A SENTIMENT ANALYSIS IN THE FIELD OF BENGALI TEXT : A MACHINE LEARNING APPROACH

BY

Shabikun Naher Eva
ID: 191-15-12677
AND

Md. Nazmul Hossain
ID:191-15-12074

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

Supervised By

Dewan Mamun Raza
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This project titled **"A SENTIMENT ANALYSIS IN THE FIELD OF BENGALI TEXT: A MACHINE LEARNING APPROACH"**, submitted by Shabikun Naher Eva, ID No: 191-15-12677 and Md. Nazmul Hossain, ID No: 191-15-12074 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 *24.01.2023*.

BOARD OF EXAMINERS

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

SArefin 24.01.23

Dr. Mohammad Shamsul Arefin — **Internal Examiner**
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

24.01.23

Md. Sabab Zulfiker — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

24.1.23

Dr. Ahmed Wasif Reza — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that this thesis has been done by us under the supervision of **Dewan Mamun Raza, Senior Lecturer, Department of CSE**, Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dewan Mamun Raza
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Eva

Shabikun Naher Eva
ID: 191-15-12677
Department of CSE
Daffodil International University

Nazmul

Md.Nazmul Hossain
ID:191-15-12074
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First of all, we want to render our gratitude to the Almighty Allah for the enormous blessing that makes us able to complete the final thesis successfully.

We are really grateful and express our earnest indebtedness to **Dewan Mamun Raza** Senior Lecturer, Department of CSE Daffodil International University, Dhaka, Bangladesh. Profound Knowledge & intense interest of our supervisor in the field of “Machine Learning & Deep Learning” make our way very smooth to carry out this thesis. Her remarkable patience and dedication, scholarlyguidance, continualencouragement, vigorous motivation, direct and fair supervision, constructive criticism, valuable advice, and great endurance during reading many inferior drafts and correcting the work to make it unique pave the way of work very smooth and ended with a great result.

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

We would like to express thankfulness to fellow student of Daffodil International University, who took part in this discussion during the completion of this work.
We would like to express our immense thanks to the Different food application to visible us user original reviews as a result we collected raw data to make our work possible.
We would also like to thank the people who provide the done by us to collect the market real information.

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

ABSTRACT

Now a-days, online marketing and e-commerce businesses in Bangladesh were thriving. Because it is the most secure way, online shopping has replaced traditional methods of buying after the COVID-19 epidemic. It reduces the amount of time required for businesses to launch their websites. More options for purchasing goods and services online are convenient and help consumers, but it also raises questions about reliability and safety. This makes it easy for unsuspecting new customers to fall victim to fraud while making purchases online. Our goal is to develop software that uses NLP to analyze customer reviews of online shops and provides a percentage breakdown of positive to negative feedback provided in Bangla (NLP). For the research, we compiled over 2003 user reviews and feedback items. We employed KNN, MULTI, RF, SGD, and SVC, as well as sentiment analysis as classification strategies. SVC achieved 85.7% accuracy, which was higher than any other approach.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Times are changing; Bangladesh is keeping pace with the outside world and is now taking its place in the list of developed countries. One of the few reasons behind the development of a country is its information technology development and how deep the relationship between information technology and the country’s people is .New technologies introduce us to novel methods that make our daily lives comfortable, excellent, and profitable. One of these technologies is buying and selling products online. It's easy to buy and sell online from home or anywhere. E-commerce is the medium of shopping using the Internet. It is a medium where one can easily buy and sell goods at home without hassle. Everyone is now opting for e-commerce for buying and selling due to easy access to the Internet. It is becoming more popular with the advent of new online payment methods. Among them, SSL is an online payment method that is very secure and fast. Day by day, people are becoming more and more involved in online shopping. E-commerce is being widely used for this. There are many benefits, like sitting at home without any hassle and shopping as you wish very quickly. Many new problems are also being created. People of our country are still not aware of internet resources and crime. As a result, online shopping is being cheated on in various ways. People looked at a product online, ordered accordingly, and did not get it as expected. It wastes time and money and is a lost cause in e-commerce. Despite all this, people are becoming more comfortable with e-commerce over time. There are many e-commerce sites in Bangladesh. Chaldal, Daraz, Evely, Bikkroy.com, and Rockmart are some of them. Due to a global epidemic, we were confined to our homes, which increased the possibility of online shopping. As more businesses move their operations online, the probability of rapacious capitalists selling defective products across the nation rises. Public opinion is crucial for identifying defective products and dishonest vendors in this situation. Even so, when an item has thousands of feedback and reviews, it is difficult for customers to determine the best. As people feel more comfortable being able to express themselves in their native tongue, the evaluations are written in Bangla. We planned to use reviews and comments written in Bangla to learn how people in that language feel about various products. We try to approach the issue distinctively. People sometimes seem to classify any amount as a point based on their assessment, so why do we need this review analysis once the exact number of participants is positive or negative? Uncertainty exists regarding how we evaluate positive and negative attitude responses using natural language processing (NLP). We have used some common machine learning algorithms for this data analysis, which have performed very well, they are Random Forest, Decision Tree, Support Vector Machine (SVM), KNN, and Logistic Regression.

1.2 Motivation

There has been a lot of growth in Bangladesh's online marketplace recently. Most consumers feel secure making purchases via the Internet. This is becoming more commonplace as a consequence of the corona pandemic. However, there are constraints on what online shoppers can do. One issue is that consumers have no way of knowing how products are generally received. Another choice is a product review, which is a study of a specific item, phenomenon, or set of documents. Reviews of products, advertisements for jobs, entire genres or industries, buildings, sculptures, designs, eateries, policies, exhibitions, and concerts are just some of the possible topics. The focus of this talk will be on presenting our evaluations of various products. Most consumers research items online by reading reviews before buying. To get a feel for a product in general, reading reviews is a smart move. Reviews of products, advertisements for jobs, entire genres or industries, buildings, sculptures, designs, eateries, policies, displays, and performances are all within the realm of possibility. The focus of this talk is on presenting evaluations of various products. The vast majority of shoppers today check out customer reviews on Amazon before making a purchase. You can learn a lot about a product by reading reviews. Further, we must decide whether or not to develop an AI capable of reading online reviews and sorting them into positive and negative categories. What, then, is a grade? It's a breeze to choose the top-rated items on a website when you use ratings. However, the ranking is determined by other factors. Examples of service indicators include product quality, delivery timeliness, and the friendliness of the delivery person. Therefore, it's not the best method for selecting a good. We've settled on using

NLP and ML to solve this issue. We are all aware that strings are not easily understood by computers. The first step is to turn the string into a number. The TFIDF algorithm was utilized here. Each comment was then put into one of several categories determined by the Machine Learning algorithm. Settings were changed for each algorithm. The optimal results were achieved with these settings.

