uppermost correctness of 89.41%

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	**1-3**
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	2
1.4 Research Questions	2
1.5 Significance of the study	3
1.6 Limitations of the study	3
1.7 Report Layout	3
CHAPTER 2: BACKGROUND	**4-10**
2.1 Preliminaries/Terminologies	4
2.2 Related Works	4
2.3 Comparative Analysis and Summary	8
2.4 Scope of the Problem	10
2.5 Challenges	10

CHAPTER 3: RESEARCH METHODOLOGY	**11-24**
3.1 Research Subject and Instrumentation	11
3.2 Data Collection Procedure/Dataset Utilized	13
3.3 Statistical Analysis	13
3.4 Proposed Methodology/Applied	24
3.5 Implementation Requirements	
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	**25-29**
4.1 Experimental Setup	25
4.2 Experimental Results & Analysis	25
4.3 Discussion	29
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	**30-32**
5.1 Impact on Society	30
5.2 Impact on Environment	30
5.3 Ethical Aspects	31
5.4 Sustainability Plan	31
CHAPTER 6: SUMMARY, CONCLUSION, AND IMPLICATION FOR FUTURE RESEARCH	**33-34**
6.1 Summary of the Study	33
6.2 Conclusions	33
6.3 Implication for Further Study	33

REFERENCES 35-36

LIST OF FIGURES

FIGURES	PAGE NO
Figure 2.1. Emotion Analysis	4
Figure 3.1: Dataset in CSV format	13
Figure 3.2: Proposed Methodology	14
Figure 3.3: highest frequent words in favorable appraisals	16
Figure 3.4: common frequent words in Unfavorable appraisals	17
Figure 3.5: highest repeatedly words in Unfavorable appraisals	18
Figure 3.6: common frequent words in Unfavorable appraisals	19
Figure 3.7: logistic regression graph	20

LIST OF TABLES

FIGURES	PAGE NO
Table 2.1: Comparative analysis with previous work	8
Table 3.1: confusion matrix for binary classification	22
Table 4.1: Logistic Regression ML classification result	26
Table 4.2: Multinomial naïve Bayes ML classification result	27
Table 4.3: Support Vector ML classifier Result	27
Table 4.4: Decision Tree ML classifier Result	27
Table 4.5: Random Forest ML Classifier Result	28
Table 4.6: ML classifier score vs 5-fold Cross-Validation scores	28

CHAPTER 1

Introduction

1.1 Introduction

Movie reviews are a crucial tool for assessing a film's effectiveness while assigning a movie a number or star score lets us know about the measurable accomplishment or disappointment of a movie, a group of movie evaluations gives us a deeper qualitative perspective on the film's various elements. It is simpler for viewers to find what they're looking for when movies are categorized or categorized. Therefore, we need a system that categorizes cinemas by their type so that viewers may easily recognize the content. On social media, anyone may quickly and openly publish their ideas. Such social data can be examined and used in education to foster quality instruction. As "An individual with Favorable or Unfavorable moods," sentimentality is defined. Sentiment investigation is the act of removing this user-produced text information to ascertain users' attitudes regarding any specific person, commodities, or event. Everyone values opinions since they are critical in making decisions, whether for themselves, businesses and services, governments, or any other entity in the world. Nowadays, the majority of individuals investigate products online before making purchases. They read customer reviews before deciding what to buy. When businesses sought the people's or customers' feelings in the past, they secondhand to behavior studies and belief polls, which were expensive, time-consuming, and required human resources. These attitudes can be favorable, unfavorable, or even unbiased. Since 2000, judgment investigation has gained much popularity as a study topic. Compared to other reviews, including product reviews, sentiment analysis of movie reviews is thought to be more difficult. Finding phrases that express emotion and understanding the connection between textual reviews and their effects are both made possible by sentiment analysis. One such instance is the impact of online movie evaluations on box agency receipts. We reach-me-down movie evaluations as of IMDb, the most widely used source of celebrity, television, and film material, for our experimental evaluation. IMDb has appraisals for more than 3.5 million films. IMDb users rate movies and provide ratings for how beneficial they are. However, IMDbS' individual rating procedure, which in go is constructed on reviewers' ratings, determines a movie's overall rating. Concern over the methods used to deal with absent standards in the dataset is also developing. The suggested approach types usage of Natural Language Processing, and Machine Learning Classifiers.

1.2 Motivation

This study was done to ascertain the widespread acceptance of a certain movie. There are many cinemas released each year. However, not all movies are excellent. Despite the fact that there are many great movies, not all of them are worth sharing. Examine the film's overall evaluations. Analyze how people feel about the movies to help viewers comprehend them in depth and to help people pick the perfect movie for them. For this project, opinion research is the best approach. You can obtain public opinion both offline and online. Then, after receiving feedback, we might train our computers to determine the result.

1.3 Rationale of the Study

We survey the general population about their preference to see a particular movie using sentiment analysis. Most of the time, we are unaware of the specific drama, characters, settings, and sights that moviegoers enjoy. It is challenging to communicate with everyone. We might receive input from the general public by releasing top-notch films that will help people in the future study more or think more clearly. To make high-quality movies, we must therefore obtain audience opinion. We begin a study on feelings analysis of movie evaluations as a result. Among the Predictive Analysis classifiers, we employ are Logistics Regression, Multinomial Naïve Bayes, Support Vector Classifiers, Decision Tree and Random Forest.

1.4 Research Question

These are the research questions:

- What is Sentiment Analysis?
- How does sentiment analysis in film reviews work?

- Make a model that can determine whether a movie review is favorable or unfavorable?
- What issues does sentiment analysis in film reviews have?
- How can the outcome be more precise?

1.5 Significance of the study

The results of our study are crucial for film evaluations. In this research, we tried to use NLP and some machine learning models to categorize movie reviews. The goal of the project is to classify movie reviews more quickly and precisely.

1.5 Limitations of the study

There are a couple of restrictions through this project that we hope to get past in the future. We employed feature selection approaches and fewer classifiers than we would like to in the future.

1.7 Report Layout

- The project's objective, justification for the study, research question, meaning of the study, and study limitations are all outlined in Chapter 1
- Introduction, Related Work, and Research Summary in Chapter 2.
- contains a data set by means of associated data collecting and processing procedures in Chapter 3.
- System Design and Analysis in Chapter 4.
- The research's outcomes and Debate in Chapter 5.
- Summary, Assumption, and hereupon in Chapter 6

CHAPTER 2

Background study

2.1 Preliminaries

Sentiment analysis is the process of finding and distinguishing opinions, feelings, attitudes, perspectives, and evaluations from a specific input of data. The main goal of perception investigation is to categorize the division of a text sampled from a verdict or page. Results from user-submitted reviews are based on the majority of extracted positive, negative, and neutral attitudes.

