

**Bangla Book Review Analysis with Different Word Embedding and Machine Learning**

**BY**  
**Abdul Al Motakabbir**

**ID: 203-15-3864**

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

Supervised By

**Md. Sazzadur Ahamed**  
Assistant Professor  
Department of CSE  
Daffodil International University

Co-Supervised By  
**Mr. Abdus Sattar**  
Assistant Professor & Coordinator M.sc  
Department of CSE  
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**27 JANUARY 2024**

## APPROVAL

This Project/internship titled “**Bangla Book Review Analysis Different Word Embedding And Machine Learning**”, submitted by **Abdul Al Motakabbir**, ID No: 203-15-3864 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 **27 January 2024**.


### BOARD OF EXAMINERS

**Dr. Sheak Rashed Haider Noori**

**Professor and Head**

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

**Chairman**

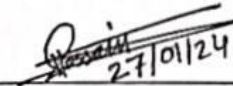
 27/1/24

**Dr. Arif Mahmud**

**Associate Professor**

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

**Internal Examiner**

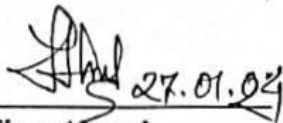
 27/01/24

**Mr. Shahadat Hossain**

**Senior Lecturer**

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

**Internal Examiner**

 27.01.24

**Dr. Firoz Ahmed**

**Professor**

Information & Communication Engineering  
Rajshahi University

**External Examiner**

## DECLARATION

I hereby declare that this thesis has been done by me under the supervision of **Md. Sazzadur Ahamed, Assistant Professor**, Department of CSE, and co-supervision of **Mr. Abdus Sattar, Department of CSE** Daffodil International University. I 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:**



---

**Md. Sazzadur Ahamed**

Assistant Professor

Department of CSE

Daffodil International University

**Co Supervised by:**

---

**Mr. Abdus Sattar**

Assistant Professor & Coordinator M.sc

Department of CSE

Daffodil International University

**Submitted by:**



---

**Abdul Al Motakabbir**

**ID: 203-15-3864**

Department of CSE

Daffodil International University

## ACKNOWLEDGEMENT

First of all, I want to render my gratitude to the Almighty Allah for the enormous blessing that made us able to complete the final thesis successfully.

We are grateful and express our earnest indebtedness to **Md Sazzadur Ahamed, Assistant Professor**, Department of CSE Daffodil International University, Dhaka, Bangladesh. The profound Knowledge & intense interest of our supervisor in the field of “Machine Learning & Deep Learning” made our way very smooth to carry out this thesis. Her remarkable patience and dedication, scholarly guidance, continual encouragement, 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 paves the way for work very smoothly and ended with a great result.

I would like to express our gratitude wholeheartedly to **Dr. Sheak Rashed Haider Noori**, Professor, and Head, of the Department of CSE, for his kind help in finishing our thesis and also to other faculty members and the staff of the CSE department of Daffodil International University.

I would like to express thankfulness to a fellow student of Daffodil International University, who took part in this discussion during the completion of this work.

I would like to express our immense thanks to the Different applications for the visible user original review as a result we collected raw data to make our work possible.

We would also like to thank the people who provided the done by us to collect the market real information.

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

## **ABSTRACT**

Different product review analysis recently attracted the attention of natural language processing specialists, thanks to positive customer comments and reviews spread around the web. The increasing number of e-commerce sites leads to a rise in the purchasing rate of various commodities. An illustration of this is the rapidly growing interest in literature among the general public. In today's era of internet technology, Bangladesh's e-commerce and online marketing sectors are already robust. Take online product reviews as an example; they've really taken off as a go-to resource for shoppers. Some say that a book is a person's closest companion. Books are indispensable for individuals as they offer insights into the external world, enhance literacy skills, and bolster memory and intellect. I aim to evaluate and rank reviews from Bangladesh and provide precise details on books and online bookstores. I objective is to aid book enthusiasts in making informed decisions when purchasing books and finding reliable online merchants. The word2vec and FastText algorithms were utilized to transform text into numerical values. I used five different classification algorithms, which are as follows: MNB, KNN, RF, SVC, and XGboost Classifier or Multinomial Naïve Bayes. An accuracy of 85.92% was attained by the Support Vector Classifier (SVC) using the FastText method.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Approval	i
Declaration	ii
Acknowledgments	iii
Abstract	iv
List of Figure	viii
List of Table	ix

## **CHAPTER**

<b>CHAPTER 1: INTRODUCTION</b>	<b>PAGE NO.</b>
	<b>1-5</b>
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 Report Layout	4
1.8 Expected Outcome	5
<b>CHAPTER 2: BACKGROUND</b>	<b>6-9</b>
2.1 Introduction	6
2.2 Related Work	6
2.3 Comparison of Related Work	8
2.4 Research Summary	9
2.5 Challenges	9

<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>10-13</b>
3.1 Introduction	10
3.2 Data collection	11
3.3 DataPre Preprocessing	11
3.4 Dataset Labeling	12
3.5 Tokenization	12
3.6 Algorithm Implementation	13
<b>CHAPTER 4: RESULT ANALYSIS</b>	<b>14-20</b>
4.1 Introduction	14
4.2 Experimental Result	14
4.2.1 Multinomial Naïve Bayes Algorithm	16
4.2.2 XGB Classifier	16
4.2.3 SVC	17
4.2.4 Random Forest	18
4.2.5 KNN	19
4.3 Evaluation	20
<b>CHAPTER 5: SUMMARY, CONCLUSION AND FUTURE WORK</b>	<b>22-23</b>
5.1 Impact on Society	22
5.2 Impact on the Environment	22
5.3 Ethical Expect	22
5.4 Sustainability Plan	23
<b>CHAPTER 6: SUMMARY, CONCLUSION AND FUTURE WORK</b>	<b>24-25</b>
6.1 Summary of the Research	24
6.2 Conclusion	24
6.3 Recommendation	25

