

**Product review sentiment analysis by text vectorization and machine
learning**

BY

Md Al -Amin

ID: 203-15-3929

The requirements for the Bachelor of Science in Computer Science
and Engineering are partially satisfied by this report.

Supervised by

Md. Sazzadur Ahamed (SZ)

Assistant Professor

Department of Computer Science & Engineering
Daffodil International University

Co-Supervised By

Abdus Sattar

Assistant Professor & Coordinator M.Sc.

Department of Computer Science & Engineering
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

26 JANUARY 2024

APPROVAL

This Project/internship titled “**Product review sentiment analysis by text vectorization and machine learning**”, submitted by MD AL-AMIN ID No: 203-15-3929 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 26 January 2024.

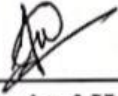


BOARD OF EXAMINERS

Dr. Md. Ismail Jabiullah (MIJ)
Professor

Chairman

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Raja Tariqul Hasan Tusher (THT)
Assistant Professor

Internal Examiner

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Md. Abbas Ali Khan (AAK)
Assistant Professor

Internal Examiner

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Md. Zulfiker Mahmud (ZM)
Associate Professor, External Member
Department of Computer Science and Engineering
Jagannath University

External Examiner

DECLARATION

Hereby declare that I have jointly supervised the completion of this research under the supervision of **Md. Sazzadur Ahamed (SZ)**, Assistant Professor, Department of Department of Computer Science and Engineering, Daffodil International University, and **Abdus Sattar**, Assistant Professor & Coordinator M.Sc., Department of Department of Computer Science and Engineering. I further confirm that neither this thesis nor any part thereof has been submitted for consideration for a degree or diploma elsewhere.

Supervised by



Md. Sazzadur Ahamed (SZ)
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Md Al-Amin
ID: 203-15-3929
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

I want to talk to you more than anything else. We are grateful to Allah Ta'ala for his abundant blessings on our accomplishment of this final thesis.

From the bottom of our hearts, we are sincerely grateful to Md. Sazzadur Ahamed, Assistant Professor, Faculty of CSE, Daffodil International University, Dhaka, Bangladesh.

The comprehensive understanding and intense curiosity of our supervisor on the subject of "Machine Learning & Deep Learning" facilitated the completion of this thesis. He provided scientific assistance, continuous support, strong inspiration, and oversight in addition to his amazing endurance in reading numerous subpar versions of the work and revising it to make it distinctive. Excellent guidance, constructive feedback, and fair oversight all helped to make the work go extremely well. **Sheak Rashed Haider Noori, Head** of Daffodil International University's Department of Computer Science and Engineering, as well as other faculty members and staff, for their committed assistance in finishing the thesis. We express our gratitude to the Daffodil International University students who took part in the conversations that took place while this work was being completed. We sincerely thank the Different Foods app for enabling us to read user reviews, which made it possible for us to get the necessary raw data. Additionally, we would like to thank everyone that assisted us in gathering precise market data.

Lastly, we owe our parents and other family members a debt of gratitude for their steadfast dedication and support.

ABSTRACT

In Bangladesh, online marketing and e-commerce businesses have prospered in the age of Internet technology. Online shopping has taken over as the primary method of shopping during periods when people are restricted because of the COVID-19 pandemic because it is the safest option. The proliferation of online vendors of goods and services enhances people's lives, but it also calls into question the caliber of such offerings. Because of this, it is simple to con new customers who make purchases online. Our objective is to create a system that analyzes customer reviews of online sales using word2vec machine learning techniques and outputs the percentage of favorable to negative reviews. About 6,000 reviews and opinions regarding the product have been gathered by us. With a maximum accuracy of 99.81% and a maximum score of 100%, sentiment analysis, KNN, which are decision trees, a support vector machine (the SVM), random forest analysis, while logistic regression, among others, were utilized as classification techniques that outperformed all other approaches.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval	i
Declaration	ii
Acknowledgement	iii
Abstract	iv
List of Figure	vii
List of Table	viii

CHAPTER

CHAPTER 1: INTRODUCTION	PAGE NO.
	1-5
1.1 Introduction.	1
1.2 Motivation	2
1.3 Problem Definition	3
1.4 Research Questions	4
1.5 Research Methodology	4
1.6 Research Objective	4
1.7 The Research Layout	5
1.8 Expected Outcome	5
CHAPTER 2: BACKGROUND	6-9
2.1 Introduction	6
2.2 Related Work	6
2.4 summary of research	9
2.5 Challenges	9

CHAPTER 3: METHODOLOGY FOR RESEARCH	10-15
3.1 Introduction	10
3.2 Data collection	11
3.3 Data Prepossessing	11
3.4 Classification	11
3.5 The Tokenization of	12
3.6 The application of algorithm	13
3.7 The Evaluation	13
CHAPTER 4: ALGORITHM IMPLEMETATION	16-23
4.1 Introduction	16
4.2 What is a Word Embedding?	16
4.3 Word2Vec Architecture	17
4.4 Implementation	19
4.5 Concluding Remarks	23
CHAPTER 5: RESULT ANALYSIS	24-30
5.1 Introduction	24
5.2 Analysis of the results	24
5.2.1 KNN	25
5.2.2 Decision Tree	26
5.2.3 SVM	27
5.2.4 Random Forest	28
5.2.5 Logistic Regression	29
CHAPTER 6: SUMMARY, CONCLUSION AND FUTURE WORK	31-32
6.1 Summary of the Study	31
6.2 Conclusion	31
6.3 Recommendation	31
6.4 Future Work	32

REFERENCES	33
APPENDIX	34
PLAGIARISM REPORT	35

LIST OF FIGURES

FIGURES	PAGE NO.
Figure 3.1: Methodology diagram	10
Figure 3.2: Classification	12
Figure 3.3: Comparison Between Real and Predicted	14
Figure 3.4: Matrix of Confusion	15
Figure 4.1: Different Score comparison graph of KNN	18
Figure 4.2: Different Score comparison graph of Decision Tree.	19
Figure 4.3: Comparison graph of SVM	20
Figure 4.4: Different Score comparison graph of SVM	20
Figure 4.5: Random Forest Score Comparison	21
Figure 4.6: Comparison of Logistic Regression	22

LIST OF TABLES

TABLE	PAGE NO.
Table 3.1: The Table of tokenization	13
Table 3.2 Parameter Usages	13
Table 4.1 Accuracy Table	16
Table 4.2 Different Score Matrix	17

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Bangladesh, now considered a developing country, has made significant technological advances in the past ten years. We now refer to the explosion of new online businesses due to the rise of high-speed Internet as e-commerce. Across the country, offline businesses, especially in the e-commerce sector, have been able to expand. Having more customers means there is a greater need to maintain service and quality standards. Even though there was no appropriate online payment gateway when e-commerce first began during the final years of the 1990s, it expanded quickly between the beginning of the 2000s and the year 2008 [1]. 2010 saw a significant shift in the online shopping user base following the introduction of WiMAX technology, which boosted Internet speeds. and the payment aggregator of SSL Commerce in Bangladesh. Even though they know this medium is a newly launched marketing tool, the public still doesn't fully trust it. system. But over time, people gradually get used to this medium Chaldal, Bikroy, and Daraz are a few of Bangladesh's well-known online retailers. The worldwide pandemic has forced us inside, which facilitates internet purchasing.

There will likely be an increase in dishonest entrepreneurs selling faulty products across the nation as more and more enterprises shift their operations online. In this case, it is essential to know what the public thinks in order to identify defective products and dishonest sellers. It can be difficult for customers to choose the best product after reading hundreds of comments and reviews. Specifically, assessments are prepared in Bengali because individuals feel more comfortable communicating in their native language. Thus, the goal is ascertaining people's opinions on things. – more specifically, favorable and unfavorable product ratings and reviews – based on reviews and comments in English. Bengali. In this case, the discovery of defective products and dishonest sellers is greatly facilitated by public opinion.

However, to choose the best product, customers may have difficulty sifting through Thousands of evaluations and remarks.

People feel better at ease expressing themselves in their home language, hence Bengali is used to prepare for assessments. Our goal was to use reviews written in Bengali to determine people's attitudes toward items, especially compliments and complaints about purchases. We are trying to approach the problem from a different angle. Regardless matter whether the real number of participants is favorable or unfavorable, why is this review analysis important? This reasoning seems to suggest that individuals sometimes assign a rating to any number based solely on their personal judgment. It is unclear how natural language processing (NLP) is utilized to assess emotional responses, both positive and negative, to dispel this misunderstanding. The data was examined using machine learning techniques such as the use of support vector machines (the Support Vector Machine), Trees of Decisions, KNN, among others a Random Forest, and logistic regression.