1.3 Problem Definition

The majority of people's time in Bangladesh is now spent shopping for necessities. As the online world and e-commerce are becoming more commonplace in Bangladesh, so does the number of products available to buy online. Online shopping through e-commerce sites is rapidly gaining popularity. To provide the highest quality of service to customers while simultaneously saving time, this research was carried out. Machine Learning and NLP were used in this work. The effort has produced several problems that need to be solved. Due to the sensitive nature of our research topic, gathering relevant data has proven to be a formidable challenge. We did research by checking out several online stores. And I made sure to keep track of every product review in my head. The feedback we got was both good and bad. To develop this feature, we rely on the data provided. There were a total of 2,003 comments collected in Bangla. This is the raw data we have, and it's full of mistakes like typos and extra punctuation, and even emojis. So that our method could learn effectively, we removed all of these sounds during the preprocessing phase. After doing some preliminary processing, we used the TFIDF technique to convert the text to a numerical form. Several different Classification Machine Learning algorithms were used to keep up with the rapidly evolving numerical formats required by our task. Both the positive and negative aspects of each statement are identified and labeled. When we are done with the training phase, we will assess our performance using raw, untampered data. Our method is superior to rival approaches in the testing phase. Each tier was represented by its unique graph.

1.4 Research Questions

- What methods are used to gather data?
- Which contributed to the evolution used for the whole project's creation?

- How do the product's positive and negative values get represented?
- The machine learning process's ability to predetermine Positive and Negative classes.
- Is it feasible to apply the learned skills and reasoning to a functional e-commerce platform?
- How does one go about applying your idea to actual people?

1.5 Research Methodology

The study of research methodology entails systematically planning an investigation by a researcher to ensure the reliability of the results and the achievement of the study's stated aims. In this section, we'll go over the steps in our process, from gathering raw data to applying an algorithm to classifying that data—training models and assessing algorithms.

1.6 Research Objectives

- That is, classification methods can be used to dissect consumer sentiment analysis.
- To guide buyers to the best product for their needs.

1.7 Research Layout

The substance of our study is as follows:

An Unprecedented Start: The primary aspect of the preliminary inquiry plays a significant role in the process as a whole, making it very vital. In addition, the chapter delves into the thought processes that led us to conclude that conducting the research was the best course of action. The description of the issue is the most important component of this chapter. In this part of the article, we discuss the issues that have arisen with the research as well as the challenges that come with writing product reviews.

A Fresh Beginning An input analysis is what makes up this, and it gives a high-level description of the work that has been done in this area previously. Within this section, the work that was done utilizing machine learning that is relevant to this is broken down and explained.

In Chapter 3, a technique or procedure is dissected into its component pieces, and a comprehensive overview of the method or procedure is presented. What findings did you come to based on the inquiry that was carried out in this section?

In Chapter 4, an examination of the findings, will provide the solution to the problem. It is made up of the findings that were obtained via the graphical analysis.

In Chapter 5, the investigation that I've been carrying out has reached its conclusion. In this section, the outcomes of the model are broken down and discussed. By supplying further evidence, this section of the text additionally substantiates the reliability of the relationship. In addition, the online implementation of the concept and performance is provided in this section. The final half of this chapter delves into the constraints that were placed on the work. Additionally, an encoding of the potential of the inquiry was performed**.**

1.8 Expected Outcome

- We can determine whether a customer feels positively or negatively about a product.
- First, we'll do everything in our power to save customers time, and second, we'll do everything in our ability to present the most excellent product possible based on the consumer's needs.
- To show the results of any product review's emotional tone, we built a robust online API.

1.9 Summary

This chapter outlines the primary components that make up our organization's framework. We place the utmost importance on this particular chapter. In this chapter, we will offer an overview of our general framework, as well as a few related frameworks, our inspirations, our objectives, and our commitments to this framework. In addition, we will discuss a few related frameworks. In this chapter, we take a look not only at our overarching framework strategy, but also at the ways in which we might get out of this particular jam.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

The history of the study provides background for the concepts presented here. Therefore, the survey knowledge tickles the reader's attention and explains why the research topic is meaningful. For instance, in a study's introduction, you might talk about how different students' families' socioeconomic status affects their study habits or the range of their final grades. The best person to decide what information to include in the study's background is you, so take this only as an example. This chapter summarizes many competent specialists' efforts in the preceding area.

2.2 Related Works

The majority of today's services can be found online. People are free to share their thoughts and feelings on the web. Researchers usually use the object to determine how people are feeling. This subject has been discussed in several different languages and cultural contexts. The following works are references to aid in developing our task.Reviews and star ratings are essential resources for book readers. We aim to assess the quality of the language evaluations provided by Hamidur et al. in Bangla[1]. And to give readers valuable information about books and online bookstores so they can find the books they want to read and get the best service possible. Six thousand two hundred eighty-one unprocessed data points are used for the machine's training process. Their study classified binary (positive or negative) sentiments using machine learning and deep learning techniques. Text organization is done using natural language processing (NLP). Techniques like decision trees, the k-nearest neighbor algorithm and some of the other machine learning algorithms that provide good language detection are used to categorize human opinion.With LSTM, we achieved a record-high 97.49% accuracy. Safin et al. report that people in Bangladesh regularly use social media to voice their opinions in Bengali[2]. Using only textual information to put them in a category is impossible. It's not easy to label content shared on social media. Bengali text is more challenging to decipher than other language versions. Search, filter, and organize these comments based on the post's sentiment and rank them on the social media sites where they are posted. They utilized sentiment analysis to figure out why specific posts were so persuasive. A model was created to categorize Bengali posts using most of the common machine-learning techniques. They used a Bengali-speaking algorithm that produces the most exact answers when classifying social media posts. The best results can be obtained with a precision of around 88% using a logistic regression algorithm.M. T. Akter et al. proposed a model based on machine learning that can detect good, bad, or both comments[3]. They collect data from Bangladeshi e-commerce websites and use traditional machine learning algorithms on their collected dataset. KNN outperforms the other methods on every metric they consider adequate. In their modelsk-nearest neighbor achieves a 0.96 f1-score, or 96.25 % accuracy in detection.Attitude-based sentiment analysis approach was used in a Bangladesh study by Rahman et al. Bengali sentiment analysis is improving and is currently the main area of study[4]. Here, resource gathering is challenging. It isn't easy to complete tasks in Bengali, such as data collection, organizational speech evaluation, linguistics as part of the speech classifier, and many others.As part of their review of a cafe, they employed aspect-based studies to get comments from crickets. SVM has a reasonable and reliable of 71% and 77% for obtaining and exploring intensity in restaurants and insects, including both. To analyze Hindi, Mittal et al. created a technique with a good validity of 82.89 % and poor reliability of 76.59 %[5]. Since database consistency was a priority, they decided to conduct an emotional assessment and expand the database's scope. An instructional program is described here that looks into the Rom Pakistani person's emotions through the lenses of games, technology, food and cuisine, theater, and politicians. More than ten thousand sentences are taken from five hundred different online talks. The program's overarching objectives are (1) creating a brain library for Roman Urdu sentiment classification and (2) assessing the efficacy of analytical attitude techniques using regulation and N-gram (RCNN) models. In the Bangla language, S. Chowdhury et al. designed a model that could automatically remove a person from the group if they spoke a different language[6]. After using the proposed approach to a thousand and three hundred rows of col-selected data, the Support vector machine achieved a performance of

93% with its unique properties. Assumptions, emotions, and written perspectives may come together in what we call "sentiment analysis" (SA). Regarding popular dialects, SA preparation represents the most challenging aspect of the competition. Social media sites like Facebook frequently discuss the same piece of organic matter from multiple perspectives. The customer voiced his opinion on the subject in question in the news comments section of a daily publication.