6.4 Future Work	25
<b>REFERENCES</b>	26
<b>APPENDIX</b>	28
<b>PLAGIARISM REPORT</b>	29



## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO.</b>
Figure 3.1: Diagram of Methodology	10
Figure 3.2: Different steps of data pre-processing	11
Figure 4.1: Evaluation	14
Figure 4.2: Confusion Matrix	15

## LIST OF TABLES

<b>TABLE</b>	<b>PAGE NO.</b>
Table 2.1 Related work comparison table	8
Table 3.1 Dataset labeling	12
Table 3.2 Tokenization Table	12
Table 3.3 Parameter usages	13
Table 4.1 Accuracy table using FastText	15
Table 4.2 Accuracy Table Using Word2Vec	15
Table 4.3 Classification reprot of multinomial NB	16
Table 4.4 Classification Report of XGB classifier	17
Table 4.5 Classification Report of SVC classifier	18
Table 4.6 Classification Report of Random Forest Classifier	19
Table 4.7 Classification Report of KNN classifier	20

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

To foretell the political leanings of blog posts or public opinion polls, sentiment analysis (also called information extraction) gathers people's opinions, feelings, and attitudes on topics that are widely known. The process of getting online in Bangladesh is currently rather straightforward. Many first-time consumers prefer to peruse evaluations written about a product before committing to a purchase. One way to find out how a user feels about a given topic is to utilize sentiment detection. Tweets, reviews, comments, posts, and bulletins are all examples of textual content that can be given a positive, neutral, or negative polarity. One way to find out how a user feels about a certain topic is to utilize sentiment detection. Tweets, reviews, and comments are pieces of text that can be "polarized" into positive, neutral, or negative aspects. Books are among the most popular products sold online, reflecting a change in consumer tastes that has occurred in recent years.

As more and more individuals are able to communicate through the Internet, it becomes easier for some to spread false information. After seeing ads for bookstores online, people are wary about making purchases. Previous reviews and ratings left by customers for these products and service providers are also taken into consideration. When asked about learning a new language, most Bangladeshis would rather study Bangladeshi. Since that is my mother's advice, I don't mind. In the larger context of the Internet economy, reviews play a crucial role. Its goal is to find out, using a ratio system, whether a book got more Positive or negative reviews using machine learning algorithms. More and more academics are focusing on sentiment analysis because the field of natural language processing (NLP) relies on it. Collect 1600 reviews of Bengali books, split them down the middle, and label them as either good or negative. After that, I employ Random Forest Tree (RFT), Multinomial Naïve Bayes (MNB), Support Vector Classifier (SVC), and Stochastic Gradient Descent (SGD), among others.

## **1.2 Motivation**

"Internet bookstore" is a term that has recently gained traction in Bangladesh. There will always be a market for online bookstores as long as there are people using the Internet. I are all now physically closer to one another because of technology. Online bookstores have grown rapidly due to e-commerce. Going to brick-and-mortar bookstores is something most people would rather not do these days. They would rather have everything done for them and live as simply as they can. Online bookshops have also simplified the process of acquiring both digital and physical books for consumers. Making a purchase is as easy as a few clicks for customers. Another choice is to read reviews of books. Assessments of documents, cases, objects, or phenomena are what it is all about. Feedback on various subjects such as books, articles, buildings, sculptures, designs, food, politics, showpieces, performances among others is appreciated. This talk primarily examines literary reviews. Most shoppers glance at the review before buying. A book review gives a broad summary of a book, allowing readers to have a general understanding of its content. Nevertheless, it is a procedure that requires a significant amount of time. Reading every single comment might sometimes be tedious for a consumer. The above explanation demonstrates the necessity of addressing this matter in a manner that optimizes persons' time and enables them to complete their work. Ultimately, I have decided to employ Natural Language Processing (NLP) and machine learning techniques to address this particular problem. Algorithms lack the ability to comprehend strings directly, as is often understood. Initially, it is imperative to convert the string into a numerical representation. In this case, we employed the TFIDF technique. A Machine Learning algorithm was employed to categories each comment. We employed distinct parameters for each technique. I selected these settings based on their ability to yield the most favorable results.

## **1.3 Problem Definition**

Over the course of more than ten years, individuals have utilized the internet within the confines of their residences. The advent of the internet has had a profound impact on people of all age groups, ranging from the elderly to youngsters, and from instructors to new hires. Each individual has their own unique approach to understanding and utilizing

this technology to meet their specific needs. The internet surpasses all prior forms of media in terms of entertainment, visualization, commerce, information gathering, education, and gaming. The Internet has evolved into the most efficient and economical method for accessing the worldwide network. Various enhancements such as compelling advertising, real-time videos, efficient operations, and other components have been included. The internet has evolved into a highly effective instrument for marketing and sales. The internet has emerged as the primary platform for retail enterprises to showcase and sell their products. Individuals are increasingly growing more at ease with the act of procuring books from digital merchants. In order to save time and give buyers with the best book possible, this research is being carried out. In this research, I utilized Machine Learning and Natural Language Processing. Many difficulties have arisen because of this undertaking. Since I am addressing human emotions, data collection is a very challenging for my study. A number of online bookstores were combed through for data. Plus, I made sure to gather every single review from every single book. I had a mix of good and bad feedback. This data is what I use for the feature. I received around 28,45 comments in Bangla. My dataset contains many forms of interference, such as duplicated words, excessive punctuation, and emoticons. In the preprocessing phase, I eliminated all the irrelevant and unwanted data to ensure optimal learning of my algorithm. After performing preprocessing, I utilized the word embedding technique to transform the string into a numerical representation. After completing the task of establishing numeric formats, I employed multiple Classification Machine Learning methods due to the classification nature of my project. Two types of sentences are presented here: positive and negative. Upon completion of the training phase, I assess my progress by acquiring untrained, raw data. My technique surpasses others in the evaluation process.

