

**MACHINE LEARNING BASED SENTIMENT ANALYSIS ON MOVIE  
REVIEWS**

**BY**

**A. A. Md. Minhajur Rahman  
ID: 181-15-1889**

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

Supervised By

**Al Amin Biswas**  
Senior Lecturer  
Department of CSE  
Daffodil International University

Co-Supervised By

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



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**SEPTEMBER 2022**

## APPROVAL

This Project titled “Machine Learning Based Sentiment Analysis on Movie Reviews”, submitted by A. A. Md. Minhajur Rahman, ID No: 181-15-1889 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 Bachelor of Science in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-09-2022.



**Dr. S M Aminul Haque**  
**Associate Professor & Associate Head**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

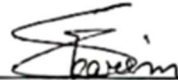
**Chairman**

## BOARD OF EXAMINERS



**Fahad Faisal**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



**Fourcan Karim Mazumder**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



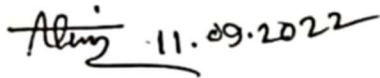
**Dr. Mohammad Shorif Uddin**  
**Professor**  
Department of Computer Science and Engineering  
Jahangirnagar University

**External Examiner**

## DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Al Amin Biswas, Senior Lecturer, Department of CSE Daffodil International University**. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

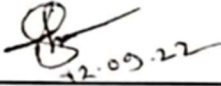
**Supervised by:**

 11.09.2022

---

**Al Amin Biswas**  
Senior Lecturer  
Department of CSE  
Daffodil International University

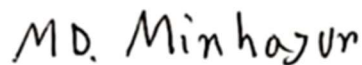
**Co-Supervised by:**

 12.09.22

---

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

**Submitted by:**



---

**A. A. Md. Minhajur Rahman**  
ID: -181-15-1889  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First i express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project successfully.

I am really grateful and wish my profound my indebtedness to **Al Amin Biswas, Senior Lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of my 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.

I am very grateful and wish my most profound debt to **Md. Sabab Zulfiker, Senior Lecturer**, Department of CSE, Daffodil International University, Dhaka. In-depth knowledge and in-depth interest of our manager in the field of “*Natural Language Processing*” to undertake this project. His unwavering patience, expert guidance, and enthusiasm, constructive criticism, valuable advice, and a lot of low draft learning and correction at all stages made it possible to complete the project.

I would like to express my deepest gratitude to my Parents, my Family, and the Head of the CSE Department “**Professor Dr. Touhid Bhuiyan**”, for his kind assistance in completing my project and for the other members of the faculty and staff of the CSE department of Daffodil International University.

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

Finally, I must acknowledge with due respect the constant support and patients of my parents.

## ABSTRACT

Sentiment analysis is a recent area of research where vast amounts of data are analyzed to offer insightful information about a particular subject. It is a strong instrument that can benefit organizations, customers, and even governments. Nowadays, analyzing the emotional content of movie reviews is popular. The approach heavily relies on textual emotion recognition. Analysts in Machine Learning (ML) and Natural Language Processing (NLP) have investigated a variety of approaches to execute the procedure with the highest level of accuracy. There are three stages to the sentiment analysis procedure for movie reviews. We must first gather reviews from online platforms. The collected data will then be analyzed. Finally, we will have all of the data that has been processed regarding the tone of those reviews. The results of the model analysis may aid the consumer in understanding how viewers feel about a certain film. To conduct this research, movie reviews were subjected to sentiment classification methods. For review sentiment classification, we looked at SVM, Naive Bayes, Logistic Regression, and K-NN, four supervised machine learning methods. Empirical results with a large number of reviews in the training dataset show that the SVM model performs better than the Naive Bayes, Logistic Regression, and K-NN approaches. The SVM method achieved an 87.46% accuracy, an 87.21% precision rate, and an 87.46% recall rate.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
 <b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-3</b>
1.1 Introduction	1
1.2 Motivation	1
1.3 Rationale of the Study	2
1.4 Research Questions	2
1.5 Output	2
1.6 Project Management and Finance	3
 <b>CHAPTER 2: BACKGROUND</b>	<b>4-7</b>
2.1 Terminologies	4
2.2 Related Works	4
2.3 Scope of the Problem	6
2.4 Challenges	7
 <b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>8-16</b>
3.1 Research Subject and Instrumentation	8
3.2 Data Collection Procedure/Dataset Utilized	8
3.3 Applied Methodology	8
3.3.1 Importing dataset	9
3.3.2 Data Cleaning	10
3.3.3 Feature extraction Using Vectorizers	10
3.3.3.1 Count Vectorizer	10
3.3.3.2 TF-IDF Vectorizer	11
3.3.4 Naïve Bayes Method	12
3.3.5 Support Vector Machine Method	12
3.3.6 K-Nearest Neighbors Method	13

3.3.7 Logistic Regression	14
3.3.8 Product perspective	14
3.3.9 Model Interfaces	15
3.3.10 Memory Constraints	15
3.4 Implementation Requirements	15
3.4.1 Hardware Requirement	16
3.4.2 Software Requirement	16
<b>CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION</b>	<b>17-25</b>
	17
4.1 Experimental Setup	17
4.2 Experimental Results & Analysis	24
4.3 Discussion	
<b>CHAPTER 5: IMPACT ON SOCIETY AND SUSTAINABILITY</b>	<b>26-27</b>
5.1 Impact on Society	26
5.2 Ethical Aspects	26
5.3 Sustainability Plan	27
<b>CHAPTER 6: CONCLUSION</b>	<b>28-29</b>
6.1 Summary of the Study	28
6.2 Conclusions	28
6.3 Future Work	28
<b>REFERENCES</b>	<b>30</b>

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 1: Methodology for sentiment analysis on movie reviews.	9
Figure 2: Review's word count for sentiment analysis on movie reviews.	18
Figure 3: Word Cloud (Positive) for sentiment analysis on movie reviews.	19
Figure 4: Word Cloud (Negative) for sentiment analysis on movie reviews.	19
Figure 5: ROC Curve for all applied models.	21
Figure 6: Confusion matrix generated by Support Vector Machine.	22
Figure 7: Confusion matrix generated by K-Nearest neighbor.	22
Figure 8: Confusion matrix generated by Logistic Regression.	23
Figure 9: Confusion matrix generated by Bernoulli Naïve Bayes.	23



## LIST OF TABLES

<b>TABLES</b>	<b>PAGE NO</b>
Table 3.4.1: The requirement of hardware for this project	16
Table 3.4.2: The requirement of software for this project	16
Table 4.2: Performance evaluation metrics for all applied models	24

# Chapter 1

## Introduction

### 1.1 Introduction

Sentiment analysis is the evaluation and classification of sentiment found in textual data using textual analysis tools. Sentiment analysis models take into account feelings and emotions (angry, pleased, sad, etc.), goals, as well as polarity (positive, negative, or neutral) (such as interested vs. not interested). Businesses utilize sentiment analysis in a number of contexts, especially for marketing purposes, because consumers are curious about whether movie reviews are favorable or negative. Among the uses are brand monitoring, market research, customer service, customer feedback, and social media monitoring ("sentiment analysis"). The ability to forecast their businesses' outcomes is a priority for many multinational corporations.