C. Feng et al. found that digital media, such as websites and specialized mobile applications, are gradually replacing traditional print newspapers[6]. The quality of long-form articles can be automatically evaluated by news recommendation systems, and suggestions for further reading can be made to readers based on this and other characteristics. The authors of this study combed the available literature from 2001-2019 and came up with a list of 81 interconnected factors, which they then sorted into six broad classes before discussing in greater detail. Concerns from a wide variety of news industry difficulties are allayed by the fact that 60% of news proposal frameworks use a hybrid approach and 66% consider tiny conversations almost datasets. This is the first comprehensive look at the metrics used to recommend news. In conclusion, a number of potential future study directions are proposed, all of which have the potential to improve the story-suggesting abilities of news organizations.

Online product criticism is becoming more sophisticated every day. Acknowledging a person's achievements is greatly aided by auditing and findings of this kind. Prediction analysis can unearth nuanced information. Based on the information provided, we found that there isn't any significant book review movement in our country. Whenever the two works are compared, our model has the most extensive collection, the best goal, and outstanding achievements in more disciplines. Our content could end up on the internet.

2.4 Research Summary

Fellows of several research groups look at what studies have been performed in the field of historical assessment. Our group is working efficiently. Although the referendum was not entirely calm, it is thought that following the division, it might be more prudent to consider the finer details of buying various items.

2.5 Challenges

One of the greatest challenges researchers confront during this procedure includes organizing various data sets requiring assistance. Additionally, we utilized a few powerful and appropriate ML software tools that produced a large dataset valuable in our research for quality assurance. An additional difficulty in our country is the lack of enough funds and organization. Applying a Machin learning perspective to the internet environment is one of our most significant obstacles.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Our research paradigm comprises six steps: collecting data, evaluating it, putting the program into action, testing the method, and making web apps. Our research diagram is shown in Figure 3.1.

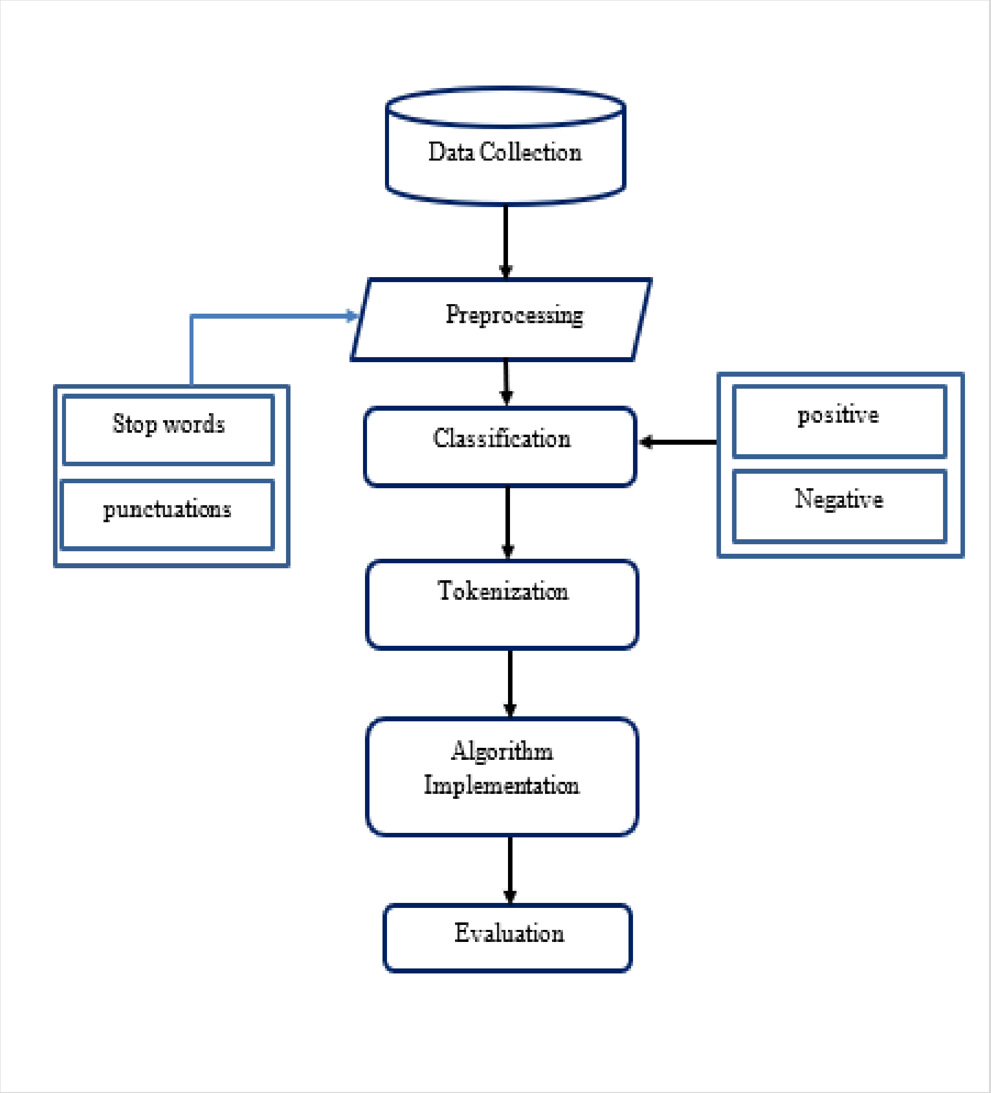


Figure 3.1 Methodology Diagram

3.2 Data Procedure

Research depends on the collection of information. To succeed in business, product reviews must be kept private. Also, we need to make sure we're getting our data from trustworthy places. The thoughts of products are the source of information used in our research. We compiled this data from customer reviews posted on Facebook and websites selling books online. Following our mandate, we collected responses only in Bangla. We mainly focused on e-commerce websites like Daraz, rokomari.com, aladaboi.com, and evaly.com.

3.3 Data Pre-Processing

As a technique for data mining, preprocessing reformats unstructured information so it can be processed effectively. Explicit knowledge relies heavily on preparatory work performed on incoming data. To create our model, we applied the KDD framework. Kamiranet al.[7] state that the four most crucial preprocessing operations are null hypothesis testing, data cleansing, data transformation, and poling. There are two subcategories within our data preprocessing phase. Both of these processes include getting rid of unnecessary syntax and stop words. Data messaging strategies were heavily utilized in our endeavor to create easily consumable data sets. To make the Bangla stop more concise, we cut out some extra information and words. Our recent comments will be used as tools to complete each procedure.

3.4 Classification

Positive and negative data were divided into two categories for our analysis. The user's emotions are considered when creating the courses. This sentence will receive a positive grade if the great analysis is sound. Like as, we've selected groups for negative evaluations. Figure 3.2 depicts how we gathered the data. 44.0% of the 3000assessments we gathered were favorable, while 56.0% were unfavorable. We can see from this graph that our dataset has the right balance to enhance our data's accuracy manually.

