

**SENTIMENT ANALYSIS OF BOOK REVIEW IN BANGLA USING NLP AND
MACHINE LEARNING**

BY

Raihan khan

ID: 181-15-11158

Abdullah Aas Suhaeel

ID: 181-15-11143

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

Supervised By

Mst. Eshita Khatun

Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

Mr. Md. Sadekur Rahman

Assistant Professor

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

5 JANUARY 2022

APPROVAL

This project titled “BOOK REVIEW SENTIMENT IN BANGLA LANGUAGE USING NLP AND MACHINE LEARNING”, submitted by **Raihan khan** and **Abdullah Aas Suhaeel** 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 (BSc) and approved as to its style and contents. The presentation has been held in 5th January 2022.

BOARD OF EXAMINERS



Chairman

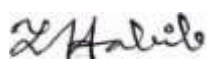
Dr. Sheak Rashed Haider Noori (SRH)

Associate Professor and Associate Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

Md. Tarek Habib (MTH)

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

Md. Reduanul Haque (MRH)

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



External Examiner

Dr. Mohammad Shorif Uddin

Professor

Department of Computer Science and Engineering

Jahangirnagar University

DECLARATION

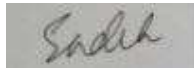
We hereby declare that this thesis has been done by us under the supervision of **Mst. Eshita Khatun**, Lecturer, **Department of CSE**, and co-supervision of **Mr. Md. Sadekur Rahman**, Assistant Professor, of **Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Mst. Eshita Khatun
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Raihan Khan
Department of CSE
Daffodil International University



Abdullah Aas Suhaeel
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 **Mst. Eshita Khatun**, 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, scholarly guidance, continual encouragement, vigorous motivation, direct and fair supervision, constructive criticism, valuable advice, great endurance during reading many inferior drafts and correcting the work to make it unique paves 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 CSE department of Daffodil International University.

We would like to express thankfulness to the 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 review 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

Sentiment polarity detection has recently piqued the interest of NLP researchers, owing to the resulting in positive of consumer comments or ratings on the internet. Due to the continued expansion of e-commerce sites, the rate of purchase of various products has grown. For example, people's interest in books is fast growing. Online marketing and e-commerce companies were already prospering in Bangladesh during this era of internet technology. For example, product reviews on the Internet have become an essential source of information for customers making purchasing decisions. Because there are sometimes too many reviews for consumers to read, figuring out how to automatically classify and determine sentiment from them has become a major research challenge.

Books are said to be a person's best friend. Books are crucial in every human's life because they provide information of the outside world, improve their reading, writing, and speaking abilities, and improve memory and intellect. Even just a few years ago, individuals in Bangladesh had to travel to the library in person to get books. Many of the benefits are straightforward, and the internet bookshop has a disadvantage in that the reader is unfamiliar with the books or with the book store itself. To avoid this, book readers prefer to rely on reviews and ratings. Our goal is to assess Bangladeshi language reviews and give accurate information about books and online bookstores so that book lovers may buy the right books to read and find better online bookstores. In this paper we show how to extract sentiment polarity (positive or negative) from Bengali book reviews using machine learning and Natural Language (NLP). We used five classification methods : Adaboost ,Decision Tree (DT), Random Forest Tree (RFT), Support Vector Machine (SVM), and lightGbm algorithm.

TABLE OF CONTENTS

CONTENTS	PAGE
Acknowledgements	iv
Abstract	v
List of Figure	viii
List of Table	ix
CHAPTER	

CHAPTER 1: INTRODUCTION	PAGE NO.
	1-5
1.1 Introduction	1
1.2 Motivation	2
1.3 Problem Definition	2
1.4 Research Questions	3
1.5 Research Methodology	4
1.6 Research Objective	4
1.7 Report Layout	4
1.8 Expected Outcome	5
CHAPTER 2: BACKGROUND	6-9
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

CHAPTER 3: RESEARCH METHODOLOGY	10-15
3.1 Introduction	10
3.2 Data collection	11
3.3 Data Pre Processing	11
3.4 Classification	11
3.5 Tokenization	12
3.6 Algorithm Implementation	13
3.7 Evaluation	14
CHAPTER 4: RESULT ANALYSIS	16-22
4.1 Introduction	16
4.2 Experimental Result	16
4.2.1 AdaBoost	17
4.2.2 Decision Tree	18
4.2.3 SVM	19
4.2.4 Random Forest	20
4.2.4 LightGBM	21
CHAPTER 5: SUMMARY, CONCLUSION AND FUTURE WORK	23-24
5.1 Summary of the Research	23
5.2 Conclusion	23
5.3 Recommendation	24
5.4 Future Work	24
REFERENCES	25
APPENDIX	27
PLAGIARISM REPORT	28