The sentiment analysis procedure includes tokenization, word filtering, stemming, and classification. Tokenization calls for the separation of text into discrete elements like words, numbers, and punctuation. The next stage is stemming, which is the act of removing prefixes and affixes from a word to convert it to its stem. After preprocessing, we use Nave Bayes, Support Vector Machine, and logistic regression to analyze the dataset. In this step, we select the best model based on accuracy. As a result, before labeling a movie as good or bad, we investigate and examine the factors that affect the ratings in our review text.

### 1.2 Motivation

The main purpose of my project is to detect and analyze movie reviews to understand overall sentiment of reviewer's in a summary. This will help the streaming service provider to know that a movie review is positive or negative. They will know the sentiment of their user's about a particular movie. It will reduce time spent reading movie reviews and learn thousands of sentiment analysis user in a movie. The data mining stream deals with Classification of status or text review As well as positive, negative and neutral.

### **1.3 Rationale of the Study**

This project's primary goal is to determine the underlying attitude of a movie review based on the assumption of textual data. In this project, we attempt to categorize whether or not a person enjoyed the film based on the reviews they provide for it. This is especially helpful for filmmakers who want to assess their overall performance based on reviews from critics and the money people are willing to pay for movies. The outcomes of this experiment can be utilized to develop a recommender system that gives viewers recommendations for movies based on their prior viewing history. Review Another use of this project would be to locate a group of viewers with similar movies taste (like or detest) (like or dislike).

This project aims to investigate several feature extraction approaches used in text mining, including keyword spotting, lexical affinity, and statistical methods, and comprehend their applicability to our issue. Along with feature extraction, we also examine various classifications and the effectiveness of approaches for representing various attributes. Finally, we arrive to a conclusion that combines feature representations and for present predicting jobs, classification techniques are the most accurate option.

### **1.4 Research Questions**

- To what extent will we be able to test our model?
- What machine learning methods exist for determining whether or not a movie review will be positive or negative?
- To what library we will resort to in order to purify our data collection?
- With what method will we achieve text vectorization?
- Which classification model will have the highest accuracy?

### **1.5 Outcome**

- Polarized of sentiment.
- Saved time
- In-depth data analysis
- Highest possible accuracy for our model
- Predict sentiment
- Optimization of Emotion

## **1.6 Report Layout**

My Report contains of the following elements.

- Chapter 1 Introduction: Introduction, Motivation, Rationale of the Study, Research Questions, Output, Report Layout.
- Chapter 2 Background: Terminologies, Related Works, Scope of the Problem, Challenges.
- Chapter 3 Research Methodology: Research Subject and Instrumentation, Data Collection Procedure, Statistical Analysis, Applied Methodology, Implementation Requirements.
- Chapter 4 Experimental Results and Discussion: Experimental Setup, Experimental Results & Analysis, Discussion.
- Chapter 5 Impact on Society and Sustainability: Impact on Society, Ethical Aspects, Sustainability Plan.
- Chapter 6: Conclusions.

## CHAPTER 2

### Background

#### 2.1 Terminologies

The performance of a movie can be evaluated in part by reading movie reviews. While a movie's numerical/star rating informs us of the quantitative success or failure of a movie, a collection of film reviews provides us with in-depth qualitative insights into the movie's many features. The strengths and weaknesses of a film are shown in a textual review. If we thoroughly examine a movie review, we may determine whether the overall quality of the film fits our expectations.

Reviewer we plan to apply sentiment polarity in this research to a collection of critic-provided film reviews. Try to determine whether they loved the movie or didn't like it in order to gauge their overall reaction. We want to be able to forecast the overall polarity of reviews using word relationships found in reviews.

#### 2.2 Related Work

We recognize that this is just one of many activities related to sentiment analysis on movie reviews. There are many projects like ours project. But those project are mostly done by peoples outside Bangladesh. Mostly they analyzed reviews on Hollywood movies. There are no data set available to analyze the sentiment of “Bangla Word”. Because of this Dhallywood movies could not be analyze by those models. I intent to add “Bangla” words on my Library to analyze the sentiment and to analysis reviews of Dhallywood movies. That’s why our model will help our local streaming company’s.

Yessenov et al. [1] suggested to use sentiment from movie reviews comment. Three machine learning-based algorithms were utilized to derive a non-linear relationship between box office and their revenue estimates, using audience opinion as both an input variable and a predictor.

Mitra et al. [2] suggested a rule-based approach using Lexion for sentiment analysis on movie review dataset, which would be more accurate than a strictly dictionary-based approach. The suggested approach has a 95% document level accuracy and an 89% sentence level accuracy.

Brar et al. [3] performed a technique for studying sentiment on movie reviews using supervised learning. They discovered that sentiment, sentimentwordnet, and sentimentlangnet are used to analyze review polarity

Reddy et al. [4] Sentiment polarity at the sentence level was determined by using logistic regression for classification. He trained his weak learners using a boosting algorithm applied to sub-tree-based decision stamps.

Agarwal et al. [5] sorted opinions using a variety of ML classifiers and feature extractors. Machine learning classifiers include Naive Bayes, Maximum Entropy, and SVM. The unigrams, bicograms, bigrams, and unigrams with portions of speech tags are all examples of feature extractors.

Go et al. [6] Make use of information gathered twitter. Learning technique for identifying emotions for approximating functions with discrete values, one can use a technique called decision tree learning, where the target function is represented by a decision tree. Trees with knowledge can also reformatted as a series of if-then rules for easier human comprehension. One of the most popular ways to learn is through widespread use makes it one of the most popular inductive inference algorithms. Assessing the creditworthiness of loan applicants is similar to practicing medical diagnosis.

Bhatt et al [7] make use of information from sources like reviews of amazon products, customer comments, and merchandise. Machine learning techniques are applied directly, together with a number of statistical feature selection methods. These results demonstrate that sentiment classification is not a task well-suited to machine learning algorithms .They prove that the presence or absence of a word is more informative of the content than the number of times a word is used.

Hu et al. [8] used a keyword-based method to analyze the content and determine the tone. The main terminology used in this method adjectives (such as awesome and awful) are used to express how someone feels. Indicators can be compiled in a list. By hand, with semi-automatic composition tools like WordNet, or through machine learning algorithms for determining which indicators are most useful based on a set of labeled examples from the target domain.

Elmurngi et al. [9] sorted reviews into categories to detect unfair reviews using machine learning techniques. This results show that the unigram-based model is not always superior, and that the optimal parameters for classification methods are data-driven.

Dang et al. [10] suggested a framework that integrates Machine Learning and Semantic Orientation techniques to produce more accurate results when classifying emotions. In particular, the learning algorithm classifiers made utilize the words' semantic orientations as an extra feature aspect.