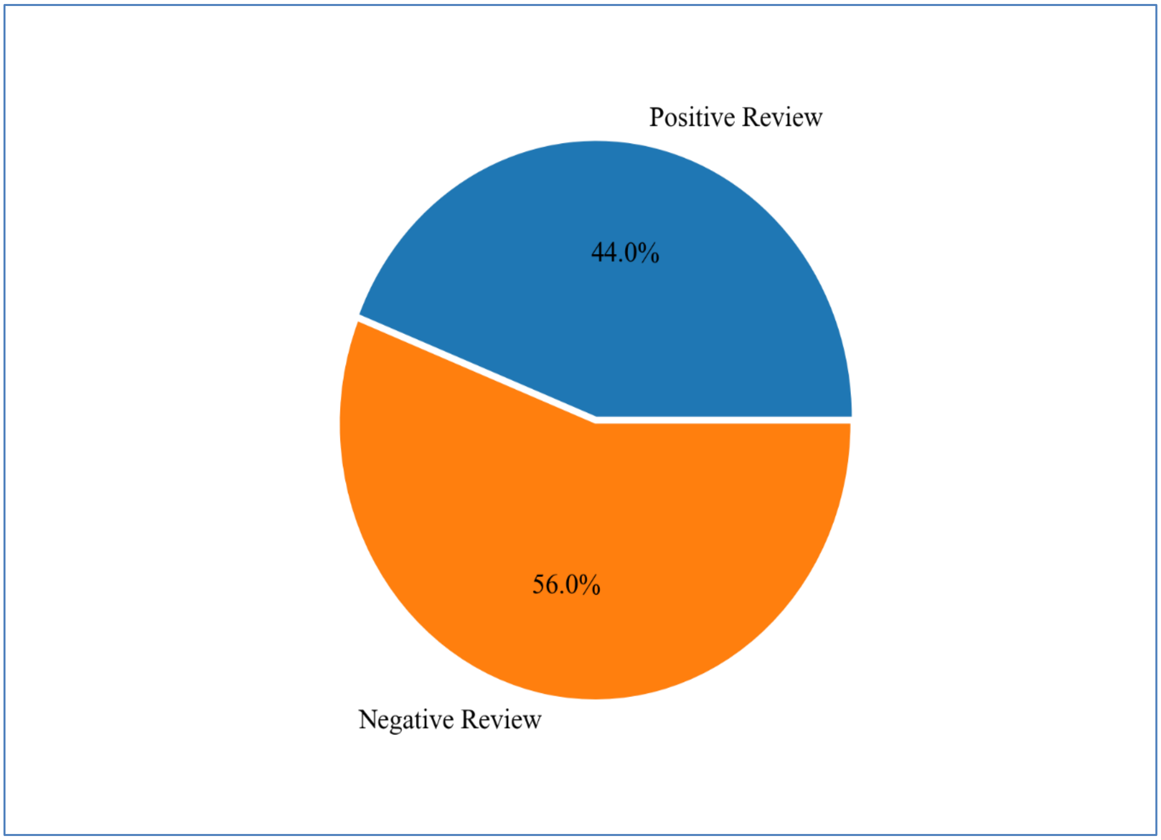


Figure 3.2: Representation of Class

3.5Tokenization

Sensitive data is converted into non-appealable "tokens" that can be utilized in a system or internal system without presenting it to the public through the technique of tokenization. Despite being unconnected numbers, the tokens' retention of initial data characteristics like length and format enables uninterrupted usage in daily operations. The original sensitive information is then kept in a secure location away from the business's systems. Tokenization is a method of dividing flag statements that may be words or symbols, as demonstrated by Safin et al. Our collection contains a lot of sentences. Instead of following sentence marks, we followed word labels to complete our task. Tokenization is additionally crucial. By tokenizing, we break up our entire statement into terms. There are many libraries for tokenization. However, we choose to utilize Scikit-Learn.

In Table 3.1, the tokenization technique is displayed.

TABLE 3.1 TOKENIZATION TABLE

Raw Data	**Type**	**Tokenized data**
তারাআসলপণ্যদেয়	Positive	‘তারা’, ‘আসল’, ‘পণ্য’, ‘দেয়’
ইয়ারফোনখুলতেইএকটামাইকখুলেগেছেপুরোটাকায়নষ্ট	Negative	’ ইয়ার’,’ ফোন’, ‘খুলতেই’, ‘একটা’, ‘মাইক’, ‘খুলে’ ‘গেছে’, ‘পুরো’, ‘টাকায়’, ‘নষ্ট’
নকলপণ্যেরওএকটারকমআছে	Negative	‘নকল’, ‘পণ্যের ‘ও ‘, ‘একটা ‘, ‘রকম’, ‘আছে’

3.6 Algorithm Implementation

In this section, we covered the integration process of the algorithm. We must execute the previous step to produce the necessary dataset before performing this. We have five different categorization techniques because our job is in the classification form. We use KNN, Random Forest, Logistics, Decision Tree, and Random Forest as our five classification methods. Table 3.2 displays the parameters that will provide the most accuracy for each approach.

TABLE 3.2 PARAMETER USAGES

Algorithms	**Details**
SVM	random_state = 42, kernel='linear'
Random Forest	Min_samples_split = 3n_estimators=100
Logistic Regression	Penalty= ‘l2’, tol = 1e-4
KNN	K = 5, random_state = 42

Decision tree	N_estimateros = 50, random_state=42,
	learning_rate = 50

3.7 Evaluation

Using an ambiguity vector and reliable data estimate, we assessed the performance of our chosen SVM algorithm. Thirty-seven fundamental data points were originally collected, but our algorithm could not learn from them. Various websites for ebook sales and Facebook book reviews were used for each of the sessions. A contrast between the anticipated and actual outcomes is shown in Figure 3.3. Our dataset includes 20 favorable reviews and 17 unfavorable ones, which are conducted by green bars. The orange color bar displays the value that our model predicts. Our model expects two additional ratings for a decent mark. Two fewer reviews are anticipated in the model of bad reviews. Our model has a slight defect like this. We can assume that our model performed well with data from the actual