LIST OF FIGURES

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

LIST OF TABLE

TABLE	PAGE NO.
Table 2.1 Comparison table of related work	8
Table 3.2 Tokenization Table	12
Table 3.3 Parameter Usages	13
Table 4.1 Accuracy Table	16
Table 4.2 Different Score Matrix	17

CHAPTER 1

INTRODUCTION

1.1 Introduction

Sentiment analysis (SA), also known as information extraction [1], is a branch of research that extracts people's feelings, attitudes, and emotions in order to forecast polarity in public opinion or textual data from microblogging sites [2] on a well-publicized issue. In Bangladesh, getting on the internet nowadays is quite straightforward. New customers like to read previously published product reviews before deciding whether or not to purchase a certain item. Sentiment detection is a technique for estimating a user's point of view on a given topic. It assigns a positive, neutral, or negative polarity to text material (in the form of tweets, reviews, comments, postings, or bulletins). Sentiment detection is a method for detecting a user's point of view on a given topic. It assigns a positive, neutral, or negative polarity to text material (in the form of tweets, reviews, comments. People's shopping habits have changed dramatically in recent years as a result of online/e-commerce sites, and the book is one of the most popular online items.

Online bookstores such as Rokomari, Boibazar, Boikhata, eBoighar, Daraz, Bookshopbd, Boi-BoiBoi and Backpack are the most well-known in Bangladesh. When technology advancement brings people closer together, it becomes simpler to deceive or lie on the Internet. People are hesitant to buy books after seeing adverts for bookstores on the internet. They also pay attention to past consumers' comments and evaluations on these items and service providers. People in Bangladesh prefer Bangladesh as a study language more frequently than not. They don't mind because it's their mother's guidance. In the broad internet economy, reviews are quite essential. Using the Machine Learning Algorithm, we want to assess and identify whether a certain book has received negative or favorable comments on a ratio basis. Since it is a major characteristic of NLP, sentiment analysis has attracted more researchers than ever before. Create a dataset of 5500 Bengali book reviews that are categorized into positive and negative attitudes. Then We use Adaboost ,Decision Tree (DT), Random Forest Tree (RFT), Support Vector Machine (SVM), and lightGbm algorithm.

1.2 Motivation

In Bangladesh, the term "online book store" has become a buzzword. Customers of online book stores are growing in tandem with the growing number of internet users. Science and technology have brought the entire globe closer together. E-commerce has aided the quick expansion of an online bookstore. People nowadays do not want to waste time traveling to a real bookshop; instead, they want convenience in their lives, and they want to live as simply as possible. Furthermore, an online book store has made it easier for clients to obtain both e-books and physical books. To place an order, a consumer only has to make a few clicks. Another method is to do a book review, which is a critical evaluation of a document, case, entity, or phenomenon. Reviews might include books, articles, whole genres or sectors, architecture, sculpture, design, restaurants, policy, exhibits, performances, and a variety of other things. The focus of this lecture is on book reviews. The majority of buyers then read the book's review before purchasing it. A book review is an effective technique to gain a general impression of a book. However, it is a time-consuming procedure. And it might be tedious for a buyer to read each and every criticism. We can see from the preceding discussion that something has to be done to tackle this problem in a way that saves individuals time while also allowing them to get their work done. Finally, we've chosen to use NLP and machine learning to tackle this challenge. Algorithms, as we all know, do not understand strings directly. We must first convert the string to numeric representation. We utilized the TFIDF algorithm in this situation. We utilized a Machine Learning system to classify each remark. We utilized different parameters for each algorithm. And we chose these parameters since they generated the best results.