#### **1.4 Research Questions**

- When collecting and working with datasets, what approaches are used?
- How well do machine learning algorithms predict which labels will be positive and which will be negative?
- Are you adept at telling good groups apart from bad ones?
- Asking, "What are the work's benefits?"

## **1.5 Research Methodology**

The following section will provide an overview of my workflow, which includes the processing of data, the processing of information, the classification of data, and the application of algorithms. The training of models and the evaluation of algorithms

## **1.6 Research Objectives**

From this work I expect that:

- Using or categorizing such classified procedures as a methodology while performing consumer analysis.
- A software application for pricing items and services should be developed using engineering tools and machine learning.
- Investigate an idea in order to back it up with proof.

## **1.7 Research Layout**

Chapter 1: This chapter includes background, purpose, issue statement, research question, methodology, and anticipated findings. Additionally, in this chapter, I detail my rationale for conducting this study.

Chapter 2: The second chapter provides a brief review of the work and a contextual examination of its history, including relevant studies and the current situation in Bangladesh.

Chapter 3: This part will demonstrate the execution of the thought process and clarify the inquiry plan by diving further into the approach or procedure.

In Chapter 4, you will get detailed instructions for putting on the proposed show, along with an exploratory outcome report.

In Chapter 5: I will describe an ethical aspect of this project.

Chapter 6: This component of the report concludes with a summary of the demonstration's performance, a comparison of precision, and an investigation.

## **1.8 Expected Outcome**

- Both good and negative feedback from customers will be handled separately.
- This will help the customer save time.
- Based on the client's choices, I will do my best to display the best book.
- To show the outcomes of each book review comment, I built a helpful web app.

## **CHAPTER 2**

### **BACKGROUND STUDY**

#### **2.1 Introduction**

The field of sentiment analysis has been the subject of a significant corpus of research, each of which has utilized a different combination of machine learning algorithms to address certain problems. In the next part, a summary of the previous efforts that were done by a variety of subject matter experts is presented.

#### **2.2 Related Works**

There were many different languages and contexts in which this subject was discussed; nowadays, nearly everything is done online; individuals voice their opinions in online forums; and the researcher magnet is often employed to gauge people's emotional states.

The authors Mittal et al. [4] developed a method for evaluating Hindi that achieved an accuracy rate of 82.89% for positive results and 76.59% for negative outcomes. In an effort to get more uniform results, they decide to evaluate emotions and increase the database size. This article details an initiative that seeks to understand Roman Urdu speakers' perspectives on a range of topics, such as politics, culinary arts, computer science, sports, and computer programming. The presentation incorporates 10,021 sentences culled from 566 distinct online conversations. Its intended purpose is to achieve two things: I will build a human-annotated corpus of Roman Urdu and then explore several sentiment analysis methods using Rule-based and N-gram (RCNN) models to do emotional analysis of the language.

Regardless of how positive or unfavorable a person's feedback was, Chowdhury et al. [5] developed a system to remove them from the Bangla language network automatically. Using unique features from 1300 col-selected data, SVM achieved 93% in its recommended strategy. One method for bringing together subjective concepts, ideas, and language is known as sentiment analysis (SA). Currently, SA poses the greatest difficulty in the field of natural language processing. Broadcasting many perspectives on one live



unit is a prevalent practice on social networking sites like Facebook. A newspaper reader offered commentary on an event's reported facts. Feedback from internet purchases is increasing in volume daily. Consequently, people's opinions and assessments greatly affect their degrees of happiness.

The way people feel about a particular subject can be explored using an aspect-based opinion assessment, which is a type of assumption analysis. This method was employed in the investigation carried out in Bangladesh by Rahman et al. [3]. Bengali estimation research is currently regarded as a prominent area of study. Data collecting, corporate dialect studies, and vocabulary building for the voice tagger are all examples of Bengali jobs that are particularly difficult to do due to a lack of resources. They intended to conduct research based on aspects to evaluate the quality of eateries in the surrounding area and to collect feedback from cricket fans. When it comes to identifying and removing pests from restaurants and other food-related environments, SVM has the highest level of validity at 71% and 77%, respectively.

Not all businesses are good at recognizing customer wants and needs, which is a major problem for online retailers. Using machine learning to distinguish between positive and negative feedback provided by potential clients, C. Chauhan et al. [7] were able to validate their assessments. Their review of the literature led them to the conclusion that Naive Bayes produced acceptable results; nevertheless, the results differed according on the setting, the methodology, and the goals.

Both the original and modified versions of the network have demonstrated remarkable performance in a number of NLP tasks, particularly for resource-rich languages like English. Unfortunately, there is a lack of study on how to use these alternatives to classification problems in Bangladesh. The government of Bangladesh has responded by creating a paradigm for multilingual text classification transformers. Alam et al. [6] developed a CNN model to assess the emotions conveyed in Bangla language. Out of 850 data points, 350 were negative and 500 were positive; the CNN obtained an accuracy of 99.87%.