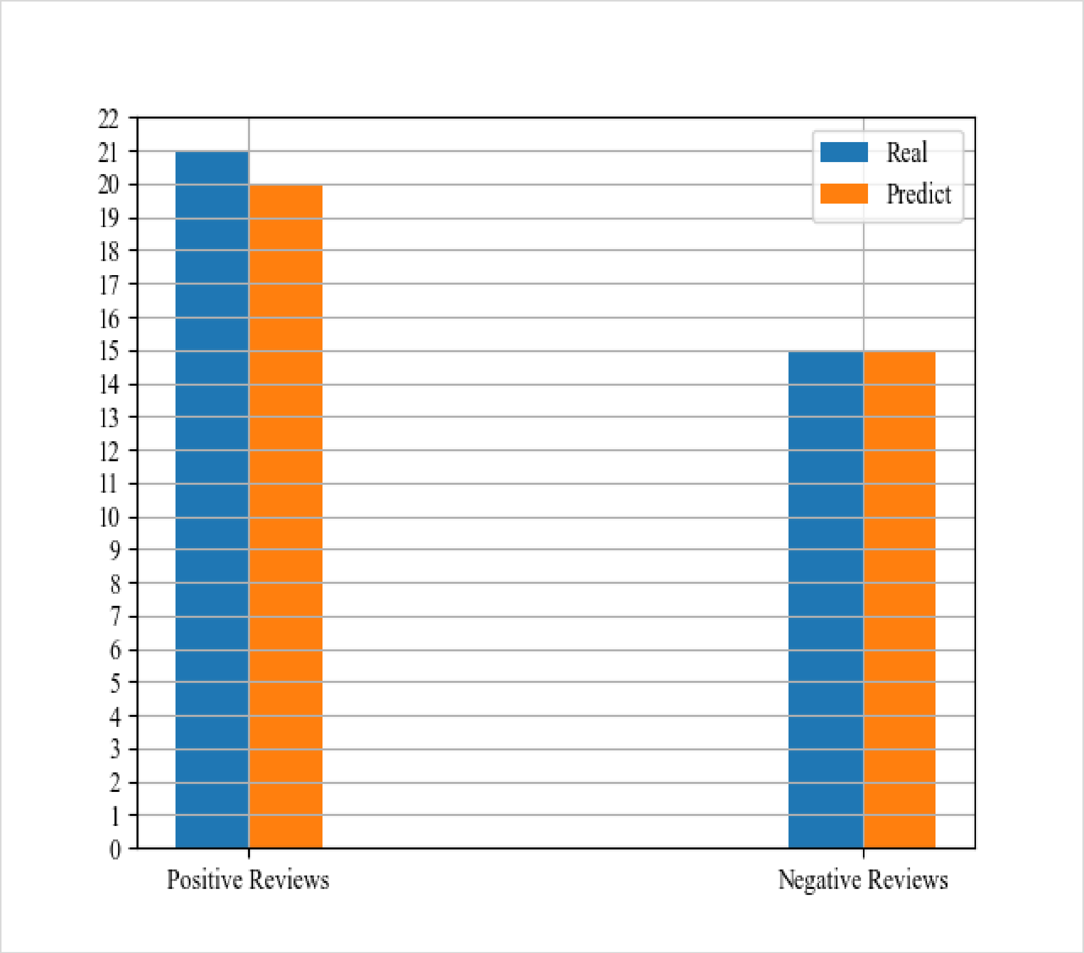


Figure 3.3: Comparison Between Real and Predicted

Figure 3.4 represents our real dataset persistence. Our collected data showed mixed positive and negative reviews of people involved in online shopping. The method conducted a 54.1% positive review and a 45.9% negative review in our data set. The data is most similar to real collected data. People gave their reviews in online media as their own mindset there is no grammatical fellow. So, our data is very much complicated for the understanding of our applied method but our method does better our aspect.

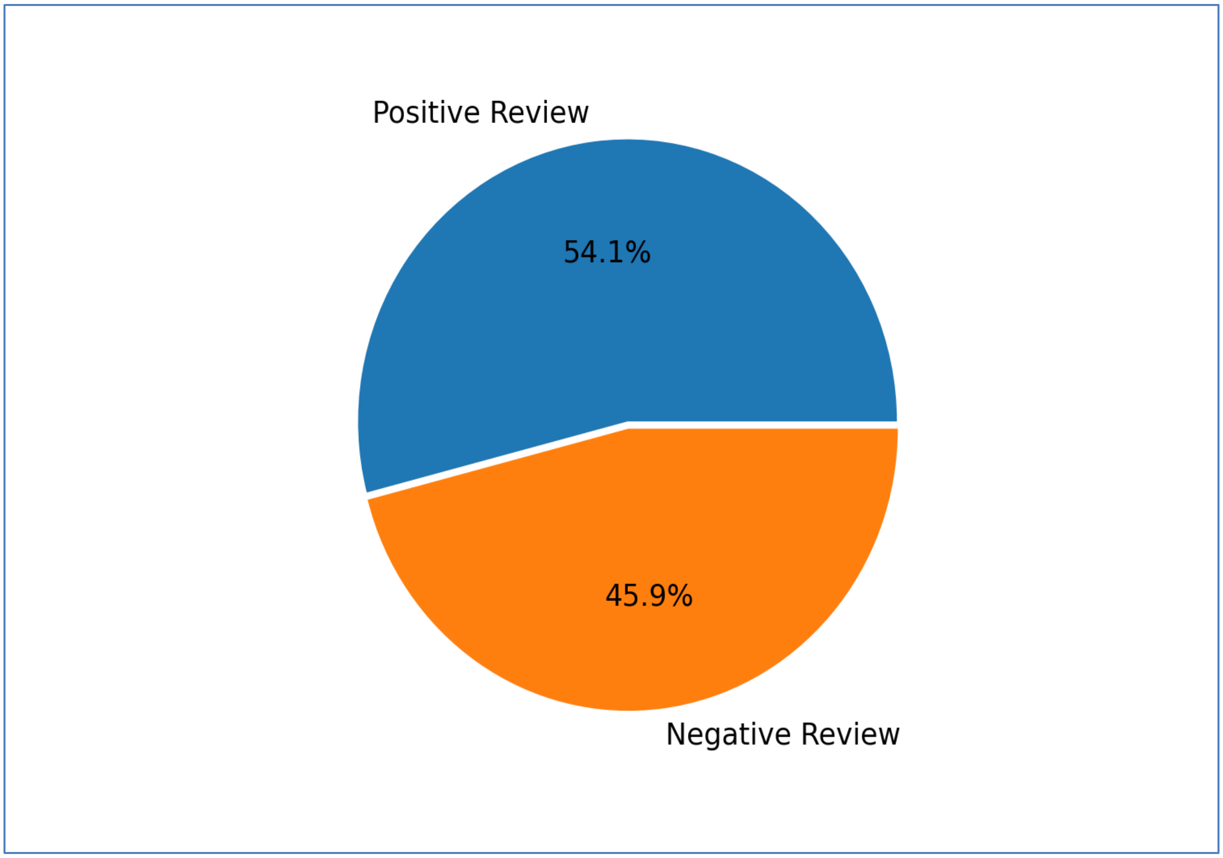


Figure 3.4 Show Real Positive and Negative Review

Figure 3.5 show the predicted data for a positive and negative review. Our algorithm performed very real state results as our collected data set. Our method can work with both visible and hidden data sets. In data analysis, there is real positive data at 54.1% and our method predicted much more at 59.5% which was silly similar to the real data set. Besides our real negative data was 45.9% but our method could predict the negative review of very few as 40.5% which was an outstanding performance.Rather than being the reason for low scores, it's a great example of our idea. Data consistency is very well we find as in this section analysis.A slight difference between the predicted and actual data does not significantly differ in accuracy. It turns out that our methods perform well on the data.