Singh et al. [11] suggested a SentiWordNet-based method, selecting adverbs, adverbial phrases, and verbs, as well as n-gram features, as language features. As an additional measure, they have utilized our SentiWordNet technique to compute the document-level sentiments for each film examined, afterwards comparing these values to those acquired using the Alchemy API. Similar comparisons are made between the movie's emotion profile and the file sentiments result. The collected findings demonstrate that, in comparison to the simple file sentiment classification, our scheme generates a more accurate and targeted sentiment profiling.

### **2.3 Scope of the Problem**

While working on this project, I ran into few issues.

- Choose the right dataset.
- Finding the best library to clean dataset.
- Selecting the best classification model.
- Train my model in a proper way.

## 2.4 Challenges

- To collect data.
- Accuracy of model.
- Predict right sentiment of reviews.
- Enormity of data takes lot of time to train.
- Quality of my model.



## CHAPTER 3

### Research Methodology

#### 3.1 Research Subject and Instrumentation

Usually, opinions are voiced in support of something. The object of observation, for instance, may be a person, a thing, a service, an organization, or any combination of these. Subcomponents are also present. As a result, entities are referred to as objects in sentiment analysis. Because objects might contain sub-elements and attributes and are hierarchical in nature, classification uses attribute-based sentiment analysis algorithms. Consequently, it is challenging for the general public to understand this technical term (qualities or elements). A straightforward term is "quality." It is employed for sentiment or opinion extraction based on attributes. It might be said in a single sentence or several sentences. The opinion is determined by how the opinion word is positioned within the text. One or more opinion words may be present in a single phrase.

#### 3.2 Data Collection Procedure

The fact that movie reviews may be found on numerous websites, including IMDB, Rotten Tomatoes, Facebook, Instagram, Twitter, and many others, made this element of the project challenging. But I had to gather a ton of information. Due to the 25000 rows of reviews, I took my time. My training data were divided into two columns. the second column is for sentiment, while the first one is for the review text (Positive and Negative).

#### 3.3 Applied Methodology

The description is given below in figure 1:

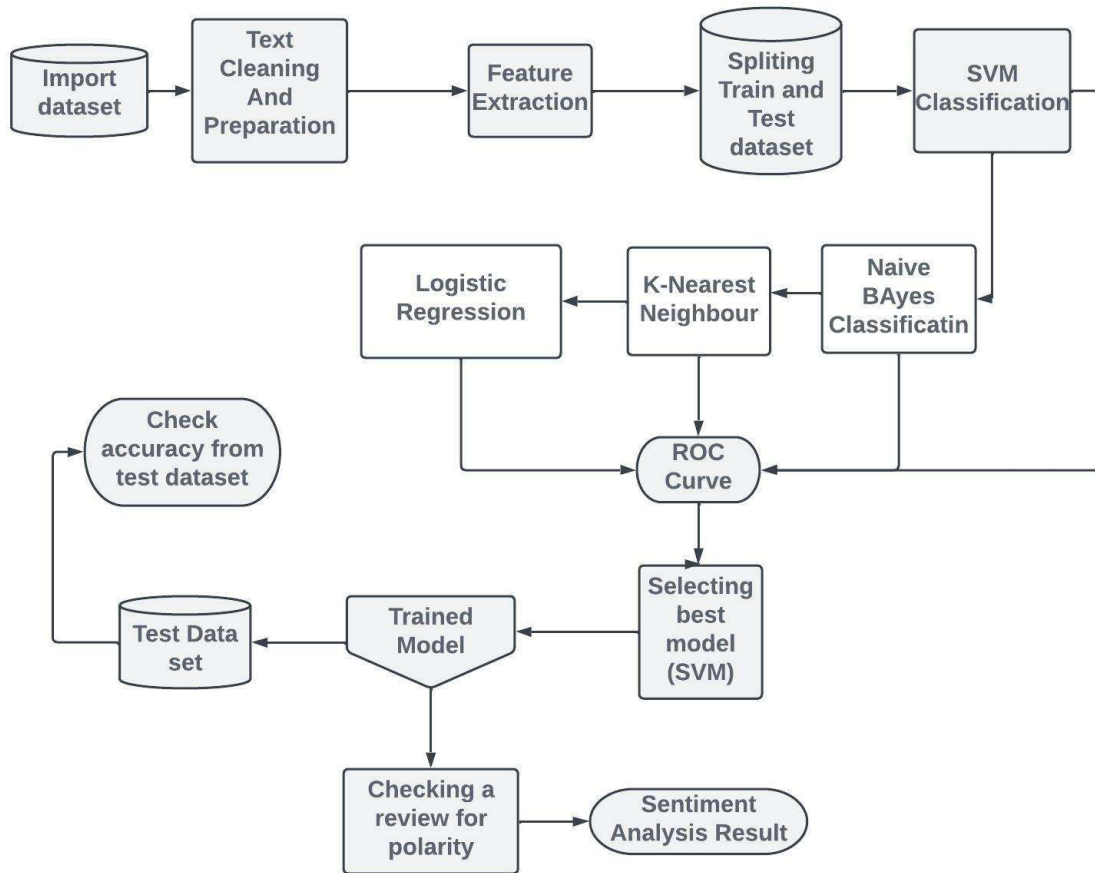


Figure 1: Methodology for sentiment analysis on movie reviews.

### 3.3.1 Importing dataset

The data science, data processing, and deep learning communities have largely adopted Pandas, an open-source Python library. It is based on the Numpy package, which is itself a multidimensional array library. Pandas is one of the most widely used packages for manipulating data, and it is included in nearly all Python distribution functions, from the OS-provided versions to commercial vendor distributions like ActiveState's ActivePython. Pandas is compatible with a wide variety of other machine learning modules within the Python ecosystem. So, I used it to bring in my training data from GitHub.

### **3.3.2 Data Cleaning**

The main activity is the data passage on the structure, which are:

- Deleting any punctuation, numerals, special characters, HTML elements, and emojis that are not part of the alphabet.
- Lower caseizing all of the letters.
- Eliminating any words that are only three characters long because they don't offer anything to the review overall.
- 'Stop words' Words like "to," "this," and "for" have been deleted because they don't accurately reflect the review's overall meaning and won't help processing.
- Adding meanings to emoticons and emoji in place of them.
- Lemmatizing all words to their original form – so that words like history, historical and historical are converted to their original words.
- History confirms that all these words are processed as the same word; Thus, their relationship becomes apparent to the machine.

### **3.3.3 Feature extraction Using Vectorizers**

Selecting an appropriate vectorial representation of the text files in order to run Machine Learning models is a major difficulty for any NLP Data Analyst. To forecast new text data, we often work with a large corpus of existing data, which may or may not be labeled, and then construct a machine learning (ML) model that can decode the underlying pattern from the strings' lexical content.