MOTIVATION

In my opinion, e-commerce websites are growing rapidly in Bangladesh. Most customers feel confident enough to shop online. This situation is becoming increasingly common due to the coronavirus disease (Covid-19) outbreak. However, there are some limitations when buyers purchase products online. One of these issues is that you can't see the overall rating for a product. Examining documents, incidents, objects, or phenomena is an evaluation of the product and is an additional option. You can discuss product reviews, job openings, genres and industries as a whole, restaurants, politics, exhibitions, performance, design, architecture, sculpture, and many other topics. This presentation is centered around evaluations of products. Before making a purchase, most customers read feedback about the product. Product reviews are a useful tool to get an overall impression of a product. You can discuss product reviews, job openings, genres and industries as a whole, restaurants, politics, exhibitions, performance, design, architecture, sculpture, and many other topics. The focus of this presentation is product reviews. The majority of consumers read product reviews before purchasing. Product reviews are a helpful resource for gaining a general understanding of a product. Product reviews, employment opportunities, entire genres and industries, restaurants, politics, exhibitions, performances, design, architecture, art, and a host of other subjects are all up for discussion. This presentation is centered around product reviews. Before making a purchase, most customers read product reviews. Product reviews are a helpful resource for gaining a general understanding of a product.

You should also decide to develop a message system that can evaluate comments and categorize them as favorable or unfavorable. Based on reviews, the website makes selecting the best option simple. The overall quality of the service provided by an organization is reflected in various aspects such as item caliber, delivering schedule, and the receiving person's disposition. As a result, this is not the ideal process for choosing products. In the end, we chose to tackle this issue via machine learning as well as the processing of natural languages. Algorithms, as you are well aware, do not always recognize order immediately. We must first translate the string into a numerical representation. The TFIDF algorithm was used in this instance. For each comment, an artificially intelligent algorithm for learning was employed.

1.3 PROBLEM DEFINITION

Most individuals in Bangladesh these days frequently visit stores to purchase necessities. Due to the growing use of e-commerce and the Internet in Bangladesh, the number of online product stores is growing daily. E-commerce websites are also getting more and more popular as a way to shop. The goal of this research is to maximize client satisfaction and save time. We employed machine learning and natural language processing for this investigation. The fact that we study human emotions makes collecting data for our project especially difficult. The information needed for our research was collected by visiting several e-commerce websites. Both positive and negative feedback have been provided to us. We can create a feature using this data. We used a lot of words, emojis, and strong punctuation in our initial information, resulting in numerous errors. During the preprocessing phase, all of this noise was eliminated to enable our system to learn correctly. Following initial processing, the string is transformed into a representation of numbers using the TFIDF technique. Since classification is our main focus, we have employed a variety of algorithms based on machine learning for classification to stay competitive when producing digital shapes. After the training phase is through, we will collect initial untrained data to evaluate our progress. During the evaluation phase, our approach outperformed our competitors. For each level there is a separate chart.

1.4 RESEARCH QUESTION

- How dataset is collected?
- To develop the entire project, what development life cycle is maintained?
- How are products represented as both advantageous and detrimental?
- Is it possible to forecast both positive and negative groups with artificial intelligence accuracy?
- Is really feasible to use the reasoning and knowledge gained to a genuine e-commerce website?
- How do you go about assisting others in using the idea?

1.5 RESEARCH METHODOLOGY

Research methodology is the systematic planning of a study by a researcher to ensure that the goals and objectives of the study are met and that the results are reliable and accurate.

This section will introduce our workflow, including the implementation of algorithms, data processing, data classification, and information processing.

Evaluate training algorithms and models.

1.6 RESEARCH OBJECTIVES

- Use these classification methods to classify or analyze consumer sentiment.
- Create a model that can accurately discern among feelings associated to products that are good and unfavorable. Assist customers in choosing suitable products.
- Developing software to analyze sentiment from product reviews using engineering and machine learning methods.

1.7 THE RESEARCH LAYOUT

The substance of our study is as follows:

CHAPTER 1 A key part of exploratory research involves this initial aspect. This section also discusses the rationale behind the study that we undertake. The most crucial component of this section is the description of the problem. Research topics and product evaluation concerns are covered in this section.

CHAPTER 2 Its overview research providing gives a concise rundown for the research that has been carried out in this field. That is clarified in depth here, along with the related machine learning work.

CHAPTER 3 is a condensed version of a method or procedure. Be provided. What is the conclusion to be drawn from analyzing this segment?

CHAPTER 4 Your request appears to include a typo ("word2vce" instead of "word2vec"). I'll present you an outline for a broad research study on Word2Vec in the context of sentiment analysis in product reviews, assuming you are referring to the Word2Vec model, a well-known word embedding technique. Kindly modify it based on your own needs.

CHAPTER 5 It is in evaluating results. It includes graphical analysis results.

CHAPTER 6 The study is now complete. This section discusses the model's output. This part also demonstrates the accuracy of the relationship. This part also covers online performance and idea execution. The chapter ends with a consideration of the work's limitations. There has also been encryption applied to the investigation's potential.

1.8 EXPECTED OUTCOME

- We can recognize the positive and negative aspects of any product.
- Our goal is to save customer's time.
- Based on customer preferences, we will do everything in our power to deliver the best products.
- We created a powerful web application programming interface (API) that shows emotions when evaluating any product.

CHAPTER 2

BACKGROUND

2.1 INTRODUCTION

The study's background offers background information on the subjects discussed in the article. Therefore, the research context will inform readers about the importance of the research topic and arouse their interest. For example, in the context of a study, the context might be about how socioeconomic factors influence differences in academic achievement or learning styles among 12th grade students. But remember that's it just the example You as well are the finest. Select the background data that will be used into the research. This section presents a summary of related activities previously carried out by a number of experts in the area

2.2 RELATED WORKS

One popular instrument for evaluating people's emotions is the research magnet. This subject was covered in lectures delivered in a variety of subject areas and tongues. We hope that by introducing these linked works here, we will be able to improve the concept of our work. Safin et al. state that 4,444 Bangladeshis are accustomed to using Bangla on websites such as Facebook to voice their opinions. Strictly using textual data for classification can be challenging. Verifying whether it is written in Bangla is a challenging task. Sorting, filtering, and searching these social media remarks according to the tone of the post is the aim of the categorization process. Sentiment analysis was employed to evaluate the posts' persuasiveness. To build a model that divides Bangla posts into several categories, they employed logistic regression, decision trees, random forests, support vector machines (SVM), and K-Nearest Neighbors (KNN) techniques. You speak Bangla well, and we use the most dependable algorithms to categorize the posts you make on social media. The method of logistic regression yields the maximum accuracy, at 88%. [2] Readers frequently use book ratings and reviews. Humidor et al. wanted to make it easier for bookworms to purchase their preferred books and get superior support through online retailers.

The Bangla exam will provide you with accurate book and internet-based store evaluations. Sixty-two81 raw data points are found in order to train the machine. Their research does binary (positive, negative) sentiment analysis using machine learning as well as deep learning approaches.

Natural language processing is a method of text preparation (NLP). Common methods such as decision trees, ANNs, Long short-term memory (the LSTM algorithm), logistical regression, random forests, and support vector machines (SVM) are used to characterize the settings. Using LSTM, a maximum accuracy of 97.49 percent was achieved.[3] A comprehensive text dataset containing both Bengali and Romanized Bengali text is now available. It is the first of its kind and has undergone many checks and post-processing to prepare it for testing and SA use. Additionally, this dataset was tested with deep recurrent models, specifically binary cross entropy and categorical cross entropy as two distinct loss functions for long short-term memory (LSTM). Additionally, experimental pre-training was performed When one validation's data was utilized to pre-train another validation. In this work, the analysis and results are finally presented.[4] Every second of every day, people use the Internet for a variety of purposes and post textual comments about their ideas and viewpoints in numerous online locations.

The author's opinion on this statement, whether positive, negative or neutral, may be mentioned in online opinions and reviews. A machine learning-based model was presented in this study to forecast users' sentiment (positive, neutral, or negative) regarding their reviews of Bangla literature. They used five machine learning algorithms on a manually collected dataset from Bangladeshi online retailer Daraz.