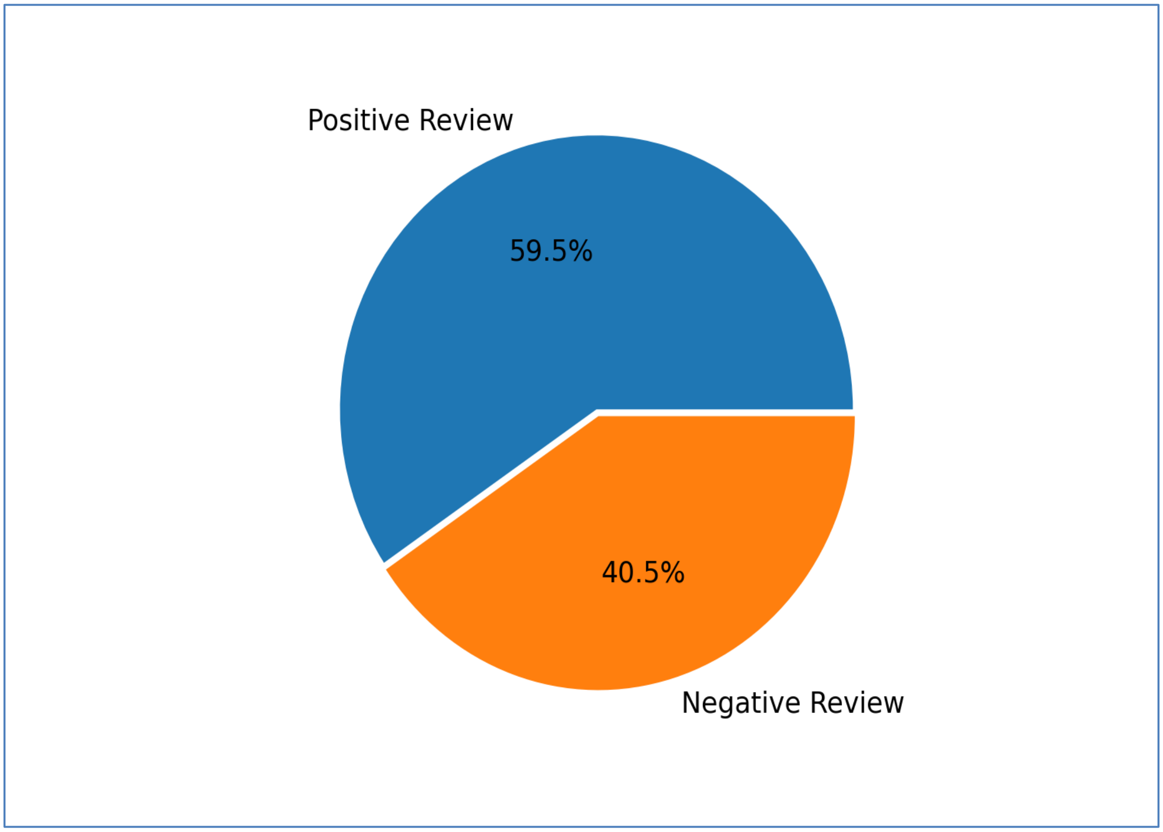


Figure 3.5 Shows Predilected Positive and Negative Review

CHAPTER 4

RESULT ANALYSIS

4.1 Introduction

Active but unfinished tasks, including repair orders, delivery dates, users to put in, and programs, are assessed by the RA mechanism. Outcome analyses that focus on the allocation of resources are one form. The effects section needs to be formatted, so the results are given without any assessment or analysis. As with other areas of the term report, help is at hand. The results are revealed, and the examination is demonstrated. Several distinct algorithms were analyzed, and we'll review our findings and recommendations. The criteria for producing this data set were selected to maximize resolution, consistency, recall, and f1.

4.2 Experimental Result

Test data usage rate		20%	25%	30%	35%
Algorithms Accuracy	*RF*	83.04	83.83	83.19	81.17
	SVC	85.79	84.23	82.70	82.31
	SGD	85.54	84.23	83.19	82.31
	Multi	81.80	81.84	80.37	80.03
	KNN	68.58	69.06	68.22	64.48

TABLE 4.1 ACCURACY TABLE

Table 4.1 shows the accuracy of the algorithms used in our research. In this case, we took test results between 20% and 35% of what we wanted to see which algorithm could do better. In the table, the yellow boxes show how accurate each algorithm is based on how much of an effect it has. All the algorithms except KNN showed good accuracy in 25% of test results. Each algorithm performed best when the test result was within 20% of all algorithms, with the highest SVC and SGD results of all algorithms used in the research.SVC and SGD outperformed all others, but SVC exceeded all others in 20–35% of test results and reached 20% of the highest accuracy.

4.3 KNN

The K-Nearest Neighbor method can be used as a supervised learning technique for tasks like classification and regression. It's a flexible technique that may be used to resample data sets or fill in missing values. A K-Nearest Neighbor is a method for predicting a new data point's classification or constant value based on evaluating its K-nearest neighbor. [11].Because, as the name implies, the closest neighbors are the reference points in the training dataset nearest to the new value. The number of such data points that we use in our execution of the procedure is denoted by KNN. Table 4.2.1 show the KNN algorithm accuracy analysis.

Label	Precision	Recall	F1	Support
Positive	0.71	0.46	0.56	173
Negative	0.68	0.86	0.76	228
Accuracy		0.69		401
Macro avg	0.69	0.66	0.66	401
Weighted avg	0.69	0.69	0.67	401

Table 4.2 KNN Algorithm

The three columns in our KNN accuracy table are precision, F1, and recall rate, with an extra row of macro and weight averages. In the KNN algorithm, for positive reviews, the height precision rate is 0.71, the recall rate is 0.46, and the f1 rate is 0.56. Similarly, the height precision rate is 0.68, the recall rate is 0.86, and the f1 rate is 0.76, as shown for the negative review. The macro and weighted average act as equal accuracy rates. The accuracy delivered by the KNN algorithm is 69% of 401 test data.

4.4 RF

It's possible that random forest is a flexible, user-friendly algorithm that consistently produces positive outcomes without the use of model parameters. The fact that it is straightforward and adaptable enough to be used for classifying and recovery estimations makes it one of the most frequently utilized computations. In this post, we'll learn how the RFAl functions, where it originates from, and how other calculations use it. It produces an "option to utilize" with regularly arranged decision-making trees through "dismissing." The ultimate result of the box technique is improved by a mixture of learning methods, which is its most significant rule. With random forest, both categorization and regression techniques can be applied.

Label	Precision	Recall	F1	Support
Negative	0.90	0.69	0.78	173
Positive	0.80	0.94	0.87	228
Accuracy		0.84		401
Macro avg	0.85	0.82	0.83	401
Weighted avg	0.85	0.84	0.83	401

Table 4.3 RF Algorithm

The RF algorithm accuracy analysis is displayed in Table 4.2. Our RF accuracy table has three columns: precision, F1, and recall rate, with an additional row for macro and weight averages. For positive reviews, the RF algorithm's height precision rate is 0.80, recall is

0.94, and the f1 rate is 0.87. Similar to the values shown for the negative review, the height precision rate is 0.90, the recall rate is 0.94, and the f1 rate is 0.87. Equal accuracy rates are produced by the weighted average and the macro. The accuracy of the RF algorithm was higher than that of the previous algorithm. The RF algorithm's accuracy for the 401-test data is 0.84%.