#### **3.3.3.1 Count Vectorizer**

In a word, this is how we implement Count Vectorizer. To determine the nature of a text from the frequency of its words, Count Vectors can be of great use. However, its main drawbacks include:

- The system can't prioritize which words to analyze and which to ignore.

- Simply said, it will only give statistical weight to the most common words in a given corpus.
- Connections among words, such as grammatical familiarity, are also not identified.

### 3.3.3.2 TF-IDF Vectorizer

The abbreviation for "Term Frequency - Inverse Document Frequency" is "TF-IDF." This metric not only measures how often a word appears in the corpus, but also provides a numerical estimate of the significance of that term. If you're looking for a better alternative to Count Vectorizers, consider using TF-IDF instead because it takes into account both the occurrence and significance of words inside the corpus. The complexity of model construction can then be decreased by eliminating the less crucial phrases for analysis.

TF-IDF is determined by the following equation (i):

For a term  $i$  in document  $j$ :

$$w_{ij} = tf_{i,j} * \log\left(\frac{N}{df_i}\right) \quad (i)$$

$tf_{i,j}$  = number of occurrences of  $i$  and  $j$ .

$df_i$  = number of documents containing  $i$ .

$N$  = total number of documents

The TFIDF algorithm is predicated on the idea that both extremely common and extremely uncommon terms in a sample are statistically irrelevant when looking for a pattern. Words that are over- or under-represented in a corpus receive low tf-idf scores thanks to the Logarithmic component in tf-idf.

### 3.3.4 Naïve Bayes Method

Naive Bayes is a classification method that uses word frequency distributions as input to determine the likelihood that a document falls into one of several categories. Phrase-and-category probabilities taken together. It presupposes that each word stands on its own. The Bayes theorem for conditional probability states that, for a provided data point  $x$  and class  $C$  by following equation (ii).

$$P(C / x) = \frac{P(x/e) \cdot P(C)}{P(x)} \quad (\text{ii})$$

In addition, if we assume that the probabilities of all of the data points in the set  $x = \{x_1, x_2, \dots, x_j\}$ . Since the likelihood of  $x$  appearing in a class is proportional to the number of  $x$ 's, not the number of attributes occurring in that class by following equation (iii).

$$P(C / x) = P(C) \cdot \prod P(x_j / c) \quad (\text{iii})$$

For this reason, the possible values of each variable must be determined before a Naive Bayes classifier can be trained. Taking place on the expected classes, estimates of which can be made using the trained data set. Classifiers using Naive Bayes produce respectable outcomes and profit from straightforward probabilistic interpretation. In an effort to convey emotion in this study, a feature selection strategy based on the categorization of online movie reviews was chosen.

### 3.3.5 Support Vector Machine Method

In order to classify linear data, the support vector machine technique might be used. It performs the transformation via a non linear mapping. A higher dimensional space for the training data. Looking for the linear ideal inside the higher dimension split the hyper plane in two. A hyper plane always provides the optimal separation between data sets from different categories. Because of the SVM, hyper plane with the help of some vectors. Russia's Vladimir Vapnik, Switzerland's Bernhard Boser, and France's Isabelle to be Guyanese in 1992. Traditional text classification is where SVM really shines. Technique for grouping things together. When a linear function is used to separate the two groups, a resulting hyperplane can be used to partition the solution. Space. Points can be divided into two groups using the graph below.

The number of possible hyper planes between the two points in the preceding example is infinite courses. At first glance, it may seem like picking a hyper plane with the largest area

under the curve would be the best option. distance between any two points in either group, maximizing the gap between them there are a total of 68 lessons. The reasoning for this approach is because a "safer" hyper plane has a greater margin. prediction mistakes due to being too close to a classification border are reduced. courses. The goal of the SVM algorithm is to locate such a hyper plane. The solution to this problem can be expressed mathematically as follows in the equation (iv) and (v).

Let  $c_j \in \{1, -1\}$  (correlating with the negative and positive) be the right kind of paper  $d_j$ . For  $C1$  and  $C2$  levels and data vectors  $X = \{x_t, r_t\}$  where:

$$R_t = +1 \text{ if } x_f * C^1 \tag{iv}$$

$$R_t = -1 \text{ if } x_f * C^2 \tag{v}$$

Using a constant  $w_0$ , determine  $w$  in such that the derivative of  $w$  and the input vector in the following (vi) and (vii) equation.

$$w_t * x_t + w_0 \geq +1 \text{ if } x_t \in C1 \tag{vi}$$

$$w_t * x_t + w_0 \geq -1 \text{ if } x_t \in C2 \tag{vii}$$

The preceding equations show that  $C1$  and  $C2$  points are located on different sides of the orthogonal a splitting hyper plane is obtained by taking the hyper plane described by the vectors  $w$  and maximizing its length  $\|w\|$ . Having the greatest possible separation among elements of either group. This discovery is supported by the data presented in Kuat Yessenov et al [1]. In optimization terms, finding this ideal hyper plane is a quadratic problem, the difficulty of which is determined by the train vectors,  $N$ , but not data set's dimensions. This leads to an intriguing conclusion: trained SVM model only considers neighboring data points predictions by dividing the hyper plane: they are indeed the support vectors, and they are typically found in much smaller regions. algorithm with better simulate the effect than using the complete dataset.

### 3.3.6 K-Nearest Neighbors Method

In the area of categorization, nearest-neighbor techniques are among the most accessible and successful. Use of algorithms. Their guiding principle is predicated on the hypothesis that, for each given collection of training set instances, if there is something new that hasn't been seen before, its category will probably be the same as the bulk of its "neighbor" events. Drawn from the sample data used for the training. Thus, the k-Nearest Neighbor technique is effective because it looks at the k nearest instances in the dataset. Application of a

preexisting classification data set to a novel event that needs to be categorized, followed by an assumption based on the classes to which the vast majority of  $k$  neighbors can be found. Distance functions between two points provide a formal definition of proximity. Two attribute values that are given to the algorithm beforehand as a parameter. Using an illustration of separation standard Euclidean distances between two locations in an  $n$ -dimensional area, where  $n$  is the dimension of the space being considered. For this data set,  $n$  represents the total number of observable characteristics.

### **3.3.7 Logistic Regression**

In this specific scenario, logistic regression is employed. That this study falls under the category both the sigmoid activity and the lack of prediction as opposed to the naive bayes algorithm and consequently we can improve our precision. This is an example of a binary logistic model. Comprises a dependent variable and two possible values, like the pass/fail dichotomy shown by the indicator variable has two predetermined values: the numbers "0" and "1" in mathematics. Calculating the log-odds for that the logistic model's "1" value corresponds to a linear interplay of a few different factors, It could be either a binary or continuous variable. Variable that doesn't stop. The resulting probability the "1"-valued quantity can swing between 0 and 1. As well as 1, which is why it's labeled that way; the logistic function, it is known to convert log odds to probability in its wake. A binary logistic regression uses the variables to the regression model contains two tiers. Multinomial logistic regression is used to model outcomes with while ordinal logistic regression is used when there are three or more possible if there are a lot of categories, you can use regression to order (such as the ordinal of proportional odds) logistic regression model. This model does not engage in statistical categorization, but can be applied to the development of a classifier by setting a threshold for input classification Those with a likelihood above the threshold make up one group, whereas individuals whose odds are below the threshold as A typical approach to constructing a binary system is to use the terms "other" and "otherwise."