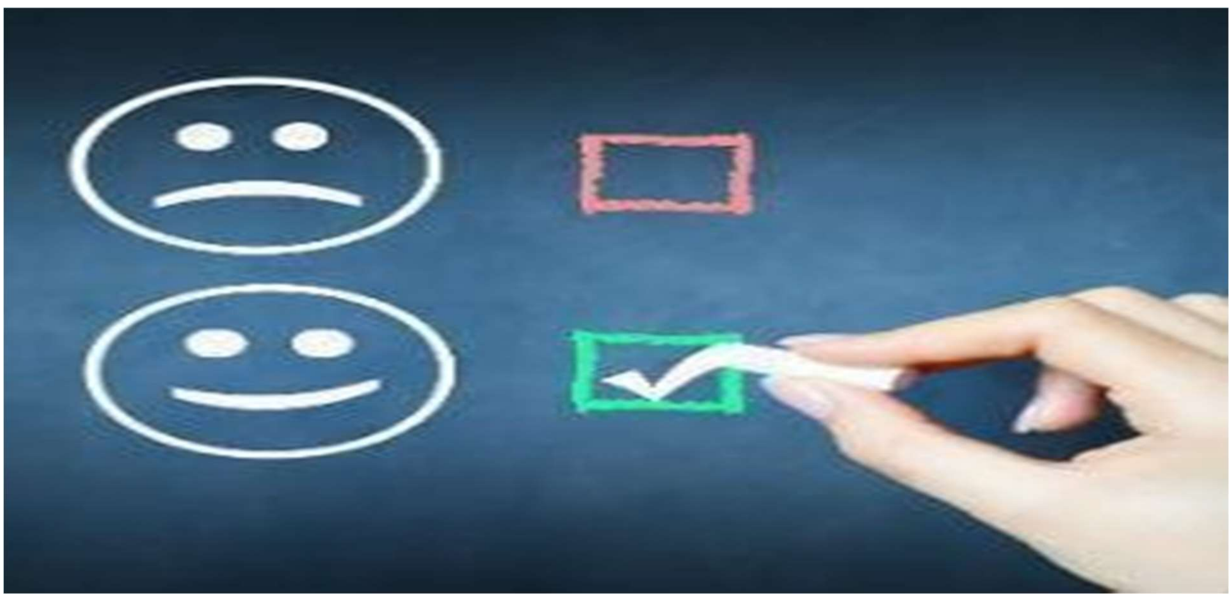

Figure. 2.1. Emotion Analysis

2.2 Related Works

In order to compare the results and have a conversation, we will look into several investigation analyses and some relevant journals:

Kumar et al. [1] used lexicons of the information from the IMDb Database, they collected information from the IMDb Movie Review Dataset and divided it into favorable and
information from the IMDb Movie Review Dataset and divided it into favorable and

unfavorable evaluations. After preprocessing the data, they applied a number of algorithms with different feature assortment approaches. With a maximum accuracy of 83.93%, Maximum Entropy is the best classifier, and Regularized Locality Preserving Indexing (RLPI) has the best accuracy at 74.66%. By chance, our approach and theirs are remarkably similar. They abruptly increased their use of feature selection and classifier methods, nevertheless, which we would want to ignore in our study.

Shaukat et al. [2] accurately and successfully identified the general viewpoint stated in movie reviews using a neural network model. The movie review database at Stanford University served as their source for the data. We used WordNet dictionaries. They used development strategies like neural networks. With a final accuracy rate of 91% and a validation accuracy rate of 86.67%, the network was able to train. In tests of comparison, their categorization model fared better than other methods.

Qaisar et al. [3] established a machine learning-based technique for detecting sentimentality in movie evaluations. Users rated movies on IMDb using a scale of 1 to 10, and the data was obtained from there and divided into training and testing vectors. They employed a variety of classifiers in their research, one of which is built on the recurring neural network (RNN) approach. They are capable of an accuracy rate of 89.9%. This method has the advantage of allowing for the suggestion of well-known films. Our investigation will make use of this model.

Sahu et al. [4] employed a lexical strategy. They used the dataset from rotting tomatoes on the polarity of reviews. The suggested model employs a number of categorization techniques, such as RF, DT, NB, etc. Bagging was the most accurate of these, with a precision of 88.57%, followed by Random Forest and k-next-door neighbors (KNN), both of which showed outcomes with an accuracy of 88.98%. Inaccurate sentiment analysis of opinion words by the present lexicons, such as WORDNET, is the fundamental weakness of the lexicon-based technique, which we seek to address in our study.

Yasen et al. [5] used Eight distinct classifiers and five dissimilar valuation metrics were merged. They gathered the data from the IMDb review archive. All techniques used for classification are under supervision. With a 96% accuracy rate, RF one of the eight does sentiment analysis with the highest level of effectiveness and accuracy. They only employed supervised learning algorithms, whereas we use both supervised and unsupervised learning methods to significantly expand our area of competence. Working with Weka restricts our productivity, which is a drawback for them.

Nagamma et al. [6] used a sentiment-aware autoregressive model, which can provide predictions in the review system that are more accurate. They obtained the information from the IMDB website. They employed clustering techniques like FUZZY Clustering. with numerous cataloging schemes. Clustering is not necessary for classification to get the same outcome. The correctness of the SVM and NB classifiers was equal. One of its key components is the request for various data removal techniques. However, their classification approach performed badly in a sizable sample.

Yenter et al. [7] inspect the sentimentality in movie evaluations They accumulated a dataset from the ACL online film dataset. To test the system, 50,000 evaluations are balanced with either a tag of 0 or 1. Several layers were applied after the data underwent preprocessing. The model's efficiency would have been far higher had they used more preprocessing techniques; still, they used fewer preprocessing techniques. which we want to become free of with our work.

Baid et al. [8] used machine learning and parts of speech approaches. We used the traditional n-gram, bi-gram, and POS bigram approaches. With a precision of 76.6%, the SVM Light classifier displayed the highest accuracy. In this research, they used a smaller dataset and fewer SVM Lite Kernels. This model's precision is subpar. which we are confident to overcome in our effort as part of our occupation.

The new hybrid classification strategy developed by Govindarajan et al. [9] performs well in terms of accuracy. The data set that is being utilized to evaluate the system consists of 200 movie reviews. After preprocessing the data, they used naïve bays and genetic algorithms. They worked on validation techniques. Accurateness of Arrangement Methods the Genetic Algorithm (GA), which is 91.25% precise, is shadowed by the NB-GA method in relation to correctness (93.20%). By a wide margin, the Hybrid classifier performs better than the Single classifier. Yet, they only observed at a minor dataset

Ali et al [10]. recommended practice for sentimentality examination was wished-for. The IMDb dataset, which consists of 50K movie review files (25K favorable and 25K unfavorable), all of which were written in English, was used to obtain the data. After preprocessing the data, they employed four deep neural networks: the multilayer perceptron (MLP), convolutional neural network (CNN), long short-term recall (LSTM), and cross-model. CNN LSTM outperformed the others with an accuracy of 89.20%, while CNN came in second with 87.70%. We'll use it to determine reviewers' feelings in our upcoming research.

Topal et al [11]. introduced an executive tool. They collected data from the IMDb dataset, which contains 157,344 reviews. They preprocessed the data before using the grouping procedure. K-means. retaining the k-means grouping techniques to categorize cinemas into sorts based on the thoughts of critics. In light of all the movie reviews, they redid the movie grouping. They make use of k-means procedures, which provide enormous dataset sizes.

Bandana et al. [12] proposed supervised learning methods, heterogeneous characteristics and features based on the Lexicon to develop this model. They used The Dataset, which was created from a variety of sources and was personally tagged. They preprocessed the data using supervised learning techniques like lemmatizing, stop words, and part of speech removal using linear SVM, as well as Naive Bayes. The exactness of NB, which is 89%, is the greatest of these. There were no other learning methods employed. By integrating supervised and unstructured approaches, we intend to greatly expand our field.

Chirgaiya et al. [13] formed a natural language processing classical to investigates the sentimentality in movie evaluations. They obtained the data from the Movie Review Dataset, which has 25.000 entries, and divided and described it using three classifiers. They used the Natural Language Toolkit to prepare the dataset for this study (NLTK). In their work, they employed a few supervised learning algorithms. With a score of 97.68%, the classifier Logistics Regression showed the greatest accuracy. For our research work, we are also attempting to build this model.

Pour Ansari et al [14] recommended a representative for sentimentality investigation using NLP. In addition to the IMDb dataset, which comprises 50000 reviews, they also employed a Multi-class dataset. As part of the preparation, they employ a Python tool to convert the transcript to a lesser case and eliminate stop arguments. For binary classification, they used various classification models. In this research, they used NLP Classification Models extensively. The closest classification method is SVM, and among those, classifiers using logistic regression have the highest accuracy (86.6%). Compared to other categorization models, their model performed better.

Different classification models were proposed by Lopez et al. [15] to predict the attitude of IMDb reviews. The IMDb movie reviews dataset served as the source for the datasets. In addition to Decision Trees (DT) and pre-processing methods like term frequency-inverse document frequency (TF-IDF), BOW, Binarization, Stop Words Removal, NGrams, fixed vocabulary, etc., they used a variety of algorithms, such as LR, MNB, NB, and Support SVM. The NB-SVM is the categorizing technique that is most accurate. If a lexicon was used, the model offers the best level of accuracy.