4.5 SGD

An all-purpose optimization process called gradient descent can locate the best answers to various issues. The basic notion is to change parameters to reduce the differential equation incrementally. An essential Gradient Descent (GD) parameter is the learning rate input parameter, which controls the size of the steps. It will take a very long time for the algorithm to converge if the learning rate is too low, and it may cause us to skip the optimum values if it is too large.

Label	Precision	Recall	F1	Support
Negative	0.91	0.73	0.81	173
Positive	0.82	0.94	0.88	228
Accuracy		0.85		401
Macro avg	0.86	0.84	0.84	401
Weighted avg	0.86	0.85	0.85	401

Table 4.4 SGD Algorithm

The accuracy of the SGD method is shown in Table 4.2.3. Precision, F1, and recall rate are the three columns in our SGD accuracy table, plus an additional row of macro and weight averages. For good reviews, the height precision rate is 0.82, the recall rate is 0.94, and the f1 rate is 0.88 in the SGD method. As demonstrated in the unfavorable evaluation, the height precision rate is 0.91, the recall rate is 0.73, and the f1 rate is 0.81. The two algorithms show height result SDG is one of them. The macro and weighted average both have the same accuracy rate. The SGD algorithm delivers 0.85% accuracy on 401 test data points. Table 4.3 SGD Algorithm.

4.6 SVC

The SVC might be the best option for categorization and recovery training. Most often, it is used in problems that need grouping. In the classifier, each data point is assigned a coordinate in an n-dimensional area where a higher value is placed on closer cooperation. Then, we classify the vector that most effectively divides the two sets.

Label	Precision	Recall	F1	Support
Negative	0.92	0.71	0.80	173
Positive	0.81	0.95	0.88	228
Accuracy		0.85		401
Macro avg	0.87	0.83	0.84	401
Weighted avg	0.86	0.85	0.84	401

Table 4.5 SVC Algorithm

Analysis of the SVC algorithm's accuracy is shown in Table 4.3. Our SVC accuracy table consists of the three metrics of precision, F1, and recall rate, plus the additional metrics of macro and weight averages. In the SVC method, favorable ratings result in a height precision rate of 0.81, a recall rate of 0.95, and an f1 rate of 0.88. Similarly, the negative review demonstrates a height precision rate of 0.92, a recall rate of 0.71, and an f1 rate of 0.80. Similar precision can be achieved by using either the macro or weighted average. It is true that SVC works very fluently for human command or review.SVC algorithms show the best results of all the applied algorithms. The SVC algorithm provides 0.85% accuracy on a set of 401 test data points.

4.7 MULTI

A standard Bayesian learning method used in NLP is the Multinomial Naive Bayes algorithm (NLP). Using the Bayes rule, the computer attempts to determine what category a piece of content (such as an email or news article) belongs to. For a specimen, it calculates the likelihood of each label and returns the brand with the highest probability.

Label	Precision	Recall	F1	Support
Negative	0.89	0.66	0.76	173
Positive	0.78	0.94	0.85	228
Accuracy		0.82		401
Macro avg	0.84	0.80	0.81	401
Weighted avg	0.83	0.82	0.81	401

Table 4.6 MULTI Algorithm

An evaluation of the MULTI algorithm's precision is presented in Table 4.5. Our MULTI accuracy table consists of precision, F1, and recall rate columns, plus a row for macro and weight averages. For positive reviews, the MULTI method achieves a height precision rate of 0.78, a recall of 0.94, and an f1 of 0.85. Similarly, it can be seen that the negative review has an f1 rate of 0.76, a recall rate of 0.66, and a height precision rate of

0.89. Both the weighted average and the macro have equivalent precision. The MULTI method provides an accuracy of 0.82 percent on a sample of 401 test cases.

4.8 Analysis

Here we apply five traditional machine learning algorithms. Here SGD and SVC gained height score. But we need to choose one algorithm for our future implementation. The SVC gained best result of both at 30%. So, we selected the SVC for our prediction.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on society

Our project has the potential to make a significant contribution to the greater good.

- By reading the remarks, people can quickly determine whether the product is favorable or unfavorable.
- People of average intelligence will have an easy time determining whether a review is good or negative.
- No one will be taken in by a lie.
- The act of shopping online will become widely accepted as a reliable source of information.
- There will be a significant rise in the number of people shopping online. Which will contribute to increased social as well as economic progress

5.2 Impact on the environment

The list that follows is comprised of the positive impacts that our project will have on the natural environment in its immediate vicinity.

By utilizing our method, either favorable or unfavorable feedback can be easily identified, leading to an increase in the number of online purchases made compared to the previous period. Those individuals who are still determining which product they should purchase will have an interest in buying it. The number of things purchased on the internet will rise as a result of this. Its dishonesty will see a significant drop as a result. Because it is an automated process that uses internet product reviews, our method does not contribute to environmental damage in any way.

5.3 Ethical Aspect

The project that we are working on does not give rise to any specific ethical problems at this time.

Anyone who want to do so will have a difficult time updating the programming code. It will not be possible to personalize it in any way at all. As a result, the process of its change will not be a simple one. Additionally, we are not permitted to use this system in a manner that is dishonest of any kind. Our method can only be used to identify positive or negative reviews of a product. As a direct consequence of this, there won't be any one-of-a-kind ethical issues to worry about. People are able to make it work for them and get the benefits of doing so.

5.4 Sustainably plan

Some of the things we're doing to be environmentally responsible are:

- Working with the different types of feedback that are still available, our long-term goals include: (Baglish, Hindi). In order to prevent regular people from being misled into purchasing other kinds of things online, this must be done.
- We would like to improve the quality of our database. Our outcomes will be more favorable.
- imogi remarks will be taken into consideration in our work. So that imogi comments can be simply uploaded with a minimal amount of data required. It will result in a decrease in the price of internet service. The number of people interested in utilizing our system will rise.
- To ensure that it is accessible to the widest possible audience, we want to offer this system in both English and Bengali.

CHAPTER 6

SUMMARY, CONCLUSION, AND FUTURE WORK

6.1 Summary of the Study

The use of machine learning for analyzing the attitudes of customer evaluations has been extensively studied in all the other languages, but little is known about the topic in the Bangla language. Though "work in predictive methods" is frequently used to describe e-learning systems, researchers have recently begun investigating the issue due to the profound effect these jobs have on our daily lives. Promising real-world applications bolster our research. The economy of Bangladesh, however, is not being studied extensively. We plan to create an application programming interface (API) that can examine all Bangladeshi product reviews.

6.2 Conclusion & Future Work

Our model is nearly as reliable, with a 95% accuracy rate as our tests, which average approximately 95 %. The SVM classifier outperforms the competition in terms of efficiency. SVM has the highest accuracy and is the best method, beating off well-known ones like KNN, Logistics, Decision Tree, and Random Forest. We collected two thousand reviews from Bangladeshis and published them on popular shopping websites. Using the approach, we provide, we may analyze the tone of comments made about an item on the web and classify them as positive or negative, at least among the Bangladeshi people. Buyers and the e-commerce administration can use reviews and ratings to make purchasing decisions. Either online store owners or customers can gain from such a strategy.

The suggestions that follow are for the further advancement of this project:

1.In our research, we only looked at good and negative comments to determine the most important trends. From now on, we'll make snarky or indifferent comments.