1.3 Problem Definition

The usage of the internet in each individual home has been widespread for more than a decade. The internet revolution has impacted people of all ages, from seniors to youngsters; from veterans to trainees, everyone has their own manner of learning the approach and applying it to their own needs. From entertainment to imagination, from buying to researching, from education to gaming, the internet outperformed all previous forms of

media. The Internet has become the most convenient and cost-effective means to connect to the global network. Attractive advertising, live videos, streamlined usages, and other features were introduced. The internet has become a useful tool for promoting and selling goods. For retail enterprises, the internet has become the new catalog for product sales. People are increasingly extremely comfortable purchasing books from online book websites. This research is introduced in order to save time and provide the greatest book for customers. We employed Natural Language Processing and Machine Learning in this study. There are several issues that have arisen as a result of this effort. Because we are working with human emotions, data collecting is a particularly delicate undertaking for our study. We gathered data by visiting several book-selling websites. And I gathered every single comment from every single book. We gathered both negative and favorable feedback. This is the data that we use as a feature. Approximately 5500 Bangla comments were collected. This is our raw data, which contains a lot of noise, such as double words, additional punctuation marks, and other emoji. We deleted all of this noise during the preprocessing step so that our algorithm could learn properly. After the preprocessing, we utilized the TFIDF technique to transform the string to numeric representation. Because our work is classification-based, we applied several Classification Machine Learning algorithms after the competition of creating numeric formats. Every sentence is divided into two groups, one positive and the other negative. We assess our work when the training state is accomplished by retrieving original data that has not been trained. Our method performs better in the evaluation step. Each stage was represented by a separate graph.

1.4 Research Questions

- What methods will be used to collect and prepare the dataset?
- Can the machine learning method properly predict Positive and Negative classes?
- Is it possible to define the positive and negative groups correctly?
- Is it possible to put in place online?
- How will this work benefit the people?

1.5 Research Methodology

This part will go through our workflow, which includes data processing, information processing, data categorization, and algorithm implementation. Algorithm evaluation, model training.

1.6 Research Objectives

- To create a model capable of accurately detecting positive and negative comments.
- To anatomize consumer analysis by applying or categorizing those classification methods.
- Using engineering tools and machine learning, create a software application to view pricing.
- Conduct research to illustrate a certain scientific concept.

1.7 Research Layout

Chapter 1: will cover the following topics: introduction, motivation, problem definition, research question, research methodology, and our project's predicted conclusion. We also explain why we opted to perform this study in this chapter.

Chapter 2: The history of this research, as well as related work and present state from the perspective of Bangladesh, will be discussed in Chapter 2. It includes a contextual analysis as well as a brief synopsis of the work.

Chapter 3: will describe research methodology.

Chapter 4: will discuss performance of the proposed model.

Chapter 5: shows result comparison and analysis.

Chapter 6: It summarizes the findings of this study. This section shows how effective the model is. The model component and web performance implementation are also included in this part.

Chapter 7: here all the references we used for this research.

1.8 Expected Outcome

- We will distinguish negative and positive opinion of client remark.
- We will save client time.
- We will attempt to show best book dependent on client decision..
- We developed a powerful online application that displays the outcome of any book review opinion.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

Different machine learning algorithms for prediction have been researched. Prediction is one of the most often used applications of Machine Learning. Many studies on sentiment analysis have been carried out. These studies targeted specific issues and utilized a range of machine learning approaches to solve them. This chapter highlights the actions that several experts in the preceding region successfully carried out.

2.2 Related Works

Almost everything nowadays is web-based. On the internet, people express their opinions. The researcher magnet is frequently used to detect people's emotions. This topic was presented in a variety of areas and languages.

Mittal et al. [4] proposed a technique for analyzing Hindi that yields 82.89 and 76.59 percent positive and negative validity, respectively. They opted to assess emotions and increase the database's coverage in order to improve the database's consistency. This article describes an educational program that analyzes the Roman Urdu people's emotions through the genres of sports, software, cuisine and recipes, theatre, and politics. It contains 10,021 sentences culled from 566 internet discussions. The goals of this project are twofold: (1) developing a human-annotated corpus for emotional analysis in Roman Urdu; and (2) evaluating feeling analysis approaches based on Rule-based, N-gram (RCNN) models.

Chowdhury et.al.[5] suggested a device that would automatically delete people from the network, whether negatively or positively, in the Bangla language. SVM performed 93% with unique characteristics from 1300 col-selected data in its proposed process. Sentiment Analysis (SA) is a mixture of opinions, feelings and textual subjectivities. SA is the most difficult natural language processing job at present. Social networking sites such as Facebook are often used to share views on a single life entity. Newspaper published news about a specific incident, and in news comments the user shared his input. The amount of

feedback received from online items is increasing every day. As a result, reviews and opinions play an important part in determining people's levels of satisfaction. Opinion mining can help you find out more information.