2.3 Comparative Analysis and Summary

Table 2.1: Comparative analysis with previous work

SL No	Author Name	Process	Strength	Result
1.	Kumar, Harish, and Darshan	Apply Cross Feature Abstraction Method	Accurateness of correctness	83.93%
2.	Shaukat, Zulfiqar and Mahmood	Dictionary and neural system	Qualitive investigation of the film many features	91%

3	Qaisar and S.M	Long Short-Term Memory (LSTM)	Explanation for up to date text-based sentimentality analyzers	89.9%
4	Sahu and Ahuja	Feature Selection and classification algorithms	Used diagonally the web.	88.95%
5	Yasen , Tedmori	Machine learning classifier	Used Numerous classifiers	96%
6	Nagamma, Pruthvi	TF and IDF with Fuzzy Clustering	They worked with double favorable and unfavorable expressions	89.65%
7	Yenter, Abishek	Deep CNN_LSTM with joint seeds	They applied dissimilar development branched	89%
8	Baid, Gupta	Machine Learning Techniques	Discovered the polarities of the evaluations	76.6%
9	Govindarajan	Cross techniques of Naïve Bayes and Genetic Algorithm	Used ensemble procedures	93.20%
10	Ali Hamid, and Youssif	Deep learning models	Deep learning with cross system.	89.2%
11	Topal, Ozsoyoglyu	Emotion Analysis	Eplains the moton of the movie emotioin charts	75%
12	Bandana, R	Heterogeneous Features, and Lexicon-Based Features	Work with two dissimilar datasets	89%
13	Chirgaiya	Natural Language Processing (NLP)	Outstanding Accuracy	97.68%
14	Pouransari	Deep Learning model	Associate different neural tensor systems	86.6%
15	Lopez, B	Different Classification Model	IMDb sentiment analysis	91.80%

2.4 Scope of the Problem

This research is helping to create a model that can recognize the positive and negative sentiment in movie reviews by utilizing data analysis, machine learning methods, and ensemble methodologies. This model can be used to anticipate the factors that influence moviegoers' decisions, which will help society address these issues. As a result, this model's only objective is to pinpoint the potential roots of banking problems. For sentiment analysis, we are creating a model utilizing machine learning and natural language processing. We therefore thought of creating a model that could predict movie reviews.

2.5 Challenges

This project is a significant challenge for us. There were many obstacles we had to overcome in order to complete the project as planned. Everything takes a very long time to complete, especially training. We need to give this a lot of time to discover people who can use the techniques taught by movie reviews to discern the sentiment.

CHAPTER 3

Research methodology

3.1 Research Subject and Instrumentation

Investigation of movie assessments via sentimental investigation is the focus included in our study.

Goggle Colab

Colab is essentially a totally mist-based, free Jupyter sketchbook atmosphere. Most significantly, Colab doesn't essential to be established, and the sketchpads you generate can have multiple team associates editing that one at once, like how you run documents in Google Docs dictionary. The fact that Colab cares about the most broadly castoff machine learning collections and that they are modest to weigh onto your notebook is its greatest advantage. It’s a held Jupyter notebook with countless allowed kinds that delivers free admission to Google dispensation incomes like GPUs and TPUs and necessitates no arrangement.

NumPy

The collection acknowledged as NumPy, or "Arithmetical Python," has multidimensional collection substances and methods for processing those collections. It also offers purposes for employment in the parts of media, the Fourier convert, and line algebra. In the year 2005, Travis Oliphant established NumPy. You can practice it for it is an exposed basis scheme. A universal- drive set for treatment collections is named NumPy. It proposes a multidimensional collection thing with unresolved rapidity as well as competencies for interrelating with these collections. It is the keystone Python unit for scientific calculating. The programmed is exposed basis. The goalmouth of NumPy is to offer collection substances that are up to 50 times earlier than conservative Python slopes.

Pandas

The most well-known Python programming language package for data operation and examination is called Pandas. Pandas provide information structures and processes for strong, versatile in order to user-friendly data analysis and manipulation as an exposed basis software library developed on the highest of Python exactly for information operation and analysis. Pandas enhance Python by enabling it to interact with data similar to spreadsheets, facilitating quick loading, aligning, manipulating, and merging in addition to other crucial operations. Data Frames, which are two-dimensional array-like data tables with one variable's values in each column and a set of those values in each row for each row, are part of the Pandas open-source package. A Data Frame may contain data of the character, factor, or numeric types. Data frames created with Pandas can alternatively be compared to a dictionary or a grouping of a series of objects.

Matplotlib

The large information arithmetical organization tool NumPy comprises the graphing set Matplotlib for Python programming linguistics. Conspiracies are entrenched in Python programmed by means of Matplotlib's thing concerned with API. Python's Matplotlib toolkit delivers a whole instrument for the construction of still, lively, and collaborative imaginings. Matplotlib makes problematic things conceivable and modest belongings informal.

NLTK

A group of public libraries and requests for statistical linguistic dispensation can originate in the NLTK (Natural Language Toolkit) Public library. One of the maximum powerful NLP libraries, it contains tackles that permit processers to understand humanoid language and reply suitably when it is cast off.

3.2 Data Collection Procedure/Dataset Utilized

IMDb dataset was used in our research. There are (50,000) movie reviews in the dataset. There are 25,000 favorable besides negative sentiment distributions in the instance. Our dataset is downloaded in CSV format and use for research purpose.

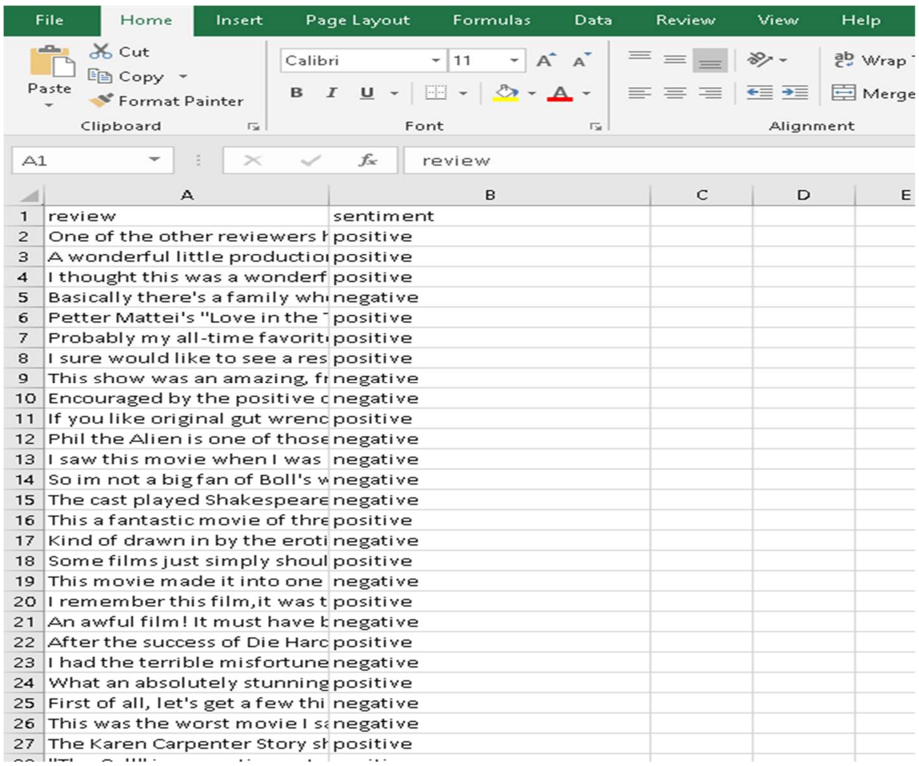

Figure 3.1: Dataset in CSV format

3.3 Statistical Analysis

- Our dataset has one type of data such as categorical.
- Dataset is saved in Microsoft excel which extension is CSV.

3.4 Proposed Methodology/Applied

Initially outline the training data, then compute the sentiment of the primary five evaluations, count the number of words in appraisals, remove duplicate data, and then apply text stemming. The following is a description of the remaining data preparation methods.