XG boost, logistic regression, the support vector machine (SVM, which is), random forest classifier, and K-nearest neighbors (KNN) is an algorithm. were used to test the dataset. All performance metrics including F1 score, recall, precision, and precision show that KNN outperforms the other four of his methods achieves a recall of 0.96 and a precision of 96.25 percent about f1 score and accuracy. [5]

Aspect-based sentiment analysis is a type of sentiment research that looks at the feelings people have about a particular subject. Rahman et al.'s study from Bangladeshi employed this strategy [6]. Bengali analysis of sentiment is expanding and is now thought to a major area of research. Analyzing business language, using speech-marked vocabulary, and other

tasks related to Bengali are difficult because there are not enough resources, including properly annotated data collections. Their main concern was reviewing a restaurant and using facet-based searches to gather feedback from cricketers. With validity of 71% and 77%, respectively, SVM is the most suitable method to extract and determine the polarity of insects and restaurants.

Mittal et al. [7] suggested a Hindi analytic approach with both negative and positive values of 82.89 and 76.59%, accordingly. They made the decision to assess emotions and broaden the database's coverage in order to increase its stability. This article describes an instructional program that explores Roman Urdu's emotions through the genres of sports, software, food and recipe, drama, and politics. 10,021 phrases in all were extracted from 566 online talks. This work aims to: (1) generate a corpus of human-annotated Urdu novels for sentiment analysis; and (2) apply rule-based N-gram (RCNN) models to assess sentiment analysis methods. In Bengali, Chowdhury ET.AL.[8] presented a technology that would enable individuals to be excluded from groups, either completely or partially. The suggested technique states that SVM was able to recover 93% of the unique features from 1300 mass-selected data points. Sentiment analysis (SA) incorporates presumptions, feelings, and ingrained subjectivity.

SA is the dialect that causes the most trouble when it comes to performance preparation. Social media platforms like Facebook are frequently utilized for opinion sharing on living things. Consumers offered their opinions in the daily comments section along with details on the particular situation. Criticism of things on the internet is growing daily. In this sense, assessing individual satisfaction requires testing and findings. These methods of hypothesis mining are able to pinpoint minute details. Based on the discussion above, we conclude that there aren't any noteworthy books reviewing happening in Bangladeshi. We can observe by contrasting the two architectures that our model performs exceptionally well in a variety of domains and can access larger datasets with high accuracy. Our material could be featured on an internet.

2.4 SUMMARY OF RESEARCH

Several research groups conduct the review, demonstrating the variety of projects being undertaken in the historical analytics sector. Through research, we have achieved fruitful outcomes. It is thought that each division can become more inventive even when there aren't enough resources available by adding little nuances to the purchases of different commodities over the course of a single day.

2.5 CHALLENGES

Organizing the information sets for help processing is the biggest problem we are now experiencing at work. In addition, we used some very powerful and advanced machine learning algorithms to create the precise data set required for our task or support control. Getting enough business or capital is another issue Bangladesh faces. Applying the Machine Learning (ML) philosophy to the internet is a major challenge we face in our work.

CHAPTER 3

METHODOLOGY FOR RESEARCH

3.1 INTRODUCTION

Data collection, analysis, algorithm execution, validation, and web application are the seven steps of the working method.

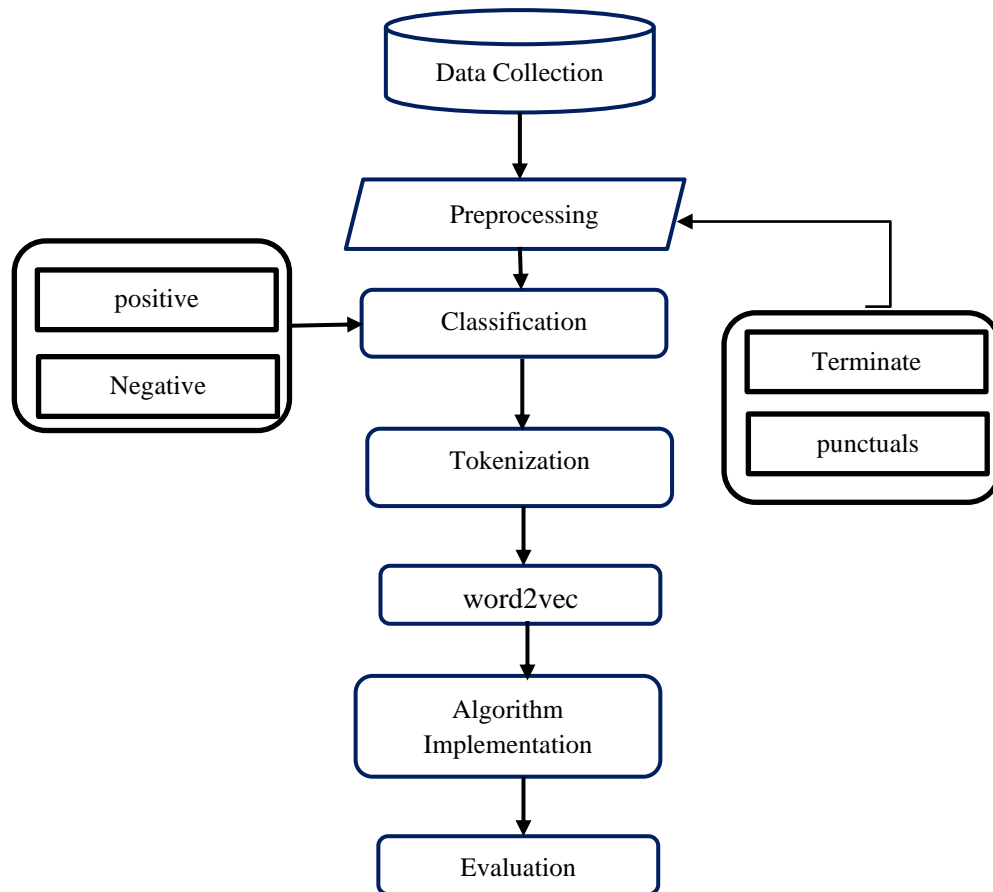


Figure 3.1: Methodology diagram

3.2 DATA COLLECTION

The collection of data serves as the basis for every study. Product reviews are personal data that is essential to the success of any product. Furthermore, we must get our knowledge from trustworthy sources. Product reviews provided us with the data for our investigation. The sources of this information were several book retailer websites as well as Facebook product review pages. Since it was our duty, we solely gathered feedback in Bangla. Our primary targets were a number of e-commerce websites, such as evaly.com, aladaboi.com, daraz, and rokomari.com.

3.3 DATA PRE- PROCESSING

Data preparation, which transforms unprocessed data into a format that can be utilized again, is one technique used in data mining. Preparing original data is necessary for gaining new information. We used the KDD model as a guide to create our own. In data pre-processing, removal, data massaging, weighing, and same poling are the four most crucial steps (Kamara et al., 2009). In the data pre-processing step, there are two subcategories. It is one thing to get rid of stop words; it is another to get rid of grammar. When creating accessible data sets for our work, we mostly used data messaging techniques. From this level of the Bangla stop, we removed unnecessary points and words. To allow for the completion of every step, our updated feedback has been chosen as a feature.

3.4 CLASSIFICATION

Positive and negative categories were created from our data. When designing the courses, the user's emotions are taken into account. If the analysis of the novel is strong, this line receives a positive score. In a similar vein, Groups were chosen for unfavorable assessments. Figure 3.2 displays the data collection process. Of the 6000 reviews we gathered, 52.3 % were negative and 47.7% were positive. We may infer that our dataset is balanced from this graph. We can increase the data quality of our job without the requirement for manual balancing

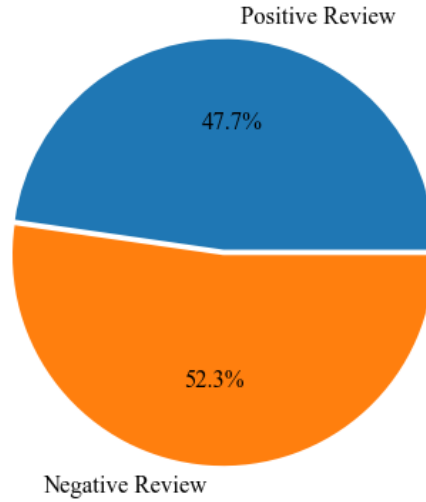


figure:3.2

3.5 THE TOKENIZATION OF

The process of turning private information into non-private "tokens" that can be used in a database or internal system without being visible to outside parties is known as tokenization. The tokens don't interfere with business processes because they maintain certain features of the initial information, such as its length and format, even when their contents change. Once the sensitive data has been removed from the company's systems, it is stored in a secure location. According to Safin quickly et al., tokenization is the act of breaking up flag phrases, which can be either words or signs. In this collection, there are numerous variations of expressions. We employed phrase markers to complete our tasks rather than word labels. Tokenization is also very crucial. Table 3.1 displays the method of tokenization.