Aspect-based Sentiment Analysis is a form of sentimental analysis that examines the sentiments surrounding a certain issue. Rahman et al. [3] did Bangladesh research using this method. Sentiment analysis is progressing in Bengali and is now considered a main research focus. In Bengali, corporate language analysis, lexicon as part of the speech tagger, and other tasks are challenging due to the scarcity of resources such as a well annotated data collection. Their focus was on a restaurant review and the application of aspect-based research to get cricket opinions. SVM has the maximum validity for extracting and discovering polarity in insects and restaurants, respectively, with 71 percent and 77 percent.

In online shopping, understanding client wants is critical, but firms may not be as informed as they should be. In order to validate their ratings, C. Chauhan et al. [7] used machine learning algorithms to distinguish between negative and positive comments from potential customers. They examined a variety of publications and found that Nave Bayes gave positive results, albeit the results varied depending on the environment, strategy, and goals.

Modeling constructed with this sort of network and its variations recently proved excellent performance in various downstream natural language processing applications, particularly in resource-rich languages like English. However, these models have not been fully studied for Bangladesh's categorization challenges. In Bangladesh, they fine-tune the multilingual text classification transformer model. In order to describe the text-based sensations given by analyzing in Bangla, Alam et al. [6] designed a model of Convolution Neural Network (CNN). CNN achieves 99.87 percent accuracy with 850 data points, 350 of which were negative and 500 positive.

Tuhin et al. [8] proposed two approaches for classifying and identifying different types of emotion from all around Bangladesh. These were ecstatic, enraged, sorrowful, terrified, enthusiastic, and sensitive. In Nave Bayes, the topical solution and the method of grouping are strategies. A data collection of 7400 Bangladesh phrases was employed, with a topical

approach providing 90% accuracy. They then compared their article to two others, both of which scored 93 percent for SVM and 83 percent for document frequency. The emotional parameter in each of the three articles was different.

2.3 Comparison of related work

Table 2.1 Comparison Table of related work

RELATED WORK	ACCURACY RATE
Sentiment analysis of hindi reviews based on negation and discourse relation	82.89%
Performing sentiment analysis in Bangla microblog posts	93%
Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation	77%
Sentiment analysis on product reviews	83%
Sentiment analysis for Bangla sentences using convolutional neural network	99.87%
Data preprocessing techniques for classification without discrimination	83%
Comparing the performance of different NLP toolkits in formal and social media text	93%
An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques	75%
employing machine learning techniques on sentiment analysis of google play store bangla reviews	76.48%
Analyzing Sentiment of Movie Reviews in Bangla by Applying Machine Learning Techniques	88.90%

Based on the discussion above, we discovered that there was no noteworthy book review activity in Bangladesh. When we compare the two pieces of work, we can see that our

model has a larger dataset with high accuracy and has done well in a variety of areas. We may use our material in a web-based platform.

2.4 Research Summary

The research described above was carried out by several research groups, demonstrating the breadth of emotional analytics research. We have effective outcomes as a consequence of our analysis. Despite the lack of resources, it is envisaged that each sector would become more resourceful by adding information on the purchase of various items after one day.

2.5 Challenges

The most difficult aspect of the task is planning the data sets for subsequent processing. To make the data set correct for our work or future processing, we employed some advanced useable ML tools. Another issue in Bangladesh is the inability to locate enough money or work. One of the most difficult aspects of our job is attempting to adapt the ML paradigm to the internet.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The working approach comprises 5 stages in the collection, study, execution of the algorithms, validation and web implementation. The chart of our work is presented in Figure 3.1.1