In their study on emotion recognition in Bangladesh, Tuhin et al. [8] put up two approaches. Excitation, outrage, melancholy, apprehension, excitement, and sensitivity were the feelings noted. Naive Bayes uses both the segmentation technique and the topical solution. With a sample of 7,400 words from Bangladesh, the researchers utilised a topical strategy and achieved an accuracy rate of 90%. Two other articles were compared to their results; one of them had an SVM score of 93% and the other had a document frequency score of 83%. Three separate research, each examining a different facet of human emotion.

### 2.3 Comparison of related work

Table 2.1 Related work comparison table

Previous work	Accuracy
Evaluating posts on Bangla microblogs using sentiment analysis [2]	93%
Evaluation of Bangla Aspect-Based Sentiment Analysis Databases and Their Foundation. [3]	77%
Applying a convolutional neural network to interpret the tone of Bangla speech. [6]	99.87%
Using supervised learning techniques, a system was constructed to analyse sentiment in Bangla text.[8]	75%
Analysing the tone of Google Play reviews written in Bangla using machine learning techniques.[7]	76.48%
Applying machine learning techniques to evaluate the sentiment of movie reviews in the Bengali language.[1]	88.90%

After analyzing the conversation regarding table 2.1, I determined that there was a lack of substantial book review engagement in Bangladesh. Upon comparing the two studies, it becomes evident that my model possesses a larger dataset, exhibits higher accuracy, and has demonstrated proficiency across several domains. No word embedding system has

been previously utilized in book review-related research. The stuff I produce could potentially be utilized on an online platform.

## **2.4 Research Summary**

This study highlights the breadth of research on emotional analytics since it was conducted by different research firms. The outcomes of my investigation are significant and definitive. Each area works to encourage resourcefulness despite limited resources by providing advice on how to get a lot of things done in a day.

## **2.5 Challenges**

Data set strategy for further processing is the most difficult part of the job. In preparation for my work or further processing, I preprocessed the dataset using highly effective machine learning methods. Obtaining adequate cash resources or job prospects is another obstacle in Bangladesh. Bringing the machine learning paradigm in line with internet standards is a challenging undertaking.

# CHAPTER 3

## RESEARCH METHODOLOGY

### 3.1 Introduction

The working methodology consists of five stages: data collecting, algorithm analysis, algorithm execution, validation, and web deployment. Figure 3.1 displays the diagram illustrating my work.

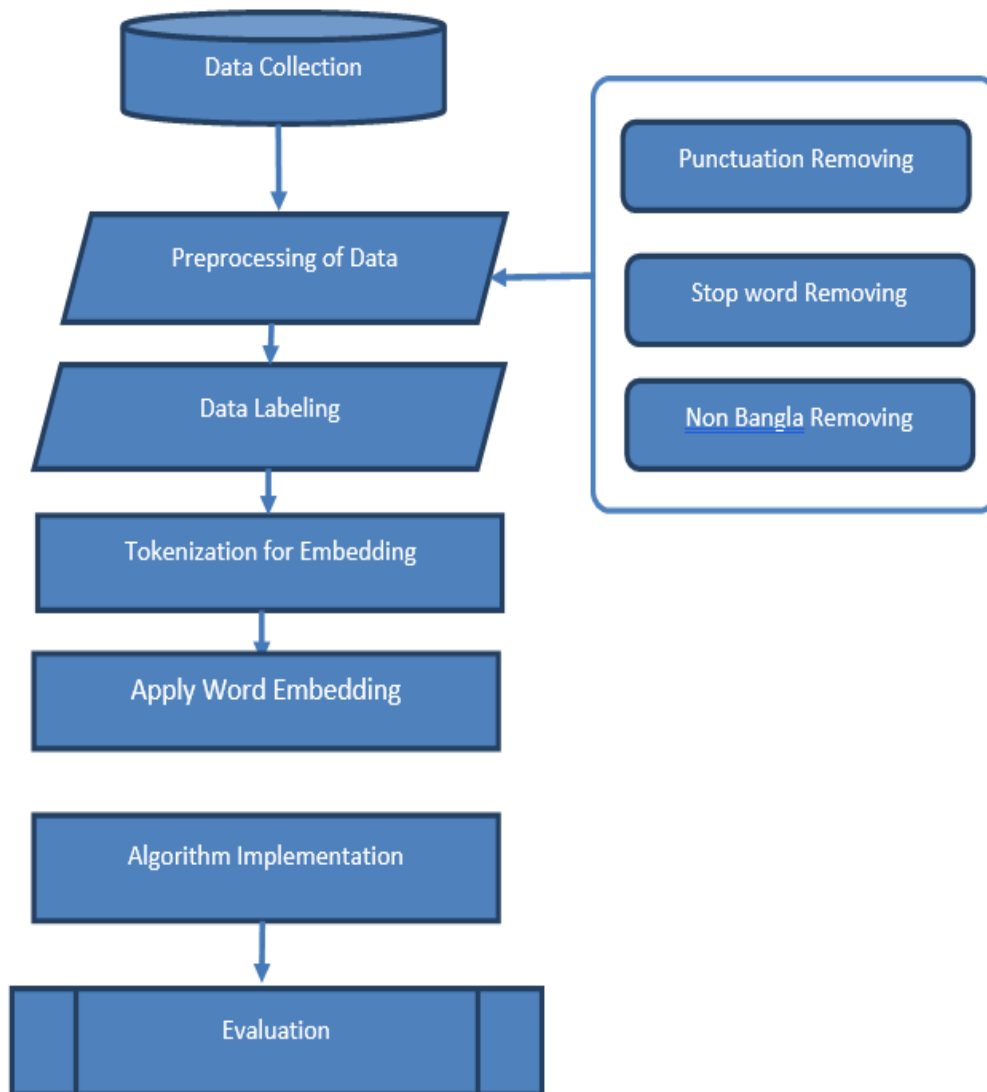


Figure 3.1: Diagram of Methodology

### 3.2 Data Collection

Gathering pertinent data is the first step in every research project. An essential piece of data known as a book review is crucial to a book's success. Additionally, I need to make sure that the places I'm getting my information from are reliable. Reviewers' feedback served as the foundation for my study. Several online book stores and review sites provided the data that was later aggregated for use on Facebook. I was able to complete the feedback collection in Bangla as that was the language we were provided. The inquiry gathered a total of 28,455 data points.