### **3.3.8 Product Perspective**

Users now have a simple way to grasp the sentiment of a review thanks to sentiment analysis in movie reviews. A trustworthy model with greater than 85% accuracy will be available to determine the tone of a review. My methodology is simple for users to use and will produce the desired outcome.

### **3.3.9 Model Interfaces**

This model can run on Google's Colaboratoty, Jupyter notebook, anaconda. The training dataset can be linked on every software mentioned.

### **3.3.10 Memory Constraints**

This program doesn't have any memory restriction even after I used 25,000 reviews to training my model.

## **3.4 Implementation Requirements**

The most important stage is to choose and select the ideal software. Think about what this model can achieve and what it actually is before deciding whether or not there is a problem with software engineering.

I had conversations with specific persons and looked up information on numerous programs' websites in order to acquire this model. I've discussed every phase of my project with my manager.

I used the Scikit Learn library to preprocess the data. A Python module called Scikit-learn integrates a wide range of modern machine learning methods for medium-scale supervised and unsupervised situations. This package specializes on bringing machine learning to laypeople by employing a high-level, general-purpose language. Usability, general performance, documentation, and API consistency are prioritized. It is shipped with a simplified BSD license and has few dependencies, which promotes its use in both commercial and educational settings.



### 3.4.1 Hardware Requirement

TABLE 3.4.1: THE REQUIREMENT OF HARDWARE FOR THIS PROJECT

Proxemics Area	Time of Target
Processor	Any Computer of modern era can run this model
Motherboard	Any modern eras motherboard
Ram	Minimum 2 Gigabyte
Internet Card	Internet Cards of any kind
Graphics Card	Video cards of any kind
Hard Disk	Minimum 50 Gigabyte
Casing	Any Type
Monitor	Colored Monitor
Keyboard	Any type
Mouse	Any type

### 3.4.2 Software Requirement

TABLE 3.4.2: THE REQUIREMENT OF SODTWARE FOR THIS PROJECT

Software	Usage
Any windows and Mac operating system	To manage all the tools, programs, and equipment while running the computer.
Google Colaboratory	To run and Train and test the model.

## CHAPTER 4

### Experimental Results and Discussion

#### 4.1 Experiment Setup

We built a Python program that executes the experiments and handles the results, based on the description from section 2. The program can be found here `!git clone https://github.com/Minhaj1889/preprocess_sprinter.git` on github. It makes use of NLTK 0.9.9 together with its WordNet bindings, machine learning tools, and movie review corpus. The packages `nltk.*`, `nltk.cluster`, `nltk.classify`, `nltk.corpus.wordnet`, and `nltk.corpus` movie reviews are specifically used. Additionally, we utilized the Python numerical library `numpy` and `optparse` to parse command-line inputs. Additionally, we utilized the `matplotlib` package to draw graphs automatically.

#### 4.2 Experimental Results & Analysis

Calculating the average size Review is one place to start when working with review text to have some understanding of the caliber of the reviews. An average of 120 words are used in each review. The graphs below show the variation in word counts for each review based on this data, which we estimate is typical of those who write lovely descriptive evaluations. Sentiment analysis is a good topic for movies and as such. Also, individuals frequently write reviews when they have strong sentiments about a movie. They either loved it or loathed it.

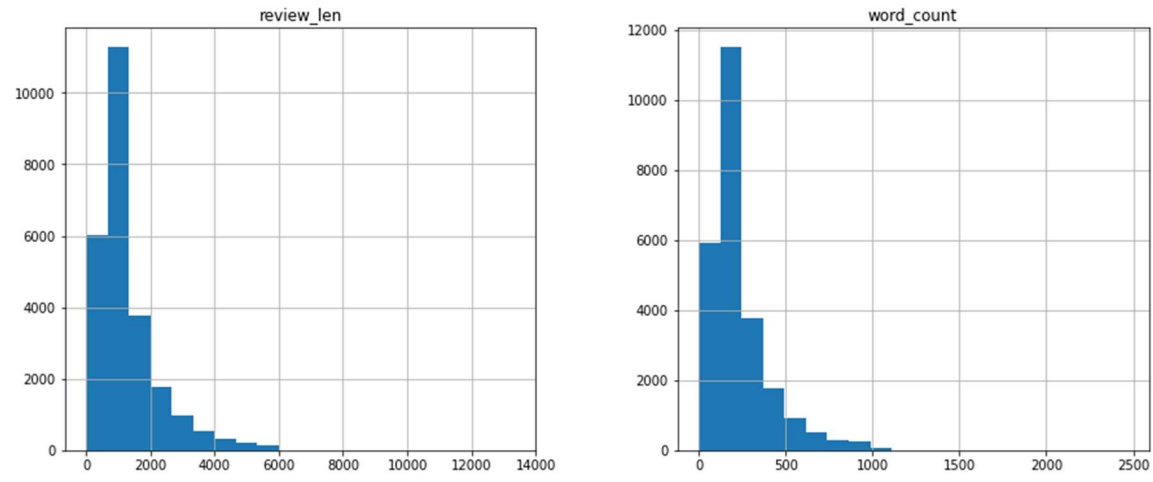


Figure 2: Review's Word Count for sentiment analysis on movie reviews.

Another intriguing metric turned out to be the number of phrases that appeared throughout the evaluation, in addition to the phrase count congruent with the abstract. Given their relative importance, some phrases are more common than others. contains a graph illustrating the variation in word occurrence among reviews as well as a list of the 20 phrases that appear the most frequently in both low and high quality reviews. In addition, there were almost 33 10,000 evaluations of each sentence on average overall. The fact that both high and poor reviews contain an excessive amount of the same phrases makes it evident from all of these logs and the graphs below that "bag of words" is not a very good version for analyzing the sentiment of reviews. The fact that there are a lot of unique words on average across all views (63,353) is also negative, thus we typically utilize between 10,000 and 40,000 unique words during training. This information also guides us to various feature extraction techniques like n-gram modeling and phrase-based TF-IDF computation.



The total number of sentences for each word across all tests is first tallied. The initial feature set was developed using the top 50,000 words as the dictionary's total word count increased to over 160,000 words. However, using 100,000 keyphrases, a different set of features can be produced in the same manner. Additionally, we produce any other word bags that show how all phrases that appear at least twice across the complete dataset are used. This guaranteed that the majority of the misspelled terms were eliminated.

Additionally, while some terms do not contribute to classification, others that appear just once in the data set might. Along similar lines, another characteristic artwork was created, but this time, the words appeared at least five times. These power 2 representations have lengths of around 34,000 and 76,000, respectively.