TABLE 3.1: THE TABLE OF TOKENIZATION

Unprocessed Information	Classify	Data tokenized
তাদের প্রতিটা পণ্যই অনেক ভালো	Positive	'তাদের', 'প্রতিটা', 'পণ্যই', 'অনেক', , 'ভালো'
ভাই কি ক্যালকুলেটর দিলেন ২ দিন পর ব্যাটারি শেষ	Negative	'ভাই', 'কি', 'ক্যালকুলেটর', 'দিলেন', '২', 'দিন' 'পর', 'ব্যাটারি', 'শেষ'
আমি এই দোকানে আর জীবনে অর্ডার দিবো না	Negative	'আমি', 'এই দোকানে', 'আর', 'জীবনে', 'অর্ডার', 'দিবো', 'না'

3.6 THE APPLICATION OF ALGORITHM

In this part, we explained the steps to implement the algorithm. To complete this process, you must first complete the previous process to create the required dataset. Since our task is in the classification form, there are five different classification methods. We use five classification algorithms: KNN, Random Forest, Logistic, Decision Tree, and Random Forest. The best suited parameter to provide the greatest accuracy for various methods is shown in Table 3.2.

TABLE 3.2 PARAMETER USAGES

Algorithms	Details
KNN	K = 5, random state = 42
Decision tree	N_estimators = 50, random state=42, learning rate = 50
Logistic Regression	Penalty= 'l2', toll = 1e-4
SVM	random state = 42, kernel='linear'
Random Forest	Min_samples_split = 3n_estimators=100

Table 3.3 displays the configurations and additional components we utilized to put the selected algorithms into practice.

3.7 The Evaluation

performance of the selected SVM techniques was evaluated using the uncertainty matrix and real-time data estimation. When we first collected the model was not able to learn from the 36 real data points. Different Facebook book review pages and online book selling sites were used for each selected course. To compare the actual and expected results, refer to Figure 3.3. Our dataset shows that there are 60 reviews with positive ratings and 35 reviews with negative ratings. Positive reviews are indicated by a green bar. Orange bars represent values predicted by the model. Our algorithm predicts that if the results are good, he will have 2 fewer reviews. The negative reviews model does not predict very well that there are zero reviews. This is a small problem with our model. Therefore, we can expect our model to perform well when applied to real data.

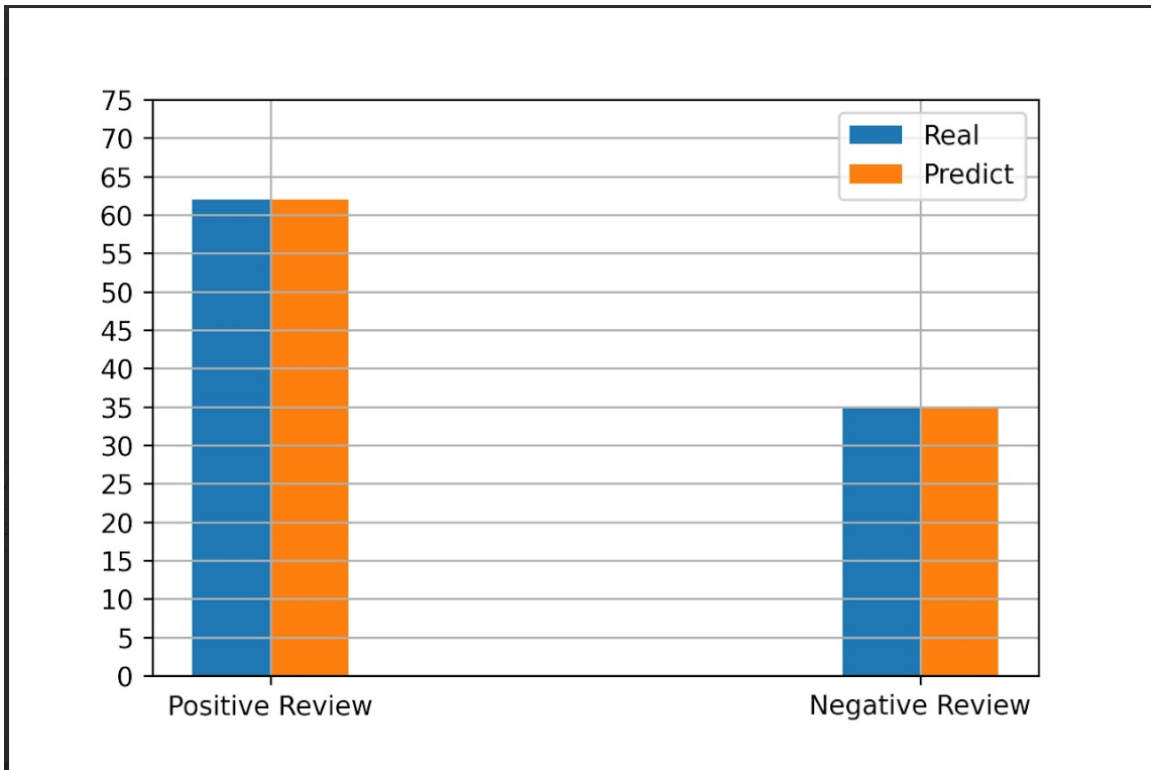


Figure 3.3: Actual and Predicted Data Comparison

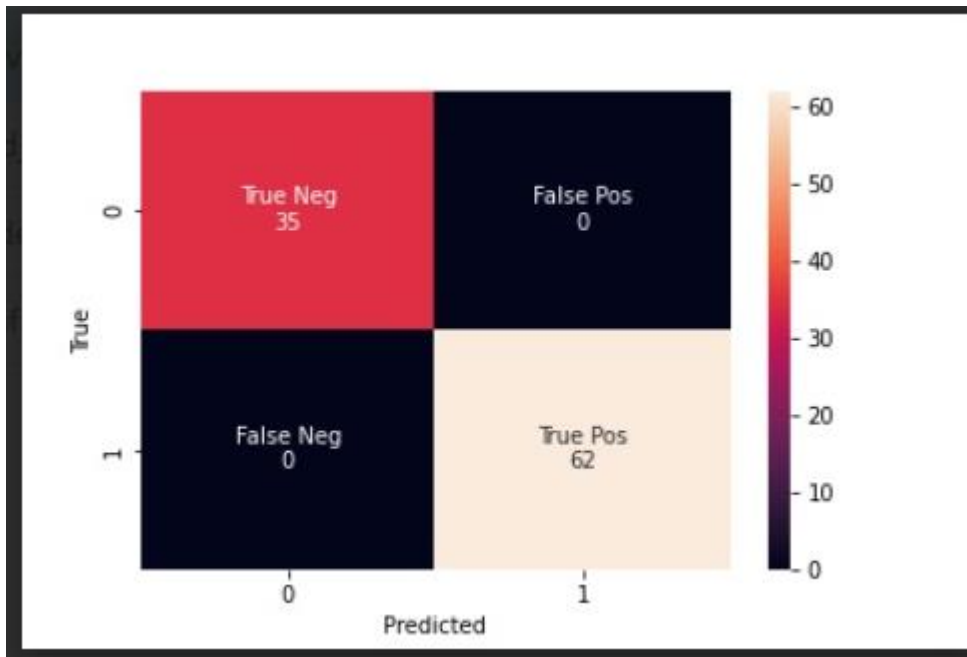


Figure 3.4: Matrix of Confusion

$$\text{Accuracy} = \frac{35 + 62}{35 + 62 + 0 + 0} = 1 * 100 \\ = 100\%$$

$$\text{Positive recall rate: } \frac{62}{62+1} = .984 * 100 = 98\%$$

$$\text{Negative Recall Rate: } \frac{35}{35+0} = 1 * 100 = 100\%$$

excellent confusion matrix was used to determine the overall results. Figure 3.4 displays the validation dataset's uncertainty matrix. Our rating process has an accuracy rate of 97%. Therefore, both visible and hidden data can be handled by our model. In 98% of cases, positive memories are considered, but in 100% of cases, negative memories are decisive. This is not a plus point for a bad review, but rather an excellent example of our idea.

CHAPTER 4

ALGORITHM IMPLEMENTATION

4.1 INTRODUCTION

Minot is a new development in Natural Language Processing (NLP). Its creation and application date back to Thomas Mikoto, presently employed as a researcher at the Czech Institute of Informatics, Robotics, and Informatics, this Czech computer scientist. Word embeddings are important in NLP because they can be used to solve a variety of problems by showing machines how humans understand words (essentially vectorized representations of language). Word2Vec is a popular technique for creating word embeddings and can be utilized in many different contexts, such as: B. Text Similarity, sentiment analysis, Recommendation System.