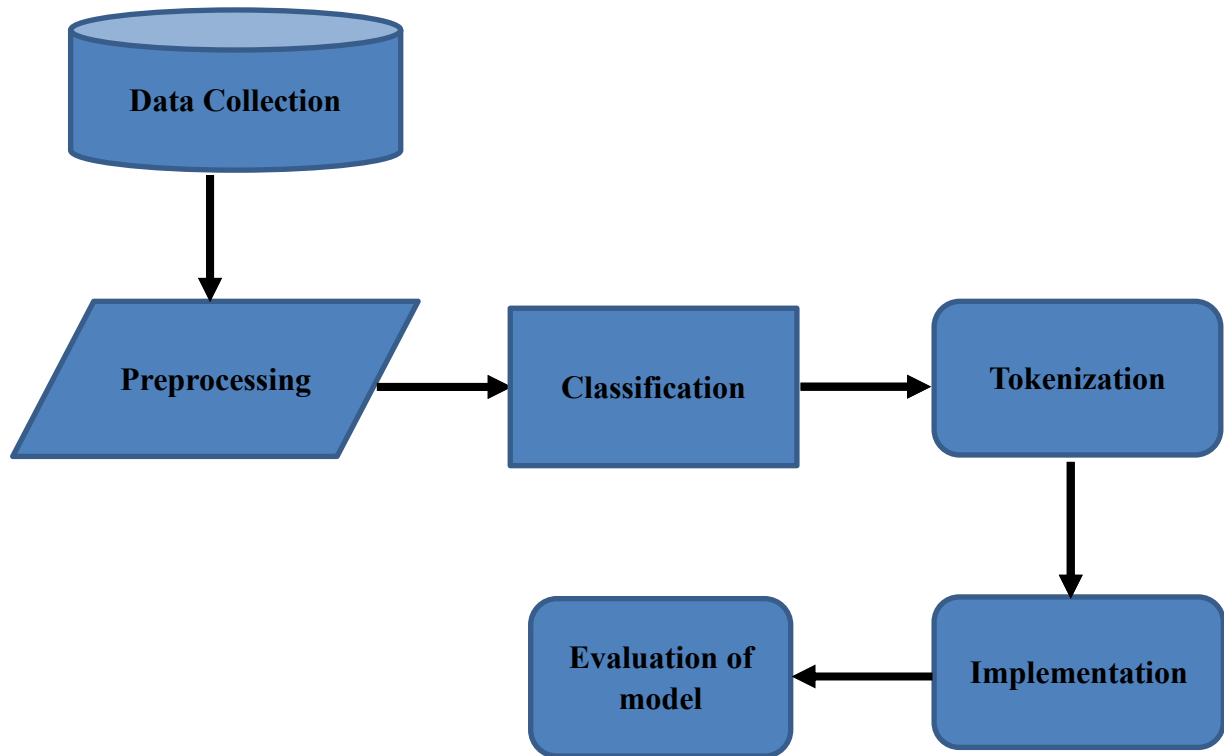


Figure 3.1: Methodology diagram

3.2 Data Collection

Data collecting is at the heart of every study. A book review is a delicate piece of information that is essential to the success of any book. We must also get our information from a reputable source. The data for our study comes from book reviewer comments. This was prepared for Facebook from numerous book vendor websites and book review pages. We only gathered comments in Bangla because that was our duty. We collect 5500 data for this research.

3.3 Data Pre-Processing

Data preprocessing is a data mining tool that converts raw data into a usable and efficient format. Information preprocessing is critical for knowledge acquisition. Our function is based on KDD. According to Kamiran et al.[9], the four most important data pre-processing procedures are elimination, data massaging, weighting, and Same poling. To develop accessible data sets, we predominantly used data messaging methodologies in our work. We deleted unneeded points and phrases from the Bangla stop in this level. We've decided to use our updated feedback as a feature to carry out all of the steps.

3.4 Classification

Our information was divided into two groups: good and negative. The courses are designed with the sentiments of the users in mind. If the novel's analysis is good, this sentence gets given a positive score. Negative assessment categories have also been chosen. This is our data

collection in Fig. 3.4.1. Our data collection includes 47.7% favorable reviews and 53% negative reviews of the 5500 data.

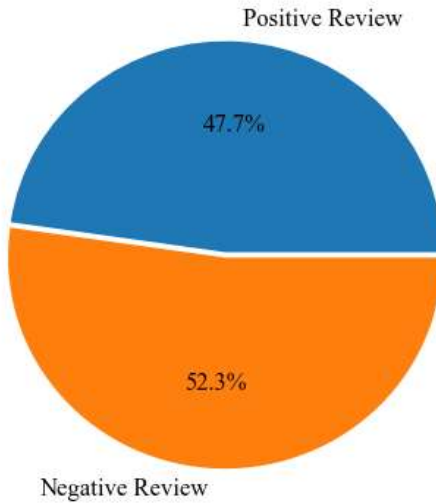


Figure 3.2: Classification

3.5 Tokenization

Tokenization is defined by Pinto et al[10] as a method of separating flag phrases, which can be words or signals. In our database, we have a lot of phrases. We did not complete our task using a phrase mark rather than a word label. It's also crucial to tokenize. Tokenization divides our entire phrase into terms. Table 3.1 shows the tokenization approach.