The TF-IDF Model: When using the two feature extraction techniques mentioned above paying more attention to the review's high-frequency sections and disregarding the rest possibly less frequent components that are more crucial to the overall polarity.

We create sound feature representations using TFIDF to account for this review. This model's feature representation is similar to the Bag of Words model in that we utilize TF-IDF values rather than frequency counts for each word. We disregard all of the frequent words used in both favorable and negative evaluations. Since they won't add much to the classification, there were more than 50 terms.

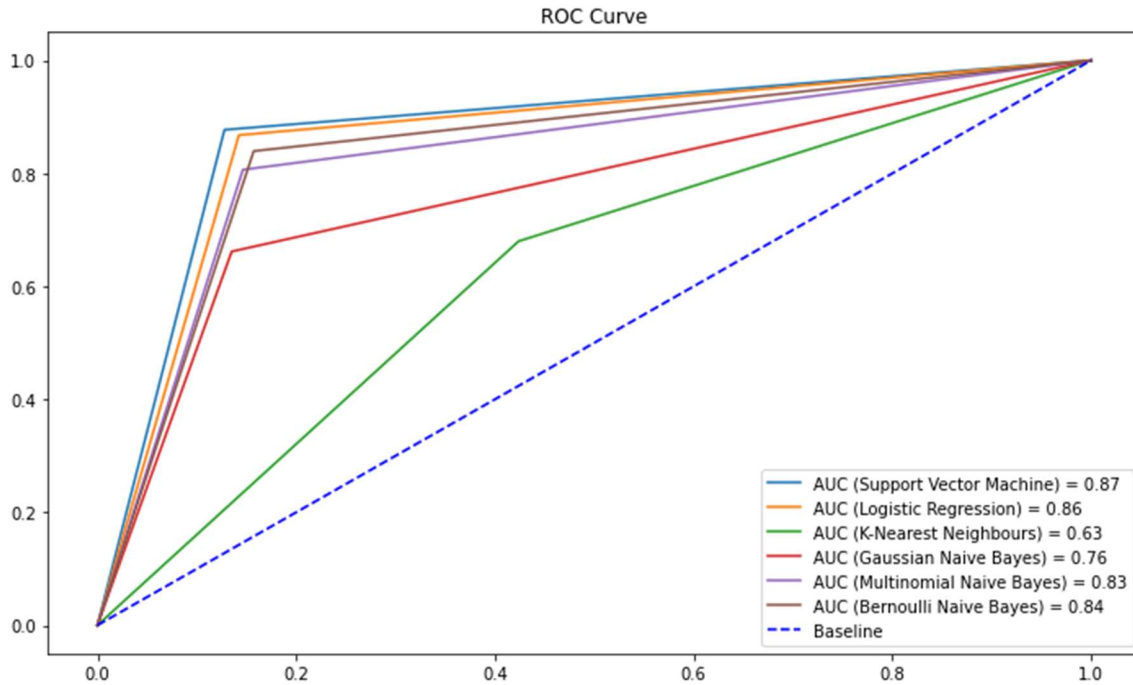


Figure 6: ROC Curve for all applied models.

In the following graph, we can observe that after training all of my classifier models, Support Vector Machine (SVM) techniques have the highest accuracy.

After testing trained model (SVM) performance with test data set. The accuracy score we got is 87.46%.

Opinion mining's efficacy can be measured in terms of accuracy, precision, and recall. Here in the context of sentiment analysis, accuracy refers to the general reliability of the models used. The ratio and mean of recall (Pos) and accuracy (Pos) accuracy rate of verified positive feedback. The ratio and precision ratio for Negative Recall and Negative Precision are, respectively. Honest criticisms, not sugarcoated. Every experiment's results should be evaluated using the equation (viii), (ix), (x) and (xi). Including precision, accuracy, and recall and f1-score.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{(viii)}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{(ix)}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{(x)}$$

$$F1\text{-score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (\text{xi})$$

I calculated the confusion matrix of Support Vector Machine (SVM), K Nearest Neighbor, Logistic Regression and Bernoulli Naïve Bayes model from 5000 data to analyze the True Positive, False Positive and True Negative, False negative numbers.

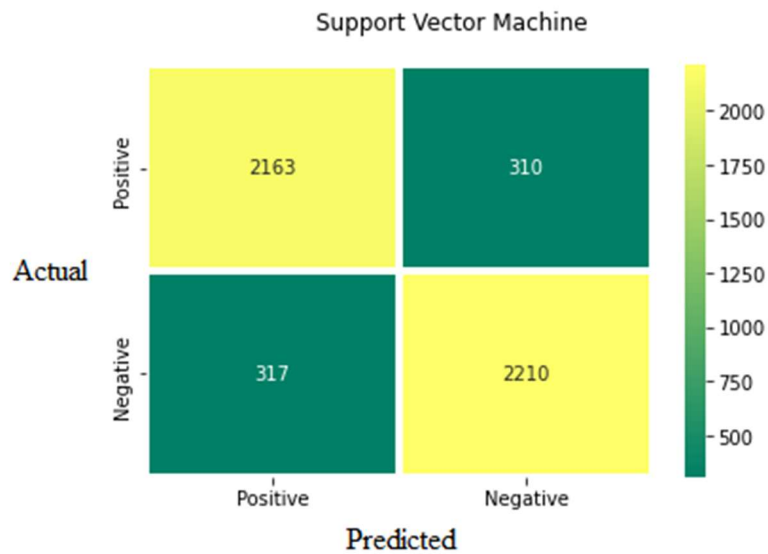


FIGURE 6: CONFUSION MATRIX GENERATED BY SUPPORT VECTOR MACHINE.

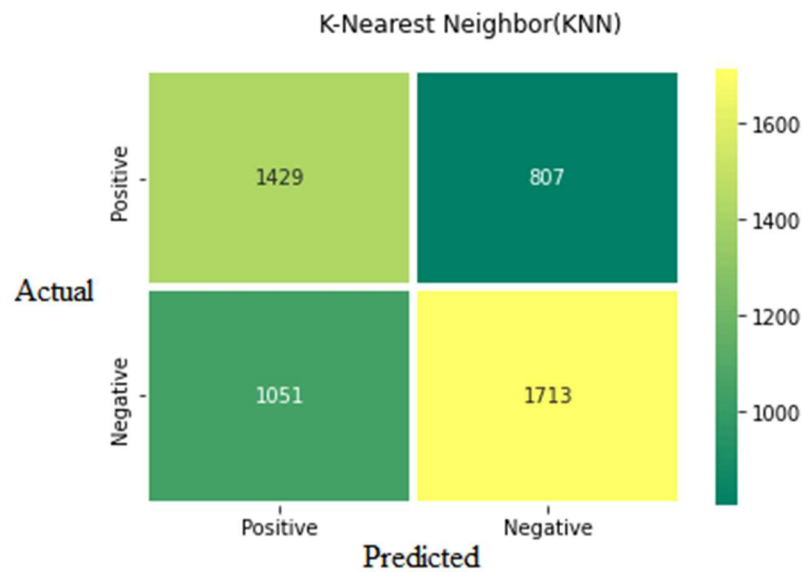


FIGURE 7: CONFUSION MATRIX GENERATED BY K-NEAREST NEIGHBOR.