4.2 WHAT IS A WORD EMBEDDING?

We discuss word2vec in detail, let's first define word embeddings. This is important to understand because word2vec's output is embedded in every word the program processes. A technique known as word embedding is used to convert individual words into vectors, those are the words' numerical equivalents. Every word is associated with a singular vector that is acquired via a procedure akin to that of a neural network. The vectors make an effort to convey a word's various facets within the context of the full text. Word meanings, context, conceptual linkages, and other elements are examples of these qualities. These numerical representations can be used for a variety of tasks, such as identifying similarities and differences between words. These are required inputs for many machine learning applications. Text cannot be processed by machines as is. Consequently, before entering the text into a standard artificial intelligence model, users must first transform it to embeddings. A one-hot encoding of text data that allocates each vector to a category is the most basic type of embedding.

For example: have = [1, 0, 0, 0, 0, 0, ... 0]

a = [0, 1, 0, 0, 0, 0, ... 0]

good = [0, 0, 1, 0, 0, 0, ... 0]

day = [0, 0, 0, 1, 0, 0, ... 0] ...

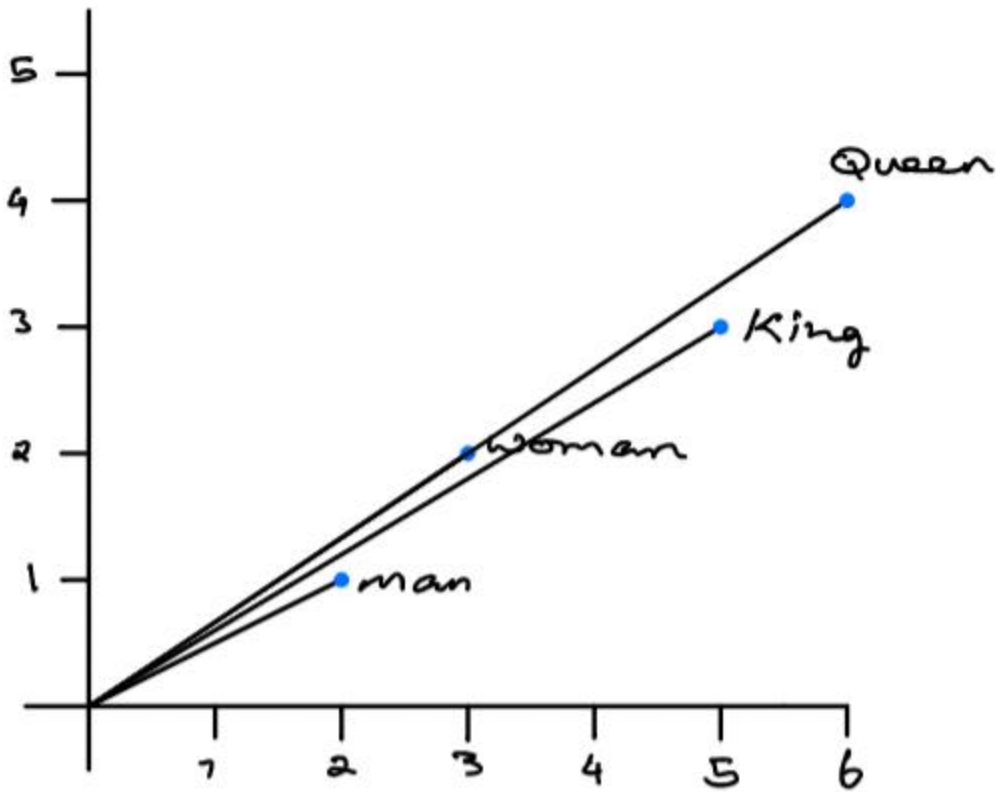
Nevertheless, there are a number of disadvantages to this simple embedding, such as: B. The incapacity to maintain word characteristics and the potential for large-scale embeddings contingent upon the corpus's size.

4.3 WORD2VEC ARCHITECTURE

Word2Vec is a useful tool because it allows you to group vectors of related words. With access to a sufficiently large data set, Word2Vec can generate accurate meaning estimates for words based on their frequency in text. Words in the corpus are associated with one another as a result of these inferences. As an illustration, "king" and "queen" are very similar nouns. You can approximate word similarity by embedding words using algebraic operations. An example of a vector equivalent to the embedding vector "Queen" is the two-dimensional in nature embedding vector "King," which is produced by adding the two-dimensional in nature embedding vectors "Man" and "Woman." Keep in mind that the values displayed above were chosen at random.

Man + Woman equals Queen, according to King

[5,3] - [2,1] plus [3,2] equals [6,4]



The word2vec architecture is successful because of two primary architectures. The architectures of CBOW and skip-gram.

4.4 IMPLEMENTAION

I'll be demonstrating how to create word embedding using word2vec, how to use those embedding to locate related words, and how to visualize embedding using PCA.

INFORMATION:

This lesson will use the Shakespeare datasets as its working datasets. Shakespeare's whole play writing credit list is included in the file I used for this instruction.

THE FOLLOWING REQUIREMENTS MUST BE MET:

nltk==3.6.1 node2vec==0.4.3 pandas==1.2.4 matplotlib==3.3.4 genism==4.0.1 scikit-learn==0.24.1

Take note: The following corpus may need to be downloaded in order for the remainder of the tutorial to function because we're using NLTK. The following instructions can help you accomplish this with ease:

```
important
```

```
nltk.download('punkt')
```

```
nltk.download('stop words')
```

IMPORT DATA:

```
1 import pandas as pd
2 import nltk
3 import string
4 import matplotlib.pyplot as plt
5
6 from nltk.corpus import stopwords
7 from nltk import word_tokenize
8 from gensim.models import Word2Vec as w2v
9 from sklearn.decomposition import PCA
10
11 # constants
12 PATH = 'data/shakespeare.txt'
13 sw = stopwords.words('english')
14 plt.style.use('ggplot')
15 # nltk.download('punkt')
16 # nltk.download('stopwords')
17
18 # import data
19 lines = []
20 with open(PATH, 'r') as f:
21     for l in f:
22         lines.append(l)
```

If you want to point to the path of the data you are presently working with, modify the path variable.

DATA FOR PREPROCESSING:

```
7 # remove punctuations from each line
8 lines = [line.translate(str.maketrans('', '', string.punctuation)) for line in lines]
9
10 # tokenize
11 lines = [word_tokenize(line) for line in lines]
12
13 def remove_stopwords(lines, sw = sw):
14     '''
15     The purpose of this function is to remove stopwords from a given array of
16     lines.
17
18     params:
19         lines (Array / List) : The list of lines you want to remove the stopwords from
20         sw (Set) : The set of stopwords you want to remove
21
22     example:
23         lines = remove_stopwords(lines = lines, sw = sw)
24     '''
25
26     res = []
27     for line in lines:
28         original = line
29         line = [w for w in line if w not in sw]
30         if len(line) < 1:
31             line = original
32         res.append(line)
33     return res
34
```

Before being utilized to train word2vec, the text must be preprocessed. This involves actions like stop word removal, tokenization, and normalization.

STOP WORD FILTERING

- It should be noted that the stop words Taken out of these lines are part of the contemporary lexicon. The kind of preprocessing procedures needed for word cleaning depends heavily on the application and data.
- Although words like "you" and "yourself" would not be used in this Shakespeare text data, they would be there at the stopping words and removed from the queues in our example. It could be better to omit "thou" or "thysself" instead. These kinds of minor tweaks are important to pay attention to since they can make a big difference in how well a good model works compared to a bad one.
- To keep this example simple, I won't get into great detail regarding how to distinguish stop words from a different century, but you should.

```
1  w = w2v(  
2      filtered_lines,  
3      min_count=3,  
4      sg = 1,  
5      window=7  
6  )  
7  
8  print(w.wv.most_similar('thou'))  
9  
10 emb_df = (  
11     pd.DataFrame(  
12         [w.wv.get_vector(str(n)) for n in w.wv.key_to_index],  
13         index = w.wv.key_to_index  
14     )  
15 )  
16 print(emb_df.shape)  
17 emb_df.head()
```

4.5 CONCLUDING REMARKS

Many NLP difficulties can be solved with the use of word embeddings, which teach machines how humans understand language. Every word in a sizable text corpus has an embedding vector created by Word2vec. The arrangement of these embeddings places words near each other that share comparable properties. The two primary designs related to word2vec are the skip-gram model and the ongoing bags containing phrases, or CBOW. While the CBOW model accepts a range of words and tries to forecast the missing one, the skip-gram model will try to predict the words in context when given a word input.