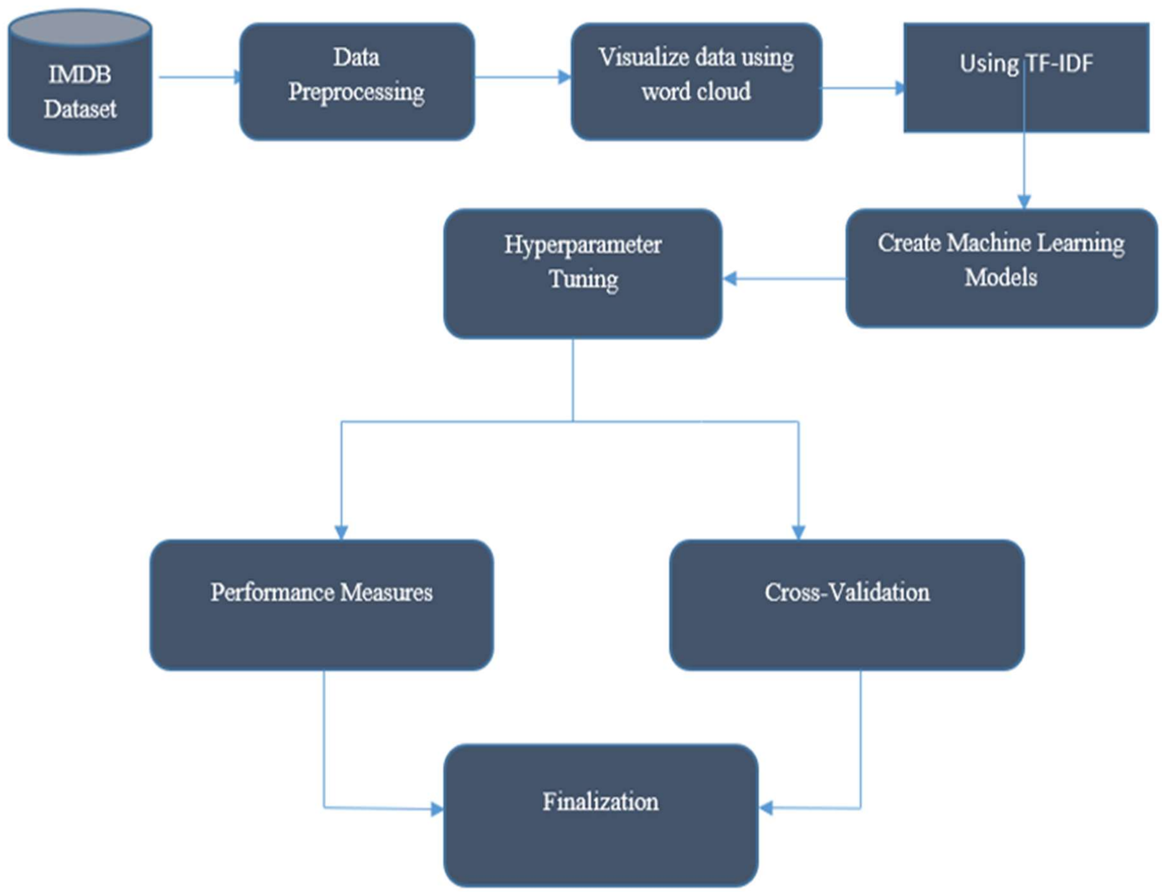

Figure 3.2: Proposed Methodology

Tokenization

Tokenization is the procedure of breach down an expression, verdict, subsection, or entire manuscript text into smaller parts, such as by way of certain words or concepts. Tokenization is used to break the string down into its component parts so that a machine can understand it. Tokenization is used to protect private information while preserving its usefulness for business purposes. We used a regular expression-based tokenizer to tokenize our material into manageable portions for our investigation.

Remove Stop words

A stop term is a commonly castoff phrase that a hunt query has erected set up to disregard, together whereas indexing admissions for penetrating and as soon as getting them as the consequence of a hunt inquiry. Arguments like "have," "a," "the," and "this" stay examples of stop words. We eliminated each stop word from the title of this work using the NLTK stop words list.

Text Stemming

By creating morphological changes to a root or base word, a stemming algorithm reduces the number of words. For example, "works," "work," and "worked" all come from the same word stem, which is "work," and all three meanings are synonyms for it. The lowering of the text categorization model happens quickly.

Word Count

The most frequent words in a text are displayed in a word cloud (also known as a tag cloud) in descending order of frequency. They provide a quick look of the most crucial terms in the text, including customer reviews, social television posts, and news stories.

Frequent Use Word

After tokenizing a text, the first figure we can calculate is the word frequency. By expression incidence, we designate the number of periods apiece symbolic happens in a text. When speaking around term frequency, we distinguish between types and tokens. Types are the distinct arguments in a corpus, whereas tokens are the words, including repeats. By means of the frequent use term technique, we verdict the greatest frequent usage word in favorable Reviews:

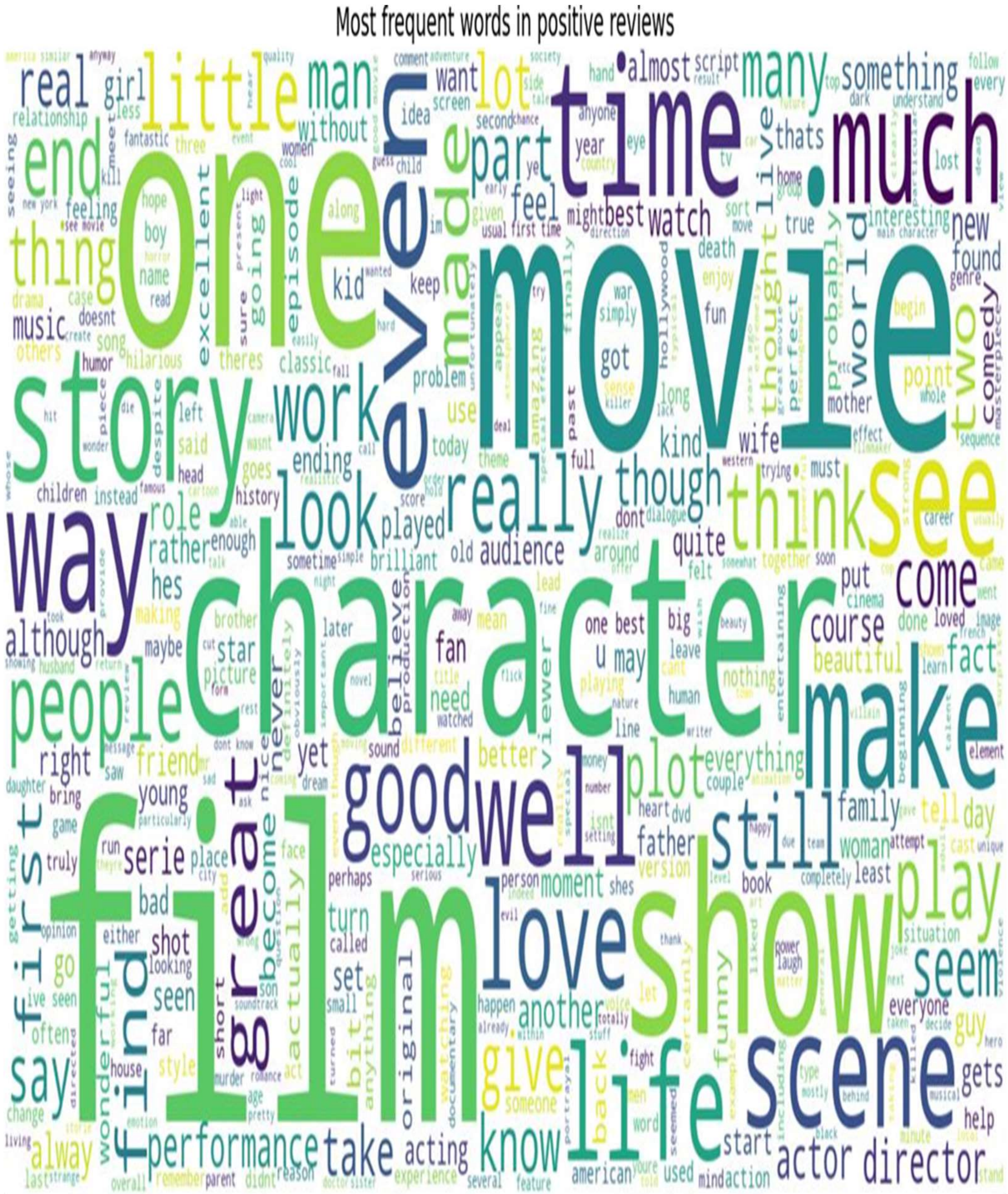

Figure 3.3: highest frequent words in favorable appraisals

Next, we classify usual words in Constructive evaluations:

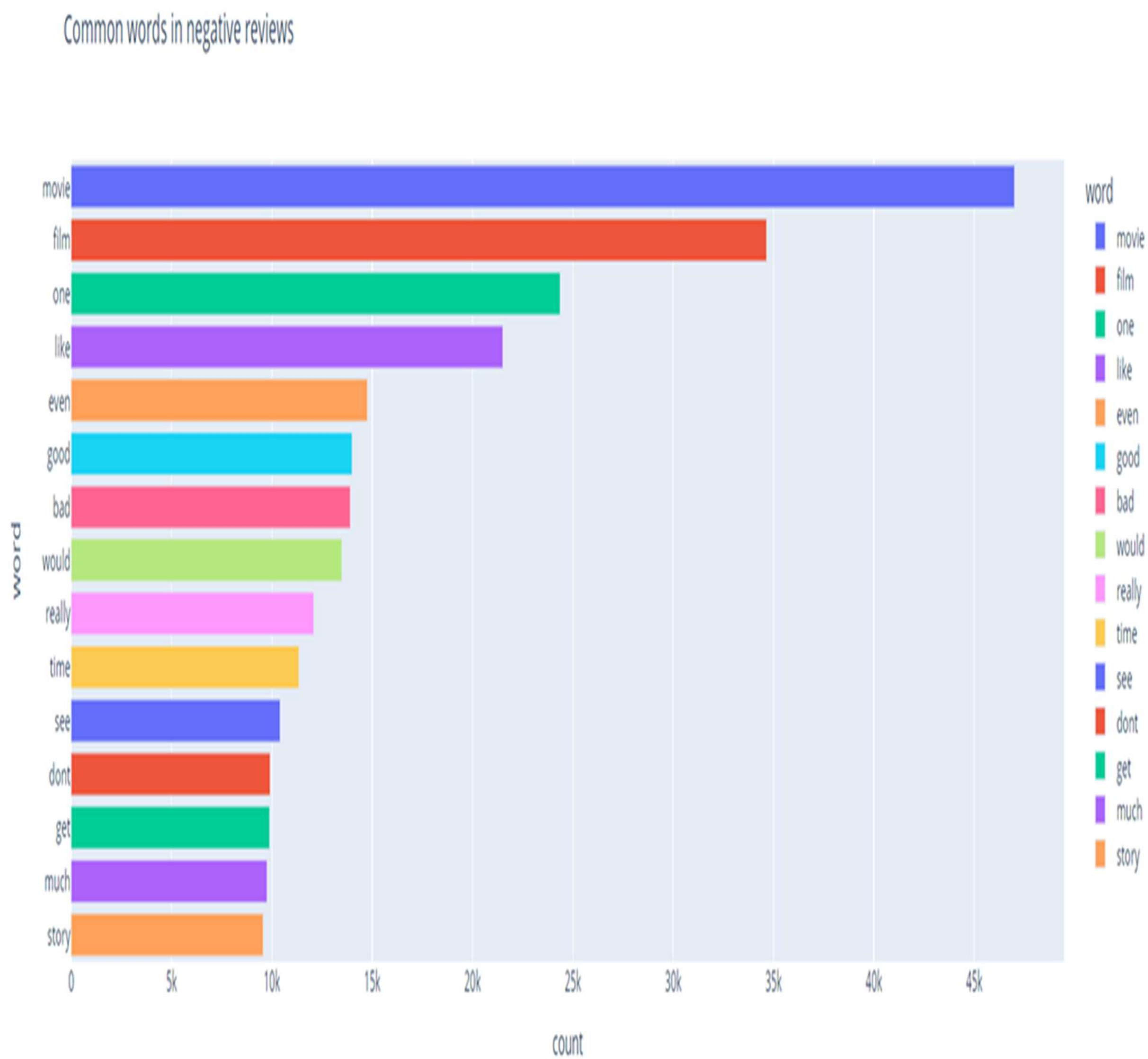

Figure 3.4: common frequent words in Unfavorable appraisals

We similarly discovered the greatest frequent usage term in Unfavorable evaluations

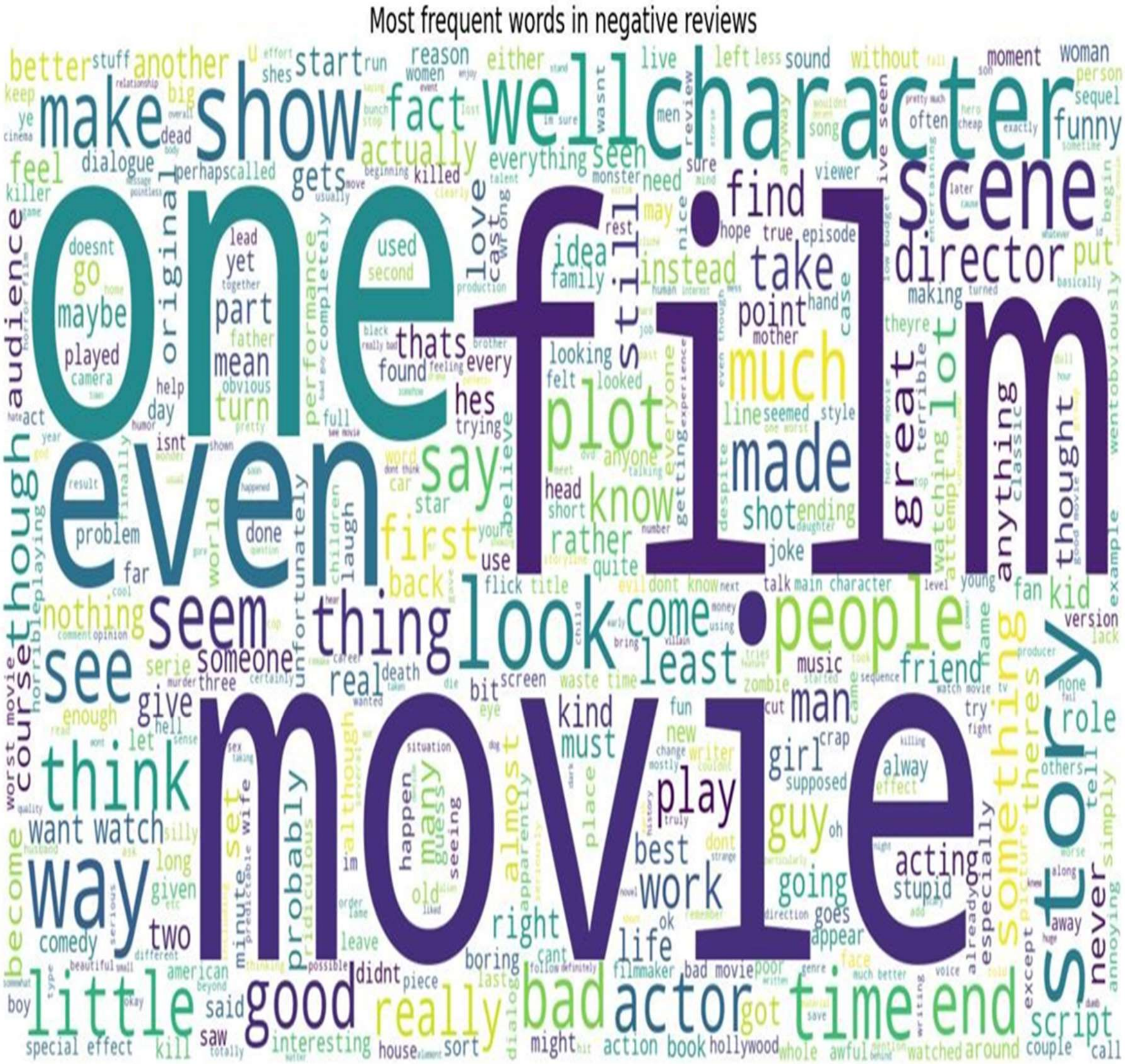

Figure 3.5: highest repeatedly words in Unfavorable appraisals

Next, we classify usual words in Unfavorable evaluations:

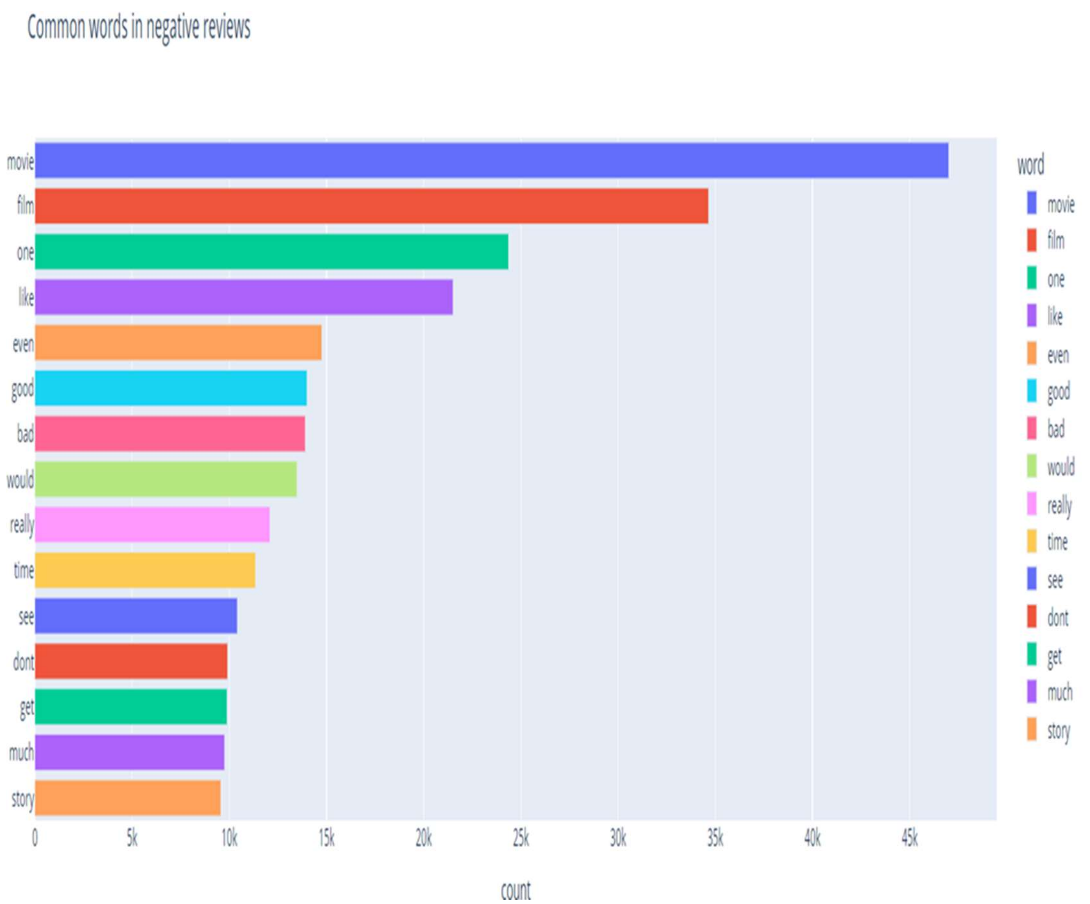

Figure 3.6: common frequent words in Unfavorable appraisals

TF-IDF

TF-IDF (term frequency-inverse text incidence) is an arithmetical amount that events how relevant a term is to a text in a group of papers. It is selfsame obliging for counting words in machine learning procedures for Natural Language Processing and has many applications, but robotic text analysis is its most important one (NLP).TF-IDF is intended for apiece term in a text by increasing two distinct metrics. the figure of periods a word seems in a text. The humblest process of decisive this occurrence is to only total the number of eras a word seems in a document. The distance of a text or the rate of the word that appears the greatest repeatedly in a document is another habit to adapt frequency. The term's regular opposite document rate of recurrence in a group of papers. This denotes how predominant or unusual a word is through all documents. A word is additional mutual the nearer it is to 0. By taking the total number of documents, dividing it by the entire number of documents covering a word, and then figuring out the logarithm, this metric can be gotten. We used a variation of ML Supervised Learning classifiers, counting logistics Regression, Multinomial Naïve Bayes, Support Vector classifier, Decision Tree, and Random Forest.

Logistic Regression model

This algorithm is helpful for our investigation because logistic regression employs a sigmoid purpose to speech the two-cataloging issue. In Logistic Regression replicas that comprise a dual condition, the outcome is typically defined as 0 or 1. When training a logistic regression model, logistic regression (LR) extracts relevant characteristics from text input. Finally, we considered how to evaluate the model's precision and predict how well it will function with hypothetical data.

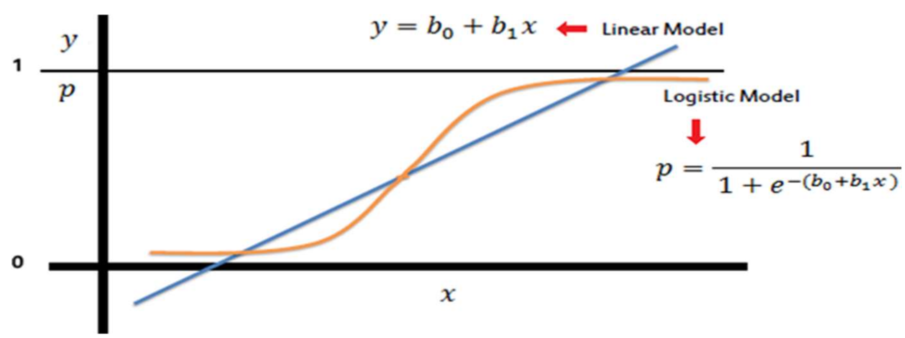

Figure 3.7: logistic regression graph

Multinomial Naïve Bayes

Multinomial Naïve Bayes applied when a multinomial distribution of the characteristics exists. Natural Language Processing uses the Multinomial Naive Bayes procedure as a probabilistic knowledge technique greatest commonly. The method, which deductions the label of a manuscript such as an electronic mail or paper article, is based on the Bayes theorem. For a given sample, it determines the probabilities of apiece tag and then produces the tag with the maximum likelihood. The Bayes theorem, industrialized by Thomas Bayes, controls the probability that an occasion will happen based on previously acknowledged circumstances.

SVC

A supervised machine learning method named the Support Vector Classifier (SVC) can be useful for organization or reversion glitches. Though classification subjects are anywhere it is most regularly applied. Once via the SVC algorithm, apiece information point is signified as a point in an n-dimensional interplanetary (where n is the number of topographies you have), with apiece feature's cost being the cost of a sure organize. Discovered a hyperplane in an N-dimensional planetary that evidently classifies the data opinions is the target of the SVC technique.

Decision Tree

The most active and well-liked method for categorization and forecast is the decision tree. A decision tree is a type of tree construction that looks like a flowchart, where each internal node characterizes a test on quality, each branch a test result, and each leaf node (terminal node) a class label.

Random Forest

When doing classification, regression, and other tasks, random forests, also referred to as random decision-making forests, are an ensemble learning technique. The class selected by the vast majority of trees in a random forest provides the output for classification issues. The mean or average forecast of the individual trees is returned for regression tasks.