2.The sentiment expressed in the last paragraph seems normal when sarcasm is present. However, in a more narrow examination, such a remark has the opposite impact because satire is challenging to predict on a computer. Therefore, we will create a technology in the coming days that can identify sarcastic comments.

3.Toward this end, we are developing a Web-based API to define the study review process.

4.To accomplish this, we relied on a machine-learning strategy. In the future, we hope to have created an autonomous system that uses deep learning algorithms.

5.Our job is conducted entirely in Bangla. In contrast, customers often post reviews written in Banglish. Because of this, we plan to begin teaching Banglish sentences so our program can understand Banglish remarks.

6.3 Recommendations

Here are some very exceptional recommendations:

1.Training data has to be quite vast if we want to have high accuracy on test data.

2.Moreover, Deep Learning algorithms such as LSTM, CNN, and Bangla Bert may be implemented with a sizable dataset.

3.Both the Django Rest Framework and the Flask Framework may be utilized for installation.

REFERENCES

[1]. M. H. Rahman, M. S. Islam, M. M. U. Jowel, 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

[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. 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,

[4]. 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.

[5]. 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.

[6]. CFeng; Wisetsri, Worakamol.Rise of Artificial Intelligence in Healthcare Startups in India Vol. 14, Iss. 1, (Mar 2021): 48-52.

[7]. 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,

[8]. [7] M. Riajuliislam, K. Z. Rahim and A. Mahmud, "Prediction of Thyroid Disease(Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 60-64

[9]. [8] T. Nguyen, D. Phung, B. Dao, S. Venkatesh and M. Berk, "Affective and Content Analysis of Online Depression Communities," in IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 217-226, 1 July-Sept. 2014.

[10]. [9] R. Razavi-Far, E. Hallaji, M. Saif and L. Rueda, "A Hybrid Scheme for Fault Diagnosis with Partially Labeled Sets of Observations," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 61-67

[11]. [10] K-Nearest Neighbor(KNN) Algorithm for Machine Learning, available at << https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning >> last accessed on 08-09-2021 at 8PM

[12]. [11] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255-1260

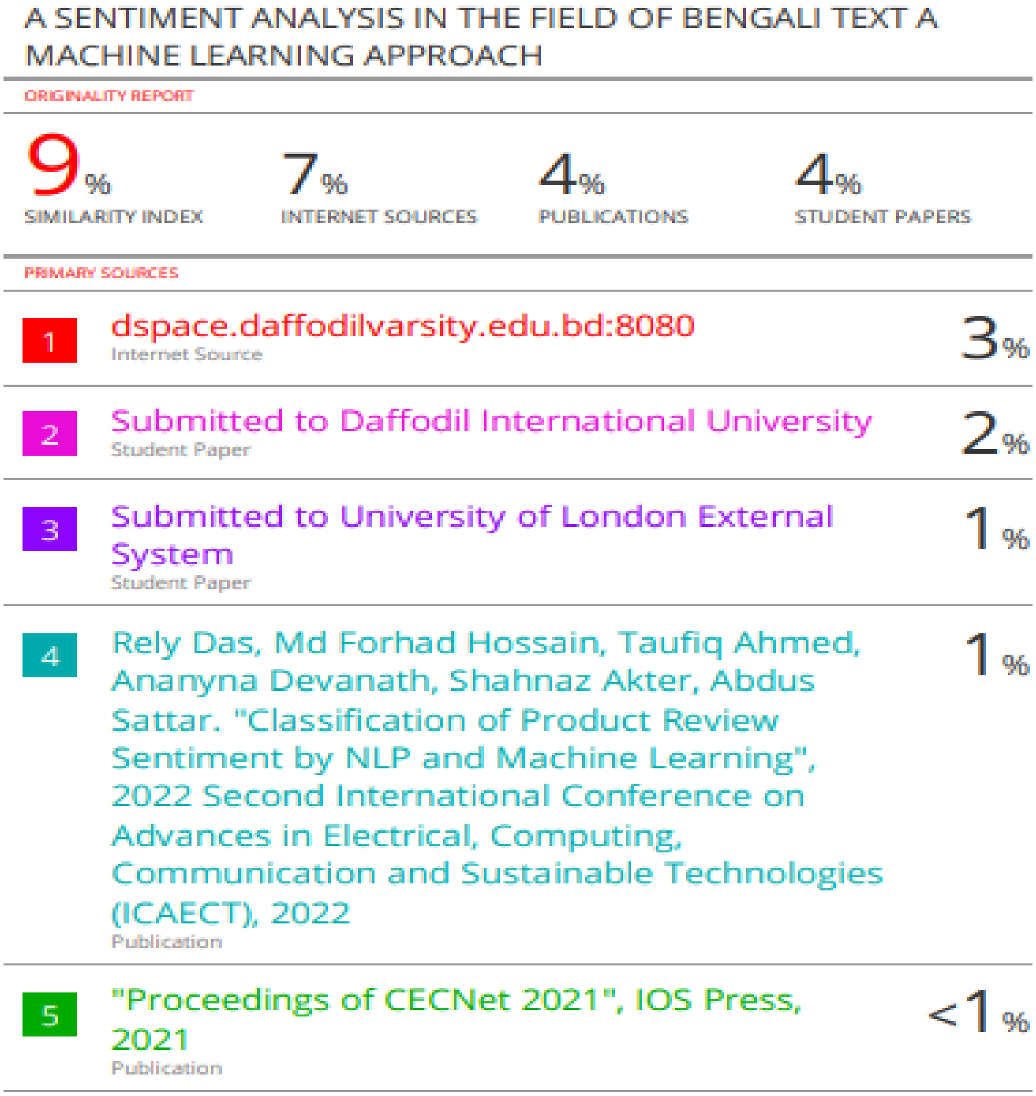
[13]. [12] Sri Krishnan, 6 - Machine learning for biomedical signal analysis, Editor(s): Sri Krishnan, Biomedical Signal Analysis for Connected Healthcare, Academic Press, 2021, Pages 223-264

[14]. [13] JOUR , Lv, Zhihan, Zhang, Zhifei, Zhao, Zijian, Yeom, Doo-Seoung, Decision Tree Algorithm-Based Model and Computer Simulation for Evaluating the Effectiveness of Physical Education in Universities , 2020, SN - 1076-2787

[15]. M. H. Aysa, A. A. Ibrahim and A. H. Mohammed, "IoT Ddos Attack Detection Using Machine Learning," 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2020, pp. 1-7, doi: 10.1109/ISMSIT50672.2020.9254703.

[16]. D. C. Grant, "Distributed detection and response for the mitigation of distributed denial of service attacks," 2018 International Conference on Information Networking (ICOIN), 2018, pp. 495-497, doi: 10.1109/ICOIN.2018.8343168.

PLAGIARISM REPORT



A SENTIMENT ANALYSIS IN THE FIELD OF BENGALI TEXT A MACHINE LEARNING APPROACH

ORIGINALITY REPORT

9%	7%	4%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	Submitted to Daffodil International University Student Paper	2%
3	Submitted to University of London External System Student Paper	1%
4	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	1%
5	"Proceedings of CECNet 2021", IOS Press, 2021 Publication	<1%