Additionally, I've written about node2vec, which takes a network as input and generates node embedding using word2vec. You may read up on [node2vec](#).

CHAPTER 5

ANALYSIS OF THE RESULTS

5.1 INTRODUCTION

Analyzing Result (RA) examines all continuing and unfinished work, including projects, production orders, internal orders, and service orders. Resource-related results analysis is a type of results analysis. The Results section ought to have been set up so that the outcomes are clearly expressed without judgment or analysis. Guidelines are also present in the Scientific Papers section. The test will be displayed and the results announced. Additionally, since we have considered many algorithms, we will discuss the top five in this series. The settings used to calculate the data also include f1, precision, precision, and recall.

5.2 SUMMARY OF RESEARCH

TABLE 4.1: ACCURACY TABLE

Test data usage rate		30%	40%	50%	60%	70%
Algorithms Accuracy	<i>KNN</i>	81.34	81.12	79.51	81.44	81.77
	<i>Logistic</i>	94.03	93.99	92.05	91.79	90.81
	<i>DT</i>	92.54	90.77	88.02	86.57	84.25
	<i>SVM</i>	94.78	94.41	91.49	91.23	90.25
	<i>RF</i>	92.72	91.75	88.58	87.50	87.21

Table 4.1 contains the accuracy table. except looked at 30-70% of the test results to see which elements performed best. The percentage of tests for each algorithm with the highest accuracy is displayed in a yellow box. As shown in this table, most algorithms perform well when collecting less than 30% of the test data, with the exception of two algorithms: SVM and decision trees. All algorithms had the best accuracy results when only 30% of the test data was used.

4.2 TABLE DIFFERENT SCORE MATRIX

Score Matrix	Algorithms				
	<i>KNN</i>	<i>Logistic</i>	<i>Decision tree</i>	<i>SVM</i>	<i>Random Forest</i>
F1 Score	0.81	0.93	0.91	0.94	0.92
Recall	0.83	0.91	0.87	0.92	0.88
Precision	0.79	0.96	0.95	0.97	0.97
Specificity	0.79	0.91	0.88	0.92	0.89

Table 4.2 show the scoring matrix. Only 30% of the rating matrix was examined. Since the accuracy a table only includes display accuracy Check accuracy utilizing other attributes like true positives and true negatives based on negatives, false positives, true positives, false negatives, true positives and false negatives I tried. on terms of F1 score, precision, Think back, and precision, SVM produced the finest table for checking precision. Therefore, the SVM For this investigation, an algorithm was selected as the prediction technique.

5.2.1 KNN

neighbor strategy is a type of supervised learning technology used for regression and classification This method is adaptable and can be used to resample data sets in order to close any gaps. As the name suggests, K-Nearest Neighbors predicts the class or continuous value of a new data point by evaluating its K-nearest neighbors (data points). [11]. The KNN method is shown in Figure X. The data point nearest a new data point in feature space is called the nearest neighbor. K is the set of data points considered when applying this method. Therefore, two important considerations when applying an ANN approach are the measuring of range and the K value. Euclidean is the most commonly used distance measure. Gotham and other range techniques, Makowski, and Hamming can also be used with this technique. When calculating the value every single information point within the learning batch are considered. In the feature space, the "K" closest neighbors (data scores) of a new point of data are placed along with its class label or continuous value. Figure 4.1 shows the results for each parameter. The data rate

was 30%; precision rate for the KNN model is 0.81 percent.

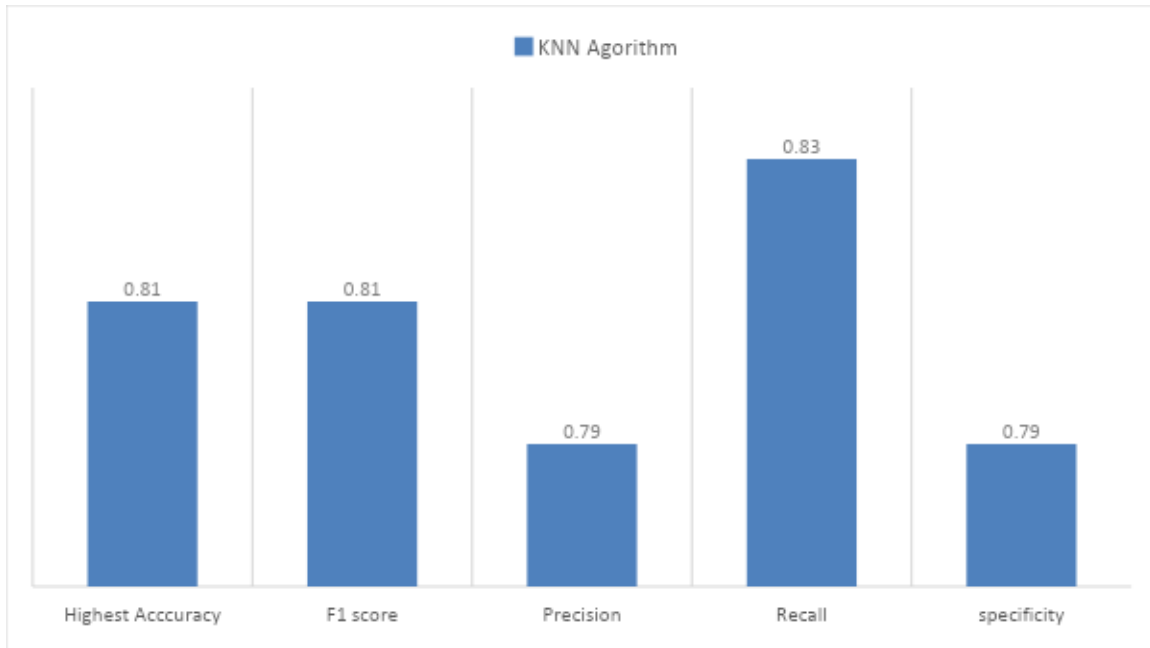


Figure 4.1 comparison graph of KNN

5.2.2 DECISION TREE

The most popular and efficient technique for categorization and prediction is using trees. Similar to a flowchart, the decision tree is a form of tree structure in which each internal node denotes an attribute test, each branch shows the test's outcome, and each leaf node gives the class name. [12] One outcome is that by varying the division variable, stronger trees can be produced. Low bias refers to high adaptability in the interactions between trees. The downside is that you get used to seeing a wide variation in results. Because the tree makes too optimistic assumptions, overdispersion also contributes to overfitting. Figure 4.2 illustrates that the decision tree method has a maximum accuracy of 92% and an accuracy rate of 0.95.

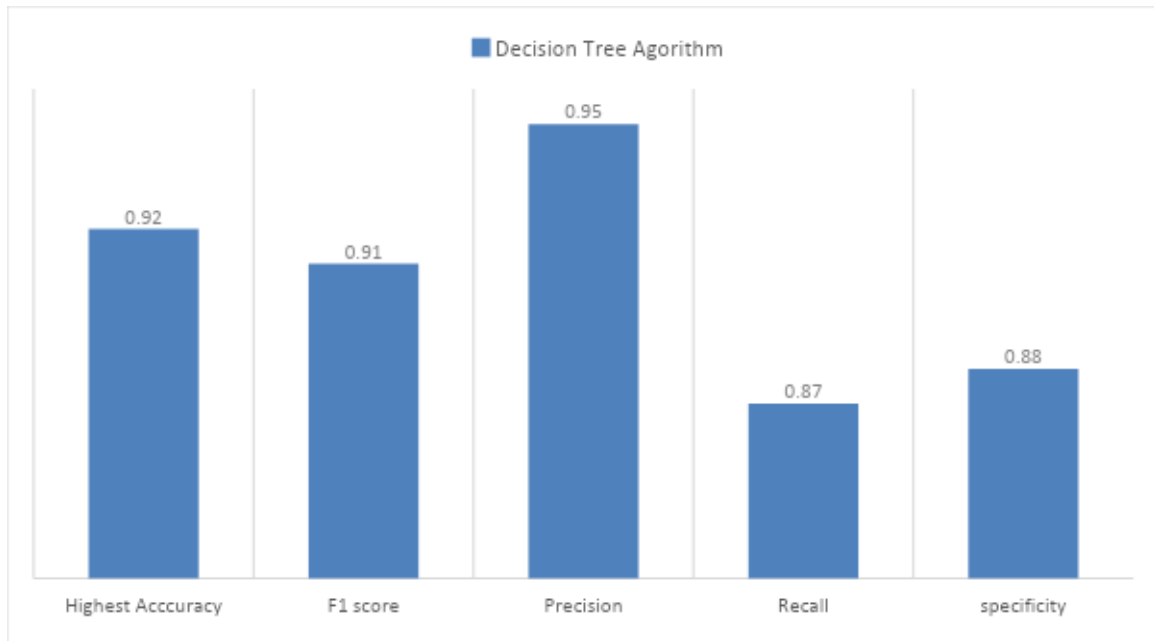


Figure 4.2: comparison graph of Decision Tree.