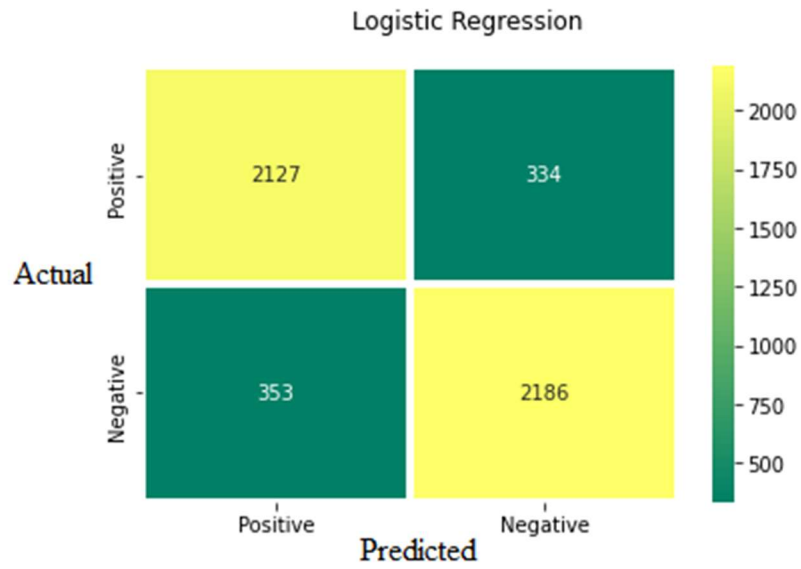


FIGURE 8: CONFUSION MATRIX GENERATED BY LOGISTIC REGRESSION.

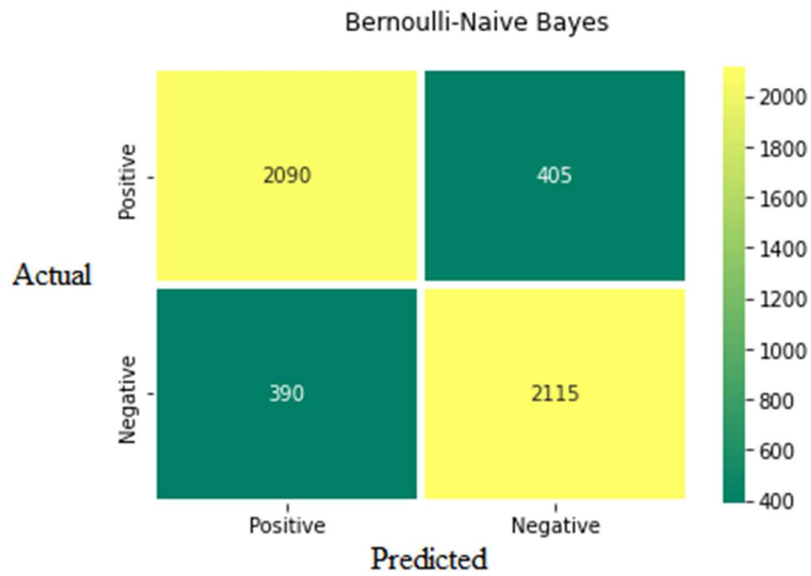


FIGURE 9: CONFUSION MATRIX GENERATED BY BERNOULLI NAÏVE BAYES.



TABLE 4.2: PERFORMANCE EVALUATION METRICS FOR ALL APPLIED MODELS

Models	Accuracy	Precision	Recall	F1-score
Support Vector machine	87.46%	87.21%	87.46%	87.34%
Logistic Regression	86.26%	85.76%	86.42%	86.10%
K-Nearest Neighbor	62.84%	57.62%	63.90%	60.60%
Bernoulli Naïve Bayes	84.10%	84.27%	83.76%	84.02%

SVM, Logistic Regression, Naive Bayes, and k-NN are the four supervised machine learning methods used in this research. The critical assessments of films posted on the internet. It has come to our attention that machine learning algorithms that have undergone extensive training can achieve remarkable results. Categorizations of the positive and negative reviews of films. The SVM algorithm can achieve high levels of precision in terms of obtain a rate of accuracy in categorization greater than 87%. For cases when the size of the training data set was only 25,000 or 200 reviews, there was a huge gap in quality between different algorithms. An extensive training set containing each of the three algorithms will perform better in sentiment categorization for reviews with 50,000 to 70,000 words. Critical analyses of the films we see.

### 4.3 Discussion

However, by choosing very few words (only the common maximum), we can also lose the textual content. Properties that may be required for the correct class. For instance, the accuracy declines if we utilize basic phrases that occur more than 10 times in a corpus. The upper limit for adequate accuracy for bigger companies, like movie reviews, is between eight and thirteen instances of the phrase. Our original algorithmic model for aspect-level sentiment classification provides a novel and distinctive approach to assembling an entire reviewer's emotional response to a film from its many individual reviews. A user-friendly, comprehensive, and insightful sentiment profile is produced. Aside from these benefits, the algorithmic formulation utilized for aspect-level sentiment profiling requires no prior

training, is easy to implement, yields results quickly, and has few to no further requirements. It's easy to use on the fly, and the resulting sentiment profile of a film is both detailed and valuable. Content-filtering, cooperative, and mixed movie proposed methods can all benefit from this operational detail.

## CHAPTER 5

### Impact on Society and Sustainability

#### 5.1 Impact on Society

The ability to analyze a large amount of data is made possible by machine learning skills. It can take more time and consistent resources to adequately educate you, even though this typically yields faster, more accurate answers that better comprehend opportunities or hazards. The usage of machines for processing substantial volumes of statistics can be increased further by combining them with artificial intelligence and cognitive technologies. My project's goal is to find and examine movie reviews in order to comprehend the reviewers' overall sentiments in a summary. This will enable the streaming service provider to determine whether a movie is receiving favorable or unfavorable reviews. They will be aware of how their users feel about a particular film. You'll spend less time reading movie reviews and discover how thousands of users participate in sentiment analysis for each film. The classification of text status or revisions, as well as positive and negative features, are all covered by the data mining flow.

#### 5.2 Ethical Aspects

Another issue is that these machine learning algorithms may function as "black boxes," making it difficult to understand how they function. It is impossible to determine the reasoning behind the machine learning algorithm's selection.

Predicting the sentiment of any product, in my instance movie reviews, is one application of machine learning. An algorithm can study a data set to figure out and forecast the tone of a certain review. We don't know how a machine learning system decides whether a review is favorable or bad, but a human can explain why the sentiment is either positive or negative. Therefore, we must rely on machine learning algorithms in all analyses even if we are aware of their limitations. The primary issue with machine learning is this. Even if it were 100% accurate, we wouldn't fully understand how the outcome was determined or how the algorithm predicted the particular outcome.