### 3.3 Data Pre-Processing

Data preprocessing is an approach to data mining that involves cleaning up raw data in order to make it more usable. Acquiring knowledge requires meticulously gathering information. The KDD framework is used to develop the function. Data massaging, removal, weighting, and Same poling are the four most significant pre-processing techniques, according to Kamiran et al. [9]. Creating easily available data sets was my main focus while working with data messaging systems. Among my preparation steps, removing punctuation, stop words, and non-Bangla words is the most crucial. To make the Bangla stop at this level easier to understand, I have removed all extraneous terms and details. My improved contribution will be clearly shown among the several procedures that must be carried out. You can see all the work we put into getting the data ready in Figure 3.2.

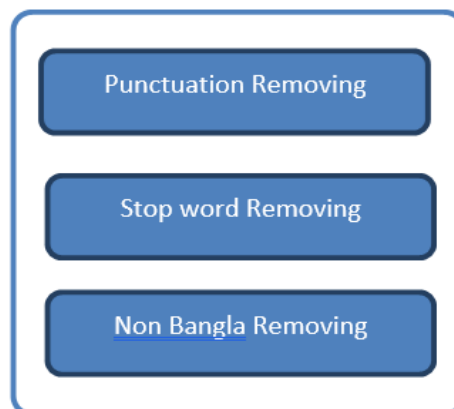


Figure 3.2: Different steps of data pre-processing

### 3.4 Dataset Labeling

I am presented with information that was categorized into two distinct groups: positive and negative. The user's emotions are considered during the course design process. If the novel's analysis is robust, this line will receive a high score. Unfavorable ratings can be classified into multiple categories. Table 3.1 depicts the dataset's labelling. By examining this table, it is evident that there are 1133 instances of negative data and 1712 instances of positive data. The training dataset has around 2845 instances.

Table 3.1 Dataset labeling

Target	amount
Negative Review	1133
Positive Review	1712

### 3.5 Tokenization

Splitting flag phrases which may be word or signal combinations is described in detail by Pinto et al. [10] as an application of tokenization. In my database, you may find several expressions. For this purpose, The use of a phrase mark has replaced the use of a word label in my work. The tokenization process is also crucial. Tokenization is the process that breaks down my sentence into its individual words. In Table 3.1, you can see the process of tokenization.

Table 3.2 Tokenization Table

Raw Data	Type	Tokenized data
বইটি আসলেই চমৎকার	Positive	'বইটি', 'আসলেই', 'চমৎকার'
ফাউল একটা বই।	Negative	'ফাউল', 'একটা', 'বই'
ভালো বই ছোট বড় সবার পড়া উচিত	Positive	'ভালো', 'বই', 'ছোট', 'বড়', 'সবার', 'পড়া', 'উচিত'

### 3.6 Algorithm Implementation

While this section was open, I covered the steps that must be taken to implement the method. Prior to beginning this phase, the previous one must be completed. In order to proceed to this level, it is essential to construct the dataset that is required. Due to the nature of my job being done in a classification manner, I possess five distinct methods for organizing the data. I use a variety of methods as classifiers, including the XGBoost method, Multinomial Naive Bayes algorithm, SVC algorithm, KNN algorithm, and many more. Table 3.2 shows the values that, when applied to each method, will allow for the highest level of accuracy.

Table 3.3 Parameter usages

Algorithms	Details
Multinomial Naive Bayes	n_informative=3, n_redundant=0, random_state=1, shuffle=True
XGBoost	random_state=0
SVC	kernel='rbf'
Random Forest	n_estimators=80
KNN	random_state = 42

# CHAPTER 4

## RESULT ANALYSIS

### 4.1 Introduction

Results from tests and other forms of empirical data are the focus of this branch of analytical study. When researching a topic, what is the first step in analysing the results? In order to show the results without any subjective evaluation or interpretation, the consequences section should be conditional. You can also find the instructions under the part that is specifically for informative publications. The results are shown, and the test is confirmed. Furthermore, out of five methods, I selected the best one after extensive analysis. Data were calculated using the following criteria: recall, precision, and accuracy, as well as f1.

### 4.2 Experimental Result

The data use rate I employed ranged from 30% to 70% for the test data. Similarly, when 70% of the data is used for testing, the training size is reduced to 30%. I utilize this methodology for the machine learning algorithm to determine the percentage that achieves the highest accuracy. According to the table, the Support Vector Classifier achieved the highest accuracy, reaching approximately 85.92%.

Table 4.1 Accuracy table using FastText

Test data usage rate		30%	40%	50%	60%	70%
Algorithms Accuracy	MNB	76.29	77.73	77.80	77.04	76.35
	KNN	82.04	80.46	80.20	80.56	78.86
	RF	82.86	82.83	81.18	81.74	80.83
	XGB	83.10	84.15	82.80	82.62	80.93
	SVC	85.68	85.92	84.50	84.44	83.44

This table was generated by using the word2vec algorithm. from this table I| can see that the Support vector classifier achieved the highest accuracy at about 84.78% using 50% test data



<b>Test data usage rate</b>		<b>30%</b>	<b>40%</b>	<b>50%</b>	<b>60%</b>	<b>70%</b>
Algorithms Accuracy	MNB	76.88	76.67	77.59	77.33	76.65
	KNN	81.34	80.72	79.35	79.51	77.05
	RF	82.16	80.90	81.89	81.39	81.50
	XGB	81.46	81.43	81.25	81.44	81.43
	SVC	83.92	83.63	84.78	85.03	83.95

Table 4.2 Accuracy Table Using Word2Vec

If I compared both table I can see that Word2Vec and FastText both of them performed similarly but FastText produced better accuracy than Word2Vec. So I have decided to use FastText as the final Word embedding model.