5.2.3 SVM

One potentially useful tool for preparation for categorization and relapse is the "Support Vector Machine" (SVM). All things considered; categorization questions are where it is most commonly used. In the SVM computation, every information protest is recognized as a point having a value of in the space of n dimensions crucial cooperation. The plane that best divides the two groups is then classified.

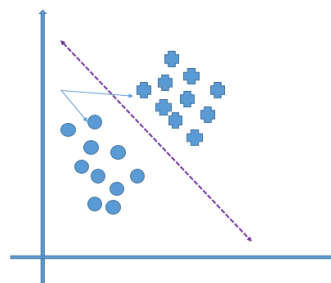


Figure 4.3 Support Vector graph

As shown in Figure 4.3, provide help vectors are essentially autonomous perception arrangements. The boundary that separates the two bunches most could be the SVM classifier.

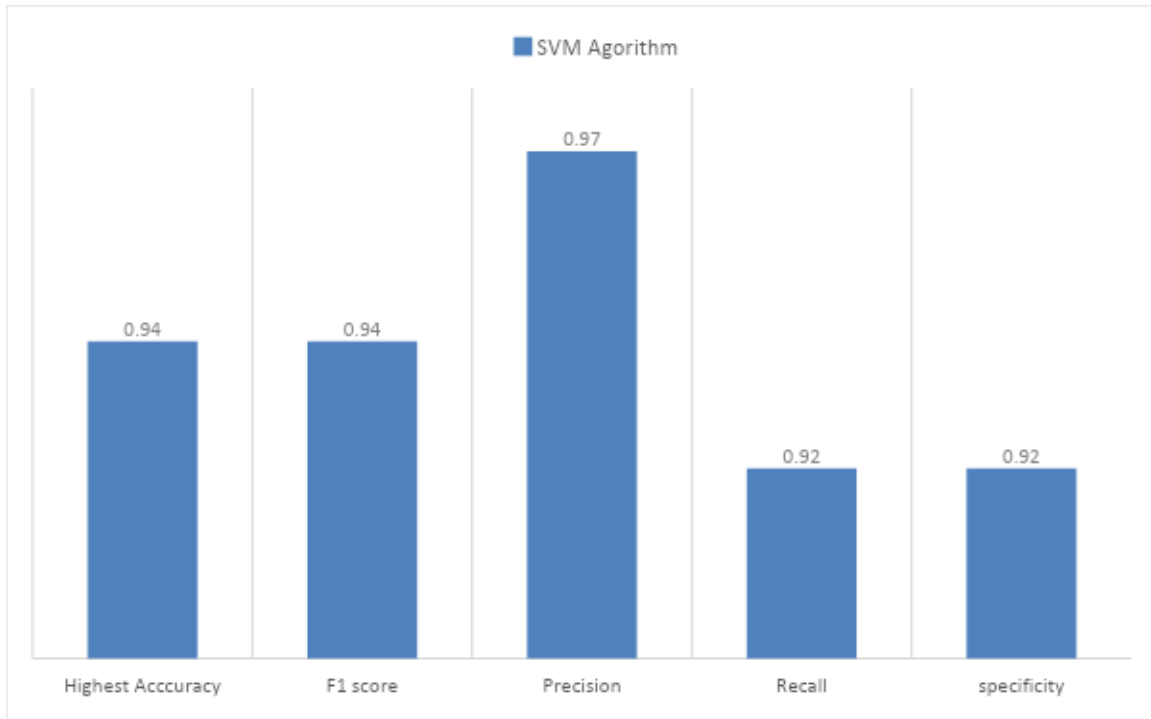


Figure 4.4 comparison graph of SVM

The finest run was determined by SVM calculations. The greatest degree of precision was 94%, and the other results was likewise extremely near to the accuracy values. Figure 4.4 shows the entire rating system graphically.

5.2.4 RANDOM FOREST

Sort forests, which do not require hyperparameters, provide A versatile and simple -to-use solution that produces excellent results in most situations. Also, due to its simplicity and adaptability, it is one of the most commonly used calculations and can be used for both classification and fallback calculations. In this essay, we will explore the inner workings of RFAI, its history, and its use in various calculations. If you create decision trees frequently, a "forest" will form when you are "rejected". The boxing technique's key tenet is the fact that the final result advances due to the use of numerous models for learning.. Random Forests can apply both a sort order and a fallback order. In our work, as a classification expert, with random

Thus, as illustrated in figure 4.5, you may obtain 92 percent accuracy and a precision rate of 97 percent.

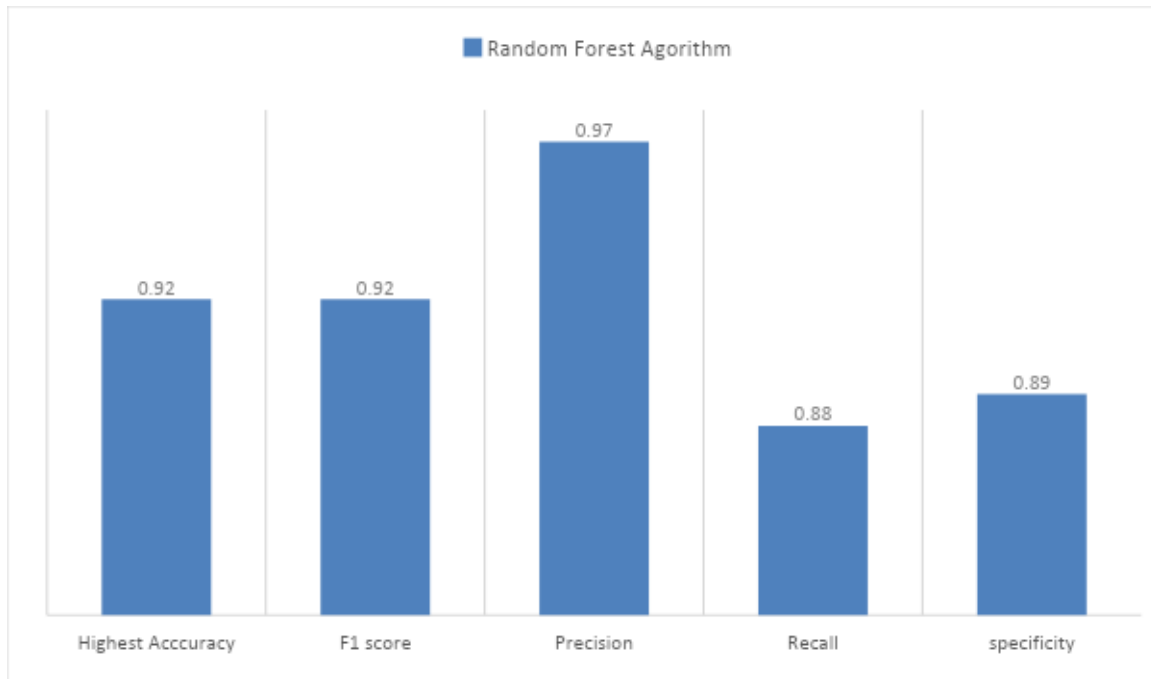


Figure 4.5 Score Comparison of Random Forest

5.2.5 LOSISTIC REGRESSION

Among the machine learning computations most frequently utilized in controlled learning approaches is fallback. It is used to determine the best way to measure a employing a group of wealthy variables to create a dependent variable that is categorical. The estimated recidivism rate predicts the yield of the dependent categorical variable. Results must be discrete or categorical. This can be any value from 1 to yes, true, or false. In any case, there is a probability value between 0 and 1, not an exact value between 0 and 1. Calculated recurrences are not used in the same way as direct recurrences. Calculated recurrence is used for classification, and direct recurrence is used to understand the difficulty of recurrence. challenge. [13] In our opinion, the

most notable points are: As visually shown in Figure 4.6, the ratio is 0.94.03 and the accuracy and correct rate are the same.