Table 3.1 Tokenization Table

Raw Data	Type	Tokenized data
বই এর গল্পগুলো অনেক সুন্দর	Positive	‘বই’ , ‘এর’ , ‘অনেক’ , ‘গল্পগুলো’ , ‘অনেক’ , ‘সুন্দর’
বইয়ের লেখাগুলো স্পষ্ট ছিল না	Negative	‘বইয়ের’ , ‘লেখাগুলো’ , ‘স্পষ্ট’ , ‘ছিল’ , ‘না’
বইয়ের পৃষ্ঠার মান ভাল ছিলনা	Negative	‘বইয়ের’ , ‘পৃষ্ঠার’ , ‘মান’ , ‘ভালো’ , ‘ছিলনা’

3.6 Algorithm Implementation

We discussed the algorithm implementation procedure in this part. To finish this procedure, we must first complete the preceding one in order to create the requisite dataset. Because our job is in the classification form, we have five distinct classification methods. For classifiers, we employ five algorithms: Adaboost, Decision Tree, SVM, LightGbm, and Random Forest. The best suited parameter to achieve highest accuracy for various methods is shown in Table 3.2.

Table 3.2 Parameter usages

Algorithms	Details
AdaBoost	n_informative=3, n_redundant=0, random_state=1, shuffle=True
Decision Tree	random_state=46
SVM	kernel='rbf'
Random Forest	n_estimators=80
LightGbm	BOOSTING_TYPE='GBDT', NUM_LEAVES=31, MAX_DEPTH=-1, LEARNING_RATE=0.1, N_ESTIMATORS=100, SUBSAMPLE_FOR_BIN=200000

Table 3.3 displays the parameters and other items we used to implement the algorithms we picked.

3.7 Evaluation

We evaluated our preferred RF technique utilizing real-time data estimation and an uncertainty matrix. We initially acquired 80 genuine data points from which our model failed to learn. For each of the classes chosen, different pages of online book sales websites and Facebook Bangla book reviews were used.