Table 4.3 Different Score Matrix

<b>Score Matrix</b>	<b>Algorithms</b>				
	<b>MNB</b>	<b>KNN</b>	<b>RF</b>	<b>XGB</b>	<b>SVC</b>
F1 Score	0.89	0.87	0.89	0.88	0.92
Recall	0.86	0.90	0.93	0.90	0.93
Precision	0.93	0.84	0.84	0.87	0.90
Specificity	0.90	0.76	0.76	0.81	0.86

The rating Matrix is displayed in table 4.3. I have completed only 30% of the scoring matrix. Various criteria have been employed to assess the accuracy of the precision desk, which is a reliable indicator of precision based on true positives and true negatives. These criteria include extremely poor, falsely excellent, genuinely high quality, and inaccurately horrible. From this table, it is evident that the SVC algorithm produced the greatest score among all the algorithms.

#### 4.2.1 Multinomial Naïve Bayes Algorithm

Bayesian dominance in distinctive dialect handling (NLP) is critically addressed by the Multinomial Naïve Bayes run the show set. The software calculates the value of a printed item's tag using Bayes' hypothesis; this may be anything from a letter or a daily newspaper piece. Returns the tag that poses the greatest threat based on the likelihood of each tag for a specific test. All of the calculations that make up the Credulous Bayes classification share the

trait of isolating each classed characteristic from all other characteristics. Table 4.4 shows the results of Multinomial Naïve Bayes's classification. The table shows that the naïve bayes algorithm achieved a maximum score of 93 in the precision section for favorable reviews. The algorithm achieves an overall accuracy of 88% when tested with real data.

Table 4.4 Classification reprot of multinomial NB

	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Support</b>
0	0.83	0.90	0.86	21
1	0.93	0.86	0.89	29
Accuracy			0.88	50
Macro avg	0.88	0.88	0.88	50
Weighted avg	0.88	0.88	0.88	50

## 4.2.2 XGB Classifier

Using the neighbor information shown in Figure 4.2, an XGB, a comprehension algorithm, correctly isolates each node. As a result, stronger trees can be driven using variable splitting. It has been demonstrated that trees can be somewhat shaped with relatively few interactions. The "high dispersion" tones are just effects, which is a major drawback. When the tree's forecasts are too optimistic, it might lead to overfitting and big disparities. When used to big datasets, selection trees yield superior outcomes. the eleventh It is possible to construct a stronger tree by modifying the department factors. Table 4.5 displays the XGB classifier's classification results. For favorable results in the recall portion, this method can attain a maximum score of 90. The test data has an overall accuracy of 88%.

Table 4.5 Classification Report of XGB classifier

	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Support</b>
0	0.85	0.81	0.83	21
1	0.87	0.90	0.88	29
Accuracy			0.86	50
Macro avg	0.86	0.85	0.86	50
Weighted avg	0.86	0.86	0.86	50

### 4.2.3 SVC

Assistance One kind of supervised learning method used for outlier detection, regression, and classification is the vector classifier. Machine learning is rife with tasks like these. By evaluating a huge number of photos, these technologies can discover cancerous cells. Similarly, a well-equipped regression model can be used to forecast future travel patterns. With its roots in Support Vector Classification (SVC), Support Vector Regression (SVR) is an extension of Support Vector Machines (SVM) that has the potential to address certain machine learning challenges. These optimized mathematical equations are the be-all and end-all, since they seek to provide the most exact and fastest outcome conceivable. The decision boundary of a Support Vector Machine (SVM) maximizes the distance between the nearest data points in the training set, giving it a distinct edge over other classification techniques. A maximum margin classifier, or hyperplane with the biggest margin, is the decision boundary that emerges from applying Support Vector Machines (SVMs). Table 4.6 shows the svc classifier's report on classification. In the recall portion, the algorithm got a maximum score of 93. With the test data, the algorithm has achieved a total accuracy of 90%.

Table 4.6 Classification Report of SVC Classifier

	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Support</b>
0	0.90	0.86	0.88	21
1	0.90	0.93	0.92	29
Accuracy			0.90	50
Macro avg	0.90	0.89	0.90	50
Weighted avg	0.90	0.90	0.90	50

#### 4.2.4 Random Forest

Random Forest is a versatile and user-friendly collection of guidelines that yield excellent outcomes even on the most extensive cases without the need for hyper parameters. Additionally, it is widely employed algorithmically owing to its straightforwardness and adaptability. This document elucidates the functioning of RFAI, its distinguishing features compared to other algorithms, and provides instructions on its use. Construct a "thicket" of trees to facilitate decision-making, as is commonly instructed in belaying. The primary factor underlying field approaches is the enhancement of the final outcome through the integration of deployment models. Random forests can be utilized as an alternative to categorical regression. The Random Forest algorithm attained an accuracy of 86% for the test data in the classification project. The recall part has obtained a maximum score of approximately 93%, however, the total accuracy is not optimal.