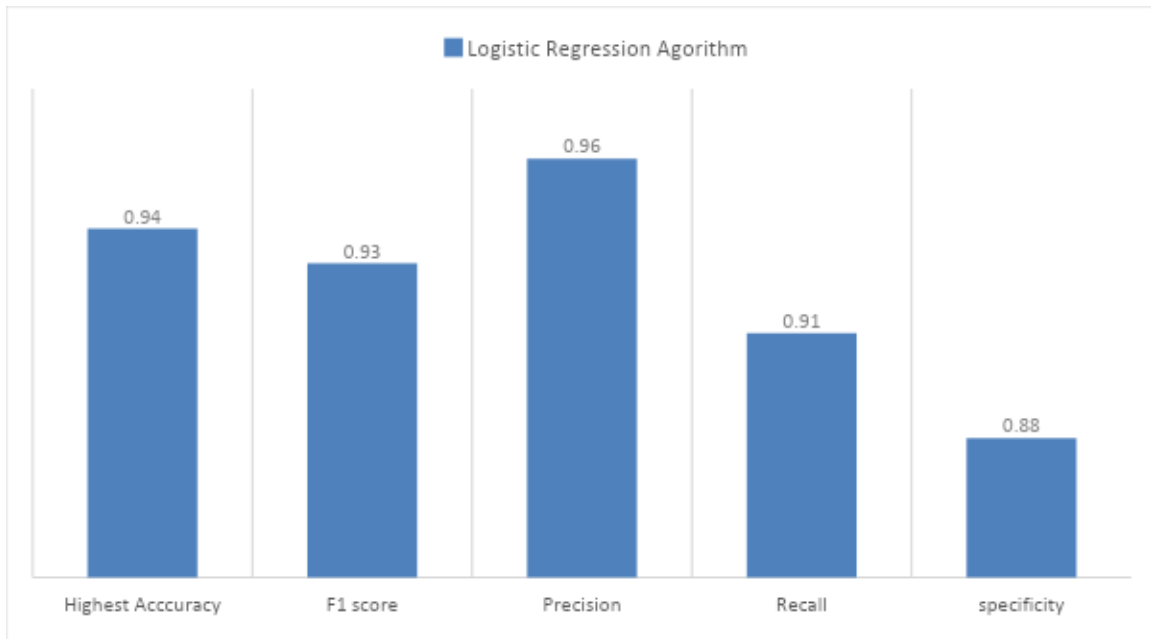


Figure 4.6 Comparison Graph of Logistic Regression

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND FUTURE WORK

6.1 SUMMARY OF THE STUDY

amount of investigation done in Bangle is quite modest compared to the large amount of research done in other languages on sentiment analysis of product evaluations using machine learning. Even though "working with a predictive approach" is another term for computer training. The fact that such occupations have a huge impact on our lives has led to the conduct of this type of research recently. There are several fascinating real-world applications that enrich our research. However, the Bangladeshi economy is relatively understudied. Our goal is to create an API system to inspect all product reviews within Bangladesh.

6.2 CONCLUSION

test accuracy is also about 94%, which is the same as the model accuracy. Among all methods, the most accurate algorithm is the SVM one. SVM won best perforation and surpassed other popular algorithms including Random Forest, KNN, Logistic, and Decision Trees. We have collaborated on his 6,000 product reviews written in Bangladesh from trusted online retailers. Using our proposed methodology, we can determine whether comments on online products are positive or negative depending on the emotions Bangladeshis express in their comments. Both customers and e-commerce authorities can choose the best products based on reviews. Both buyers and e-commerce operators benefit from this strategy.

6.3 RECOMMENDATION

Here are some very noteworthy ideas for this:

- ❖ Large training data sets are necessary to provide more precision with test data.
- ❖ Additionally, Deep learning algorithms such as CNN and Bengali Bertie, and LSTM can be employed while working with huge datasets.
- ❖ The Flask or Django rest frameworks can be utilized for deployment.

6.4 FUTURE WORK

My profession will continue to develop in the following ways going forward:

- Studies that concentrated on both favorable and unfavorable comments did not examine sarcastic or neutral comments. From now on, I will say neutral things and harsh things.
- When sarcasm is present, the sentiment expressed in the summary paragraph appears typical. However, such comments can be counterproductive in certain analyses, as sarcasm is difficult to predict in computers. Therefore, in the future we will build a system that can identify such statements.
- • In order to achieve this, we are now creating an online API for specifying analytical validation.
- To complete this task, we used Artificial intelligence techniques. Soon afterward, we'll meet up build intelligent systems using deep learning algorithms.
- We use only Bangla language in our work. On the contrary, customer comments are written on the bangle. We shorten your comments in order for our system to identify them.

REFERENCES

- [1]. (2020). Literature of Bangladesh, culture. Available: <<<https://www.bangladesh.com/culture/literature/>>> last accessed 10-10-2021
- [2]. M. R. Alam, A. Akter, M. A. Shafin, M. M. Hasan and A. Mahmud, "Social Media Content Categorization Using Supervised Based Machine Learning Methods and Natural Language Processing in Bangla Language," 2020 11th International Conference on Electrical and Computer Engineering (ICECE), 2020, pp. 270-273
- [3]. M. H. Rahman, M. S. Islam, M. M. U. Jewel, M. M. Hasan and M. S. Latif, "Classification of Book Review Sentiment in Bangla Language Using NLP, Machine Learning and LSTM," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-5
- [4]. A. Hassan, M. R. Amin, A. K. A. Azad and N. Mohammed, "Sentiment analysis on bangla and romanized bangla text using deep recurrent models," 2016 International Workshop on Computational Intelligence (IWCI), 2016, pp. 51-56
- [5]. M. T. Akter, M. Begum and R. Mustafa, "Bengali Sentiment Analysis of E-commerce Product Reviews using K-Nearest Neighbors," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 40-44,
- [6]. M. Rahman and E. Kumar Dey, "Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation," Data, vol. 3, no. 2, p. 15, May 2018.
- [7]. N. Mittal, B. Agarwal, G. Chouhan, N. Bania, and P. Pareek, "Sentiment analysis of hindi reviews based on negation and discourse relation," in Proceedings of the 11th Workshop on Asian Language Resources, 2013, pp. 45-50.
- [8]. S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts," 2014 International Conference on Informatics, Electronics & Vision (ICIEV), Dha-ka, Bangladesh, 2014, pp. 1-6,
- [9]. Kamiran, F., Calders, T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33, 1–33 (2012).
- [10]. A. Pinto, H. Gonçalo Oliveira, and A. Oliveira Alves, "Comparing the performance of different NLP toolkits in formal and social media text," in 5th Symposium on Languages, Applications and Technologies (SLATE'16), 2016: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
- [11]. J. M. Keller, M. R. Gray, J. A. J. I. t. o. s. Givens, man., and cybernetics, "A fuzzy k-nearest neighbor algorithm," no. 4, pp. 580-585, 1985.
- [12]. S. R. Safavian, D. J. I. t. o. s. Landgrebe, man., and cybernetics, "A survey of decision tree classifier methodology," vol. 21, no. 3, pp. 660-674, 1991.
- [13]. Logistic Regression in Machine Learning <<<https://www.javatpoint.com/logistic-regression-in-machine-learning>>> last accessed at 10-14-2021
- [14]. <https://towardsdatascience.com/word2vec-explained-49c52b4ccb71>

APPENDIX

The first, and certainly not least, duty was outlining the processes for the analysis. This was the initial report. It's also the first time this field has advanced in the past. Indeed, it is. Not just any job, though. Nobody who could help us was nowhere to be found. Another challenge and a big issue for us turned out to be gathering data. Having failed to find an application for processing texts that is freely available. for Bangladesh, we created a corpus of data instead. Our manual data collection process has begun. The procedure of classifying the various postings is equally difficult. After a lengthy and grueling period of work, we might succeed in achieving it.

Plagiarism

PRODUCT REVIEW SENTIMENT ANALYSIS BY TEXTVECTORIZAT AND MACHINE LERNING

ORIGINALITY REPORT

23% SIMILARITY INDEX	21% INTERNET SOURCES	14% PUBLICATIONS	10% STUDENT PAPERS
--------------------------------	--------------------------------	----------------------------	------------------------------

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	Rely Das, Md Forhad Hossain, Taufiq Ahmed, Ananyna Devanath, Shahnaz Akter, Abdus Sattar. "Classification of Product Review Sentiment by NLP and Machine Learning", 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2022 Publication	8%
3	Submitted to Daffodil International University Student Paper	4%
4	towardsdatascience.com Internet Source	1%
5	www.grin.com Internet Source	<1%
6	Submitted to HCUC Student Paper	<1%