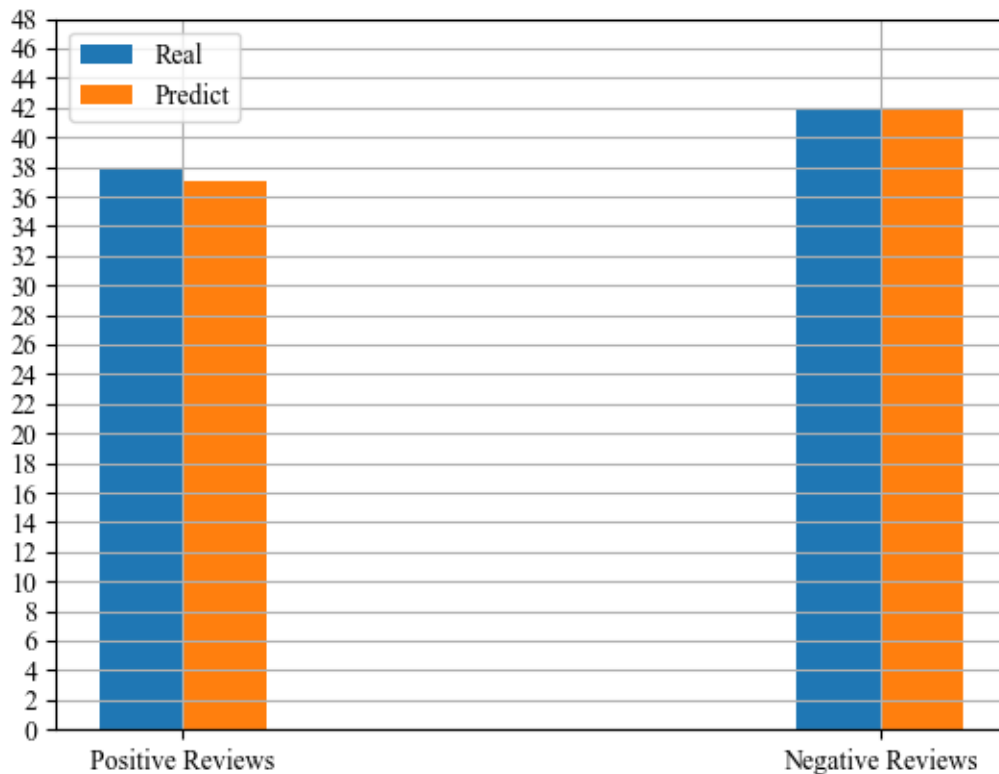


Figure 3.3: Comparison between Real and Predicted

Figure 3.7.1 shows a comparison of the actual and predicted results. There are 38 positive reviews and 42 negative reviews in our dataset, which are represented by blue bars. The color bar Orange represents the predicted value. Our model predicts one less positive ratings. e negative review model predicts the same results as the actual results. This is a minor flaw in our model. As a result, we may presume that our model worked well with real-world data. Confusion matrix can also be used to test this forecast.

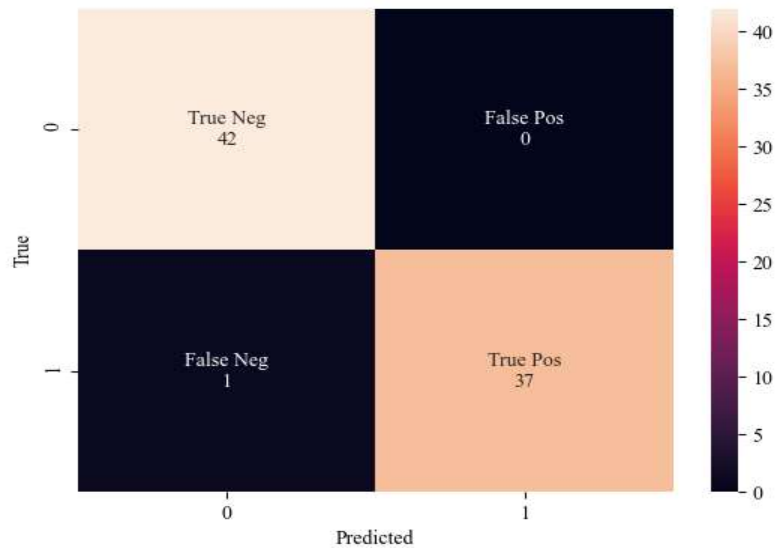


Figure 3.4: Confusion Matrix

$$\text{Accuracy} = \frac{42 + 37}{42 + 37 + 1 + 0} = 0.9875 * 100$$

$$= 98.75\%$$

$$\text{Error} = 1 - 0.9875 = 0.0125 * 100 = 1.25\%$$

Recall rate for positive:

$$\frac{37}{37+1} = .973 * 100 = 97.3\%$$

$$\text{Recall rate for Negative: } \frac{42}{42+0} = 1 * 100 = 100\%$$

To detect overall outcomes, we employed the Confusion Matrix. The validation dataset uncertainty matrix is shown in Figure 3.4. In the assessment procedure, we have a precision of 98.75%. This also means that our model works with both visible and hidden data. The percentage of positive memory is 97.3 percent, whereas the rate of negative recall is 100 percent. Rather than being a positive for bad reviews, it's an excellent illustration for our model.

CHAPTER 4

RESULT ANALYSIS

4.1 Introduction

In the analytical research, this portion mostly depends on empirical evidence and test results. When we examine a subject, what is the initial result analysis? The repercussions segment should be structured so that the outcomes are stated without any interpretation or evaluation. The guidance is also accessible in the academic papers area. The results are announced, and the test is shown. We also looked at a variety of algorithms and will discuss which ones are the best in a series of five algorithms. Precision, accuracy, reminder, and f1 were also chosen as parameters for computing the data.

4.2 Experimental Result

Table 4.1 Accuracy table

Test data usage rate		30%	40%	50%	60%	70%
Algorithms Accuracy	<i>Ada</i>	86.67	86.44	88.00	86.36	85.06
	<i>DT</i>	90.06	88.95	88.00	87.09	83.30
	<i>SVM</i>	98.55	98.49	97.20	96.79	95.04
	<i>RF</i>	99.39	99.06	97.09	95.06	92.99
	<i>Light</i>	92.67	93.09	90.98	89.15	85.19

Table 4.2.1 shows the precision table. We utilized 30 to 70% of test data to determine which item works the best. The test percentage for each algorithm that delivers the best accuracy is shown in yellow boxes. Except Ada algorithm, most algorithms perform best below 40% of the test results, as seen in this table. Under 50 ppm, lightgbm provided 88.00 ppm precision, while the RF provided 99.39 ppm precision utilizing just 30 ppm. RF still has the greatest accuracy of all the algorithms with red boxes on the table.