Table 4.7 Classification Report of Random Forest Classifier

	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Support</b>
0	0.89	0.76	0.82	21
1	0.84	0.93	0.89	29
Accuracy			0.86	50
Macro avg	0.87	0.85	0.85	50
Weighted avg	0.86	0.86	0.86	50

#### 4.2.5 KNN

Supervised learning is the method that is utilized in the K-Nearest Neighbour methodology, which is an important machine learning technology. Inferences can be made by the system regarding When a new piece of information comes in, it gets compared to things we already know. The system then groups the new info with the items that are most like it. This clever system remembers everything available. It spots and tells apart fresh bits of info by measuring their similarities. The K-NN strategy lets us quickly sort fresh information into the right group. Even though we can use this method for different tasks, like prediction and grouping, it's often used for grouping. In light of the fact that it is a non-parametric approach, the K-NN method does not make any assumptions on the data that is being discussed. A recall accuracy of approximately 90 percent was achieved by the KNN algorithm for positive cases, making it the method with the highest recall accuracy among all those that were examined. The results of the classification performed by the k-nearest neighbours (KNN) algorithm are presented in Table 4.5. When compared to the performance of all the other algorithms, this approach had the worst performance.

Table 4.8 Classification Report of KNN classifier

	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Support</b>
0	0.84	0.76	0.80	21
1	0.84	0.90	0.87	29
Accuracy			0.84	50
Macro avg	0.84	0.88	0.83	50
Weighted avg	0.83	0.84	0.84	50

### 4.3 Evaluation

In the last assessment of the model, I gathered exactly 50 comments, both good and bad ones, that my model has never encountered before. Out of the 28 favorable remarks, a total of 25 were accurately predicted by my model. Similarly, from the 22 critical views, 16 of them were correctly anticipated.

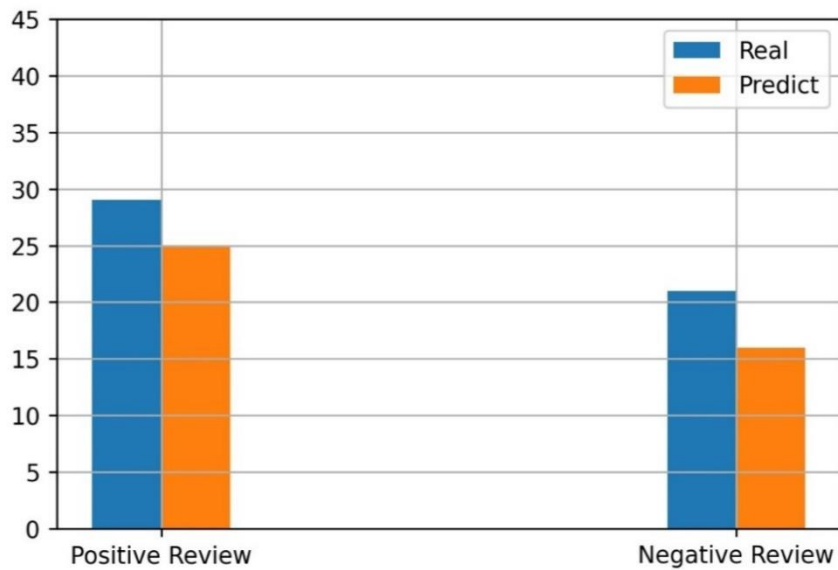


Figure 4.1: Evaluation

I employed the Confusion Matrix to ascertain the comprehensive outcomes. Figure 4.2 represents the confusion matrix evaluation procedure by using test dataset of the best model. During the review procedure, I determined that 82.00% of the test results exhibited accuracy. my technique allows for the utilization of both apparent and concealed data. The positive memory accounts for 86.21% of the total, whereas the negative recall represents 76.19%. According to this calculation, I may infer that my system is more efficient in predicting Positive reviews.



Figure 4.2: Confusion Matrix

$$\text{Accuracy} = \frac{25+16}{25+16+5+4} = 0.8203 * 100 = 82.0\%$$

$$\text{Error Rate} = 1 - 0.82 = 0.17 * 100 = 18\%$$

$$\text{Positive Recall: } 25 / (25+4) = 0.8621 * 100 = 86.21\%$$

$$\text{Negative Recall: } 16 / (16+5) = 0.7619 * 100 = 76.19\%$$

## CHAPTER 5

### IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

#### 5.1 Impact on Society

There are two societal effects of this study. With better sentiment analysis in Bengali book clothing reviews, businesses may leverage customer feedback to make better decisions. This improves customer happiness and enables the tailoring of marketing techniques for Bengali-speaking customers, so making book e-commerce more personalized and effective. Furthermore, this project will enable individuals to optimize their time. When a visitor visits to buy a book, there is no need to read the entire review of the book.

#### 5.2 Impact on the Environment

The work primarily focuses on the linguistic and computational aspects, with little implications in the physical world. However, there are potential secondary environmental benefits that can be obtained. Enhancing sentiment analysis in the realm of book e-commerce enables enterprises to efficiently control their marketing plans, hence reducing resource consumption and decreasing waste resulting from less focused advertising endeavors.

#### 5.3 Ethical Expect

- To prevent distortion and stereotyping in Bengali phrases, one must ethically analyse the intricacies of language and cultural nuances.
- Machine learning algorithms must avoid perpetuating prejudices or engaging in discriminatory practices towards specific communities.
- The process of collecting data must adhere to ethical guidelines and acquire informed consent from individuals.
- Maintaining truth and objectivity in conveying study findings necessitates avoiding sensationalism.



#### **5.4 Sustainability Plan**

This research's plan aims for sustainability. It promotes saving resources. It works for long-term impact. It helps make progress in the field. The research technique will also include ethical concerns to ensure the proper handling of data and respect ideals of equity and openness. To ensure the study findings remain relevant and applicable, I will actively engage with enterprises and language groups on a regular basis. The sustainability plan aims to expand and enrich the academic and societal impact of the research.