Confusion Matrix

A technique for briefing an organization algorithm's presentation is the misperception matrix. If your dataset has additional than two lessons or if a piece class has an unsatisfactory quantity of comments, classification correctness unaccompanied may be deceiving. You can acquire a healthier sympathy for the classification model's achievements and disappointments by scheming a confusion matrix.

Table 3.1: confusion matrix for binary classification

		Predicted Values	
True Values		Negative	Positive
	Negative	TN	FP
	Positive	FN	TP

Dimension of recall, precision, specificity, and accurateness is abundant eased by this. The TN output, which stands for "True Negatives," displays the quantity of accurately classified negative samples. discloses how several positive cases were really confidential as positive. "FP" denotes a value that is a false positive. H. how many cases that were deemed "negative" were actually "positive." The term "false negative" (abbreviated "FN") refers to the amount of positive cases that were erroneously classified as negative. The following equation can be castoff to calculate the replica's correctness using the misperception matrix:

$$\text{Accuracy} = TP+TN/TP + TN +FP+FN$$

Hyper Parameter Tuning

We decided to apply hyperparameter tuning for our study purpose Verdict a set of perfect hyperparameter standards for a learning algorithm and by means of this adjusted procedure on any data set is hyperparameter alteration. The model's presentation is exploited by means of that set of hyperparameters, which reduces a prearranged damage meaning and subsequently cutting-edge better outcomes with fewer mistakes. Be aware that the learning algorithm attempts to discover the best possible solution within the specified constraints by optimizing the loss based on the input data.

Cross-Validation

We castoff cross-validation for our study. As a result, we cast off stratified K-fold cross-validation. The dataset was separated five times and hooked on similar folds. For a entire of 40 fits, the prototypical was constructed by appropriate 5 folds to apiece of the 8 potentials. The attention quality classes are uniformly distributed across all folds thanks to a stratified K-fold. We have tried a number of approaches. We go as far as we can to accomplish what we want. To assess a machine learning model's capacity to forecast new data, cross-validation is performed. It not only tourist attractions problems like overfitting and selection bias but also sheds light on how the model will generalize to dissimilar datasets.

Implementation analysis and design of our proposed system

Data collection is the initial phase of the cutting-edge technique. The IMDb (Internet Movie Database) Dataset used for this study was found on Kaggle. We preprocessed the dataset once we acquired it to make it more precise and useful for our purpose. Utilize a word cloud to visualize data after data preprocessing.

Then, in order to get the highest level of accuracy, we create machine learning models. Hyperparameter tuning comes next. Finding the ideal set of hyperparameter values for a learning algorithm involves using the improved procedure on any dataset. In order to generate better results with fewer errors, this set of hyperparameters maximizes model performance and minimizes the specified loss function. Finally, the final output is obtained through cross-validation and performance assessment.

3.5 Implementation Requirements

Software/Hardware:

- Operating system.
- Hard disk (minimum 500 GB)
- RAM (minimum 4 GB)

Developing tools:

- Colab environment.
- Google drive
- Good internet connection.
- Any browser.

CHAPTER 4

Experimental results and discussion

4.1 Experimental Setup

We worked on a database of 50,000 reviews from IMDB. First, we trained the machine using the train data. The test data is then taken for analysis.

4.2 Experimental Results & Analysis

TP): An outcome wherever the prototype correctly forecasts the favorable lesson is devoted to as a True positive.

(TN): Similar to a genuine positive, is an outcome aimed at which the model precisely predicts the negative lesson.

(FP): A false positive is a result when the model forecasts the positive class inaccurately.

(FN): A untruthful undesirable is a consequence where the classical predictions of the negative class are imprecise.

Precision: The correct optimistic to total true positive in addition to false positive ratio is known as precision. Precision checks to see how many false positives were included in the sample. If there aren't any false positives (FPs), the model's precision was 100 percent. The precision will appear uglier when additional FPs are extra to the mix.

$$\text{Precision} = TP/(TP+FN)$$

Recall: Recall takes a dissimilar path. Recall examines the number of mistaken rejections that were comprised in the prediction process somewhat than the number of untrue positives the model predicted. Each period a forecast false negative happens, the recall rate is punished. The equations themselves are opposites because of the penalties for precision and recollection. The yin and yang of evaluating the confusion matrix are precision and recall.

Recall = TP/ (TP+FP)

F-measure: Individual once exactness and recall are together 1 fixes the F1 Score turns out to be 1. The individual at what time exactness and recall are both a strong container for the F1 score rise. An additional valuable metric than accuracy is the F1 score, which is the vocal mean of recall and precision.

F-score = 2*((Precision*Recall) / (Precision + Recall))

Accuracy: The exactness is strongminded by in-between the entire number of correct guesses by the total amount of explanations in the dataset. The precision ranges from 0.0 to 1.0, with 1.0 existence the finest. It can also be strongminded by dividing by the ERR.

Accuracy = (TN +TP)/(TN+FP+FN+TP)

First, the machine learning model's results for logistic regression,

Table. 4.2.1 shows the model classification performance for logistic regression**.**

Table 4.1: Logistic Regression ML classification result

	Precision	Recall	F1-score	support
1	0.88	0.90	0.89	7513
2	0.90	0.88	0.89	7361
Accuracy			0.89	14874
Macro avg	0.89	0.89	0.89	14874
weighted	0.89	0.89	0.89	14874

Then, we Calculated another machine learning classifier result that’s called multinomial naïve Bayes (MNB). Figure 4.2.2. shows the models classification performance for multinomial naïve Bayes.

Table 4.2: Multinomial naïve Bayes ML classification result

	Precision	Recall	F1-score	support
1	0.87	0.90	0.89	7513
2	0.86	0.88	0.89	7361
Macro avg	0.86	0.89	0.89	14874
weighted	0.86	0.89	0.89	14874

Thirdly, we calculated another machine learning classifier result that's called Support Vector Classifier (SVC). Figure 4.2.3. shows the model classification performance for the Support Vector Classifier.

Figure 4.3: Support Vector ML classifier Result

	Precision	Recall	F1-score	support
1	0.89	0.90	0.89	7513
2	0.90	0.88	0.89	7361
Accuracy			0.89	14874
Macro avg	0.89	0.89	0.89	14874
weighted	0.89	0.89	0.89	14874

Fourthly, we calculated another machine learning classifier result that's called Decision Tree. Figure 4.2.3. shows the model classification performance for the Decision Tree.

Table 4.4: Decision Tree ML classifier Result

	Precision	Recall	F1-score	support
1	0.72	0.71	0.71	7513
2	0.71	0.71	0.71	7361
Accuracy			0.71	14874
Macro avg	0.71	0.71	0.71	14874
weighted	0.81	0.71	0.71	14874

Last of all, we calculated another machine learning classifier result that's called Random Forest. Table 4.2.3. shows the model classification performance for the Random Forest.

Table 4.5: Random Forest ML Classifier Result

	Precision	Recall	F1-score	support
1	0.84	0.87	0.85	7513
2	0.96	0.83	0.85	7361
Accuracy			0.85	14874
Macro avg	0.85	0.85	0.85	14874
weighted	0.85	0.85	0.85	14874

Table 4.2.1 denotes Machine Learning Model Score and Cross-Validation Test Scores of each classifier. Cross-validation is castoff to assess a machine learning model's ability to predict new data. For the ml classifier, the Support Vector Classifier Test score is 89.22% and the Cross-Validation score is 89.41% which is the highest.

Table 4.6: ML classifier score vs 5-fold Cross-Validation scores

Serial	Classifier	Machine Learning Model Test Score	5-fold Cross Validation Test Score
1	Logistic Regression	89.00%	89.00%
2	Multinomial Naïve Bayes	86.44%	86.44%
3	Support Vector Classifier	89.22%	89.41%
4	Decision Tree	71.10%	71.31%
5	Random Forest	85.23%	85.23%

4.3 Discussion

After using 5 machine learning classifier we got the best accuracy which was 89.41% for support vector classifier. this result is accurately and precisely calculated by this technology. It accurately predicts the movie goers for the detect the positive and negative sentiment of the movies

CHAPTER 5

Impact on society, environment and sustainability

5.1 Impact on Society

The possibility of a sentiment analysis for a movie review is predicted using machine learning approaches for sentiment analysis. The usage of a system for movie reviews may have a variety of effects on society. A movie reviews system may have the ability to enhance the effectiveness and precision of the movie review procedure. Sentiment analysis of a movie review system can assist in determining whether the movie meets your preferences by automating some steps of the process. The ability to help a viewer who hasn't seen the movie decide whether or not to see it is another potential effect of sentiment analysis on a movie review system. On the other hand, there are possible issues with the use of sentiment analysis in a system for movie reviews, such as the potential for biases or errors in the data used to train the system, which might result in movie reviews that aren't accurate.