### **5.3 Sustainability Plan**

When the general public uses it professionally, the sustainability gets better and better every day. If required, we will continue to add features. And improve accuracy over the next days. In order to make the appropriate prediction, the library will also receive a lot more trendy words.

## **CHAPTER 6**

### **Conclusion**

#### **6.1 Summary of the Study**

Movie reviews are divided into good and negative poles for my project. A sizable library of movie reviews can be categorized using the system suggested in the project. The system's web-based API for sentiment analysis of movie reviews with JSON output to display results on any operating system is its strongest feature. My accuracy table shows that the system works decently. This will make it easier for the filmmakers to assess the state of your film. This API can be educated for other revisions in the future, such as those of smartphones, computers, clothing, etc.

#### **6.2 Conclusions**

With this research, we hope to measure how well existing sentiment classification methods perform in terms of accuracy. In this research, we evaluated four supervised algorithms for machine learning of varying degrees of precision and recall. Use of support vector machines, Naive Bayes, Logistic Regression, and k-Nearest Neighbors for sentiment categorization of 12500 positive and negative movie reviews. There were 12,500 processed negative comments. Results from the experiments prove that the SVM method is superior to SVM outperformed Naïve Bayes, Logistic Regression, and K-NN on the training dataset, which contained many reviews. Achieved Accuracy higher than 87%.

#### **6.3 Future Work**

Verifying if a hybrid technique may be employed by using the aforementioned classifiers in succession and combinations is a future area of research. Increase your accuracy. Rather than stopping at merely counting the amount of likes, comments, and follows, upcoming sentiment classification will strive to get to the heart of what customers are trying to convey through their online interactions. This outlook also forecasts expanded uses for sentiment classification, including but not limited to companies, personalities in the public eye, governments, charities, educational institutions, and many more. Additionally, I want to

develop a library for "Bengali words" that can parse text data in any language, including "Bengali words." This model can also be applied to the sentiment analysis of customer reviews for goods and services.

## References:

- [1] Kuart Yessenov, Sasa Misailovi, “Sentiment Analysis of Movie Review Comments”, Journal, 6.863 Spring 2009 final project, pp, 17<sup>th</sup> May 2009.
- [2] Ayushi Mitra,” Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)”, Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020), Vol.02/ No.03, pp, 2022.
- [3] Gurshobit Singh Brar, Prof. Ankit Sharma, “Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques” ISSN 0973-4562 Volume 13, Number 16 (2018) pp. 12788-12791, pp, 2018.
- [4] P. Sujan Reddy et al, “Sentimental Analysis using Logistic Regression”, Journal, ISSN: 2248-9622, Vol. 11, Issue 7, (Series-II), pp. 36-40, July 2021.
- [5] Apoorv Agarwal et al, “Sentiment Analysis of Twitter Data”, Journal, Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38, , 23 June 2011.
- [6] Alec Go et al, “Twitter Sentiment Analysis”, Conference, CS224N - Final Project Report, pp, June 6, 2009.
- [7] Aashutosh Bhatt et al, “Amazon Review Classification and Sentiment Analysis”, Journal, International Journal of Computer Science and Information Technologies, Vol. 6 (6), pp , 2015.
- [8] Hu and B. Liu, “Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining”, Conference, pp 168–177. ACM, 2004.
- [9] Elsharif Ibrahim Elmurngi and Abdelouahed Gherbi,” Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques”, Journal of Computer Science 2, vol 14 : 714.726, pp, 2018.
- [10] Yan Dang, Yulei Zhang, Hsinchun ChenA, “Lexicon Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews”, Department of Management Information Systems, vol. 25, no. 4, pp. 46-53. July 2010.
- [11] V.K. Singh et al, “ Sentiment Analysis of Movie Reviews A new Feature-based Heuristic for Aspect-level Sentiment Classification”, Conference, pp, March 2013.
- [12] Wikipedia, <<[https://en.wikipedia.org/wiki/Film\\_analysis](https://en.wikipedia.org/wiki/Film_analysis)>>, accessed at 4<sup>th</sup> sept 2022.
- [13] Kaggle, <<<https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews>>>, accessed at 3<sup>rd</sup> sept 2022.
- [14] IMDB, << [https://www.imdb.com/title/tt0499549/reviews?ref\\_=tt\\_urv](https://www.imdb.com/title/tt0499549/reviews?ref_=tt_urv)>>, accessed at 5<sup>th</sup> sept 2022.
- [15] Github, << [https://github.com/Minhaj1889/Review\\_Dataset\\_OF\\_Movies](https://github.com/Minhaj1889/Review_Dataset_OF_Movies)>>, accessed at 5<sup>th</sup> sept 2022.
- [16] Github, <<[https://github.com/Minhaj1889/preprocess\\_sprinter\\_dir](https://github.com/Minhaj1889/preprocess_sprinter_dir)>>, accessed at 5<sup>th</sup> sept 2022.
- [17] Wikipedia, << [https://en.wikipedia.org/wiki/Sentiment\\_analysis](https://en.wikipedia.org/wiki/Sentiment_analysis)>>, accessed at 3<sup>rd</sup> sept 2022.
- [18] Wikipedia, <[https://en.wikipedia.org/wiki/Crossvalidation\\_%28statistics%29](https://en.wikipedia.org/wiki/Crossvalidation_%28statistics%29)>> accessed at 3<sup>rd</sup> sept 2022.

# Test

---

## ORIGINALITY REPORT

---

**26%**

SIMILARITY INDEX

**22%**

INTERNET SOURCES

**8%**

PUBLICATIONS

**18%**

STUDENT PAPERS

---

## PRIMARY SOURCES

---

<b>1</b>	<b>Submitted to Daffodil International University</b> Student Paper	<b>7%</b>
<b>2</b>	<b><a href="https://dspace.daffodilvarsity.edu.bd:8080">dspace.daffodilvarsity.edu.bd:8080</a></b> Internet Source	<b>5%</b>
<b>3</b>	<b><a href="http://www.ijcse.com">www.ijcse.com</a></b> Internet Source	<b>2%</b>
<b>4</b>	<b><a href="http://cseweb.ucsd.edu">cseweb.ucsd.edu</a></b> Internet Source	<b>2%</b>
<b>5</b>	<b><a href="http://www.ijera.com">www.ijera.com</a></b> Internet Source	<b>1%</b>
<b>6</b>	<b>Spiros Mancoridis. "A genetic algorithm for solving the binning problem in networked applications detection", 2007 IEEE Congress on Evolutionary Computation, 09/2007</b> Publication	<b>1%</b>
<b>7</b>	<b>Submitted to Michigan Technological University</b> Student Paper	<b>1%</b>
<b>8</b>	<b><a href="http://www.ripublication.com">www.ripublication.com</a></b> Internet Source	<b>1%</b>