## CHAPTER 6

### SUMMARY, CONCLUSION AND FUTURE WORK

#### 6.1 Summary of the Study

More people are shopping at online bookstores as more folks are using the internet. Thanks to tech advances and new ideas, folks all over the world can now connect. The advent of E-trade has facilitated the exponential expansion of an online bookstore. Contemporary individuals want convenience and a genuine lifestyle, rather than investing time in visiting physical bookstores. The field of system learning has garnered some interest, but there has been a lack of comprehensive research conducted in the context of Bangladesh. Despite the prevalence of the term "vision styles" in computer education, Bangladeshi literature remains oblivious of this concept. This inquiry has lately been conducted due to those assignments causing a substantial alteration in my system's existence. Regardless, the intricacies of Bangladeshi financial matters may be of minimal importance.

#### 6.2 Conclusion

The specialized dataset is crucial to SA because of the ever-increasing number of Internet users. In order to sort Bengali book reviews into good and bad sentiments, this research introduces a sentiment classification system that makes use of many feature extraction techniques. Over 30 different book types, I sifted through 28,45 customer reviews. While analyzing book reviews for relevant terms, I gathered book data and summaries. Cost, transportation, quality, aesthetics, and contentment are the five key criteria that I have found to be present and impactful in all client viewpoints. Afterward, I developed a system that is exceptionally proficient using the most recent version. I aim for an accuracy rate of about 85% when it comes to pricing for e-book reviews. It has been established that the RF algorithm is accurate. Because of its remarkable performance, RF outperformed other popular algorithms in terms of accuracy, including KNN, decision Tree, SVM, XGBoost, and Random Forest. Booksellers and readers alike can evaluate

works for their merit or lack thereof; discerning readers can also tell which novels feature believable or pathetic protagonists. Bookstore owners gain from this strategy. Over the course of this period, there will invariably be an interchange between library and e-book purchaser research.

### **6.3 Recommendations**

There are several excellent choices available for this:

- To get more reliable results by improving the accuracy of data collecting.
- There is a striking lack of content in this booklet.
- Deep Learning should be used.

### **6.4 Future Work**

Here is the projected trajectory for the advancement of this project: my objective in Bangladesh is to investigate the sensation of a corrosive proclamation.

- I will create a sophisticated framework to execute this concept.
- My objective is to utilize an internet-based API to obtain specific analytical insights.
- In the future, it is possible to create an advanced technology that utilizes deep learning techniques to enhance intelligence.

## REFERENCE

- [1] R. R. Chowdhury, M. Shahadat Hossain, S. Hossain and K. Andersson, "Analyzing Sentiment of Movie Reviews in Bangla by Applying Machine Learning Techniques," 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), 2019, pp. 1-6, doi: 10.1109/ICBSLP47725.2019.201483.
- [2] Fang, X., Zhan, J. Sentiment analysis using product review data. *Journal of Big Data* 2, 5 (2015). <https://doi.org/10.1186/s40537-015-0015-2>.
- [3] 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.
- [4] 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.
- [5] S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts," 2014 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, Bangladesh, 2014, pp. 1-6, doi: 10.1109/ICIEV.2014.6850712
- [6] M. H. Alam, M. Rahman, and M. A. K. Azad, "Sentiment analysis for Bangla sentences using convolutional neural network," 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2017, pp. 1-6, doi: 10.1109/ICCITECHN.2017.8281840.
- [7] C. Chauhan and S. Sehgal, "Sentiment analysis on product reviews," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 26-31, doi: 10.1109/CCAA.2017.8229825.
- [8] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter and A. K. Das, "An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singa-pore, 2019, pp. 360-364, doi: 10.1109/CCOMS.2019.8821658.
- [9] Kamiran, F., Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl Inf Syst* 33, 1–33 (2012). <https://doi.org/10.1007/s10115-011-0463-8>
- [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 available at <<https://www.javatpoint.com/logistic-regression-in-machine-learning>> last accessed on 4-08-2021 at 11AM.

## **APPENDIX**

The initial objective was to delineate the protocols for the analysis, which posed several challenges. Moreover, there has been no advancement in this domain before this. Certainly. It deviated from your typical work. I was unable to locate an individual who could provide us with substantial assistance. Another obstacle I encountered was the process of gathering data, which presented a significant challenge for my team. I have initiated the process of gathering data manually. Moreover, categorizing the different jobs is a challenging endeavor.

# PLAGIARISM REPORT

## book review

### ORIGINALITY REPORT

<b>11</b> %	<b>11</b> %	<b>2</b> %	<b>6</b> %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

### PRIMARY SOURCES

<b>1</b>	<a href="https://dspace.daffodilvarsity.edu.bd:8080">dspace.daffodilvarsity.edu.bd:8080</a> Internet Source	<b>9</b> %
<b>2</b>	Submitted to Daffodil International University Student Paper	<b>1</b> %
<b>3</b>	Submitted to Higher Education Commission Pakistan Student Paper	<b>&lt;1</b> %
<b>4</b>	<a href="https://eprints.soton.ac.uk">eprints.soton.ac.uk</a> Internet Source	<b>&lt;1</b> %
<b>5</b>	<a href="http://www.aapor.org">www.aapor.org</a> Internet Source	<b>&lt;1</b> %
<b>6</b>	Merve Veziroğlu, Erkan Eziroğlu, İhsan Ömür Bucak. "Chapter 5 PERFORMANCE COMPARISON BETWEEN NAIVE BAYES AND MACHINE LEARNING ALGORITHMS FOR NEWS CLASSIFICATION", IntechOpen, 2024 Publication	<b>&lt;1</b> %