5.2 Impact on Environment

The approaches utilized to create and implement a sentiment analysis of movie reviews system will have an effect on the environment. Generally speaking, the development and operation of a system for movie reviews using electronic or digital technologies, such as computer systems and machine learning algorithms, could potentially have a negative impact on the environment due to the energy and resources needed to power and maintain these systems. However, there are other ways that a sentiment analysis of a system for movie reviews could also benefit the environment. For instance, if the system is

successful in streamlining and automating the movie review process, it may help to lessen the need for some forms of paper-based documentation and communication, which may lessen the environmental effect of paper creation and disposal. The overall environmental impact of a movie reviews system will depend on how well the technology balances its potential positive and negative effects. These potential implications must be carefully considered, and action must be taken to reduce any negative effects.

5.3 Ethical Aspects

When creating and utilizing a system for movie evaluations, there are various ethical issues to take into account. The possibility of bias in the system is one ethical concern. Inaccurate or unjust movie selections may result from biased or unrepresentative data used to train the system. For instance, if a certain demographic group dominates the training data used to train the system, the system may be more likely to approve reviews for individuals who belong to that group and less likely to approve reviews for individuals who belong to other groups. In order to prevent bias, it is crucial to make sure that the data used to train the system is representative and diverse. The possibility for the system to be exploited to deny credit to particular groups of individuals is another ethical problem. The system's design could have detrimental effects on particular categories of people as well as society at large if it disproportionately rejects reviews from those groups. It is crucial to create a system that is impartial and fair, and to take into account any potential effects on underprivileged or marginalized groups. The system's transparency should be taken into account as another ethical factor. In order for borrowers to grasp the rationale behind the decisions and to contest any biases or errors, it is critical to be clear about how the system operates and how movie reviews decisions are made. In conclusion, it is crucial to give considerable thought to the moral implications of a system for movie reviews and to take action to make sure it is fair, impartial, and transparent.

5.4 Sustainability Plan

Here is a method for movie reviews that might be sustainable:

a) **Ongoing assessment:** The performance and correctness of the movie reviews system should be regularly assessed. This can be achieved by recurring assessments or by deploying a monitoring system that continuously tracks the system's performance.

b) **Regular updates:** To increase the system's efficiency and precision, fresh data and algorithms should be added on a regular basis. This can be accomplished by adding new machine learning algorithms or by updating the training data.
c) **System performance evaluation:** The system's performance should be evaluated, and any problems should be dealt with right away. This can be achieved either by a group of data scientists who are in charge of system maintenance or by automated monitoring systems.
d) **User input:** To enhance the system's functionality and usefulness, it should seek out and take into account user input. Surveys, focus groups, and other techniques for getting user feedback can be used to do this.

CHAPTER 6

SUMMARY, CONCLUSION & FUTURE WORK

6.1 Summary

We looked at a variety of NLP classification models. There were primary components to our investigations. With the use of data pre-processing methods including tokenization, steaming, frequent usage words, and TF-IDF, we were able to represent words numerically using the IMDb dataset. Following that, we applied a variety of Machine learning classifiers, including logistic regression (LR), multinomial Naïve Bayes (MNB), Support vector classifier (SVC), Decision Tree (DT), and Random Forest (RF). Then, we employed "Hyperparameter Tuning" to improve model accuracy while decreasing model complexity. We assessed the model's performance and validated its capability using cross validation and discovered that " SVC " offer the highest accuracy at 89.41%.

6.2 Conclusions

Techniques for sentiment analysis are one of the most important foundations in the decision-making process. For effective movie review outcomes, many people rely on sentimental analysis. IMDb movie reviews are categorized as either favorable or negative. 50k reviews in total are taken into account. at that point we cast-off a variety of classifiers to achieve data pre-processing based on methods for natural language processing. Hyperparameter tuning is the utilized to make sure the best parameters are used in our investigation. SVC has the maximum accuracy at 89.44%, according to our evaluation of the model's performance and cross validation of its capabilities. Without giving away any crucial information, such as the narrative or any surprises, our study would serve to provide enough information about the movie so that the reader could make an informed judgement.

6.3 Implication for Further Study

The development of more well-organized machine learning algorithms and the use of unsupervised machine learning on unknown samples could be the main topics of future study. Additionally, we wish to test out heterogeneous features. Furthermore, there are more scopes to take into account, including a live dataset and used more preprocessing methods.

Reference:

[1] Kumar, H.M., Harish, B.S. and Darshan, H.K., 2019. Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method. International Journal of Interactive Multimedia & Artificial Intelligence, 5(5).

[2] Shaukat, Z., Zulfiqar, A.A., Xiao, C., Azeem, M. and Mahmood, T., 2020. Sentiment analysis on IMDB using lexicon and neural networks. SN Applied Sciences, 2(2), pp.1-10.

[3] Qaisar, S.M., 2020, October. Sentiment analysis of IMDb movie reviews using long short-term memory. In 2020 2nd International Conference on Computer and Information Sciences (ICCIS) (pp. 1-4). IEEE.

[4] Sahu, T.P. and Ahuja, S., 2016, January. Sentiment analysis of movie reviews: A study on feature selection & classification algorithms. In 2016 International Conference on Microelectronics, Computing and Communications (MicroCom) (pp. 1-6). Ieee.

[5] Yasen, M. and Tedmori, S., 2019, April. Movies reviews sentiment analysis and classification. In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) (pp. 860-865). IEEE.

[6] Anagama, P., Pruthvi, H.R., Nisha, K.K. and Shwetha, N.H., 2015, May. An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. In International conference on computing, communication & automation (pp. 933-937). IEEE.

[7] Yenter, A. and Verma, A., 2017, October. Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis. In 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON) (pp. 540-546). IEEE.

[8] Baid, P., Gupta, A. and Chaplot, N., 2017. Sentiment analysis of movie reviews using machine learning techniques. International Journal of Computer Applications, 179(7), pp.45-49.

[9] Govindarajan, M., 2013. Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm. International Journal of Advanced Computer Research, 3(4), p.139.

[10] Ali, N.M., Abd El Hamid, M.M. and Youssif, A., 2019. Sentiment analysis for movies reviews dataset using deep learning models. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 9.

[11] Topal, K. and Ozsoyoglu, G., 2016, August. Movie review analysis: Emotion analysis of IMDb movie reviews. In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1170-1176). IEEE.

[12] Bandana, R., 2018, May. Sentiment analysis of movie reviews using heterogeneous features. In 2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech) (pp. 1-4). IEEE.

[13] Chirgaiya, S., Sukheja, D., Shrivastava, N. and Rawat, R., 2021. Analysis of sentiment-based movie reviews using machine learning techniques. Journal of Intelligent & Fuzzy Systems, 41(5), pp.5449-5456.

[14] Pouransari, H. and Ghili, S., 2014. Deep learning for sentiment analysis of movie reviews. CS224N Proj, pp.1-8.

[15] Lopez, B., Minh, A.N. and Xavier, S., 2019. IMDb sentiment analysis.

Final Plagiarism

ORIGINALITY REPORT

20%	17%	8%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	Submitted to Daffodil International University Student Paper	3%
3	Submitted to University of Westminster Student Paper	1%
4	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1%
5	dspace.lib.uom.gr Internet Source	1%
6	link.springer.com Internet Source	1%
7	sakilansari4.medium.com Internet Source	<1%
8	Rahul Kejriwal, Mohana. "Artificial Intelligence (AI) in Medicine and Modern Healthcare Systems", 2022 International Conference on	<1%