Table 4.2 Different Score Matrix

Score Matrix	Algorithms				
	<i>adaboost</i>	<i>Decision tree</i>	<i>SVM</i>	<i>Random Forest</i>	<i>lightgbm</i>
F1 Score	0.8663	0.8998	0.9846	0.9949	0.9413
Recall	0.8213	0.8935	0.9734	0.9911	0.9049
Precision	0.9165	0.9062	0.9961	0.9987	0.9808
Specificity	0.8505	0.9037	0.9761	0.9919	0.9187

Table 4.2.2 displays the Score Matrix. We've only gone through 30% of the scoring matrix. Because the precision table only shows precision that is based on true positives and true negatives, to assess the accuracy, several attributes were utilized, such as true negative, false positive, true positive, and false negative. The RF produced the best verification of exactness table in all dimensions, including F1 score, recall, accuracy, and specificity. As a result, the RF algorithm was chosen as the prediction method for this study.

4.2.1 AdaBoost

AdaBoost is best used to improve decision tree performance on binary classification issues. AdaBoost was initially known as AdaBoost.M1 by the technique's creators, Freund and Schapire. It's been dubbed discrete AdaBoost in recent years because it's utilized for classification rather than regression. AdaBoost is a machine learning method that may be used to improve the performance of any other machine learning technique. It works well with students who are struggling. On a classification task, these are models that reach accuracy slightly above random chance. Figure 4.1 shows that the AdaBoost algorithm's best accuracy is 88.00 percent, with a precision rate of 0.9165.

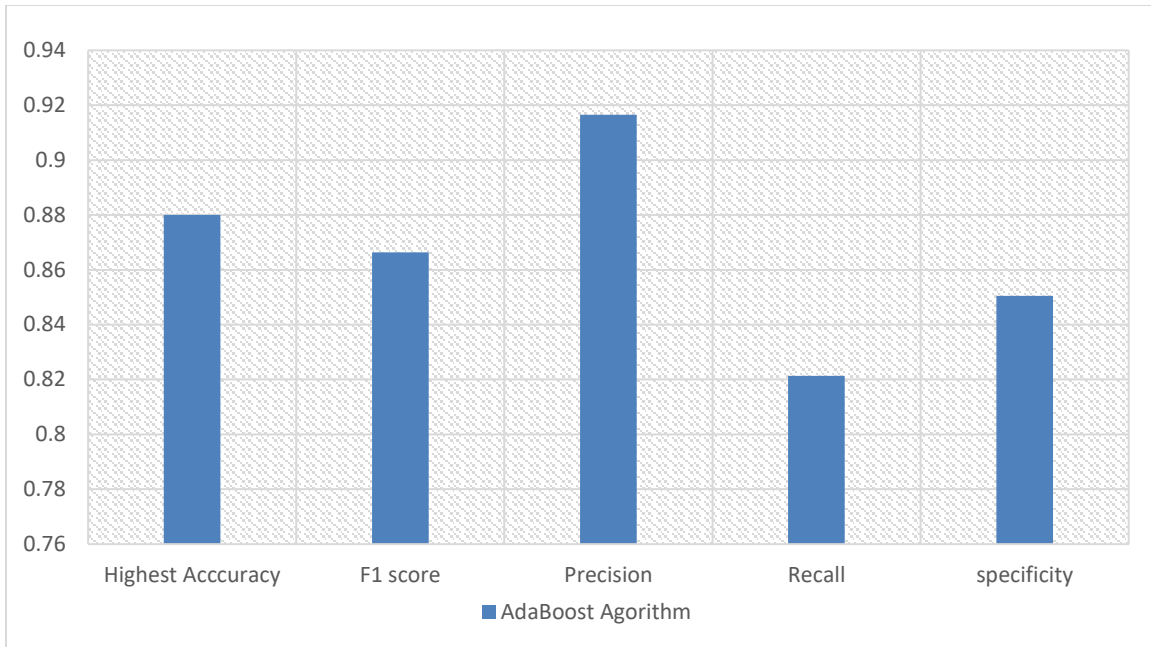


Figure 4.1 Different Score comparison graph of AdaBoost

4.2.2 Decision Tree

A decision tree is a greedy algorithm that divides each node accurately using local knowledge. One of the ramifications is that divisional variables can be used to modify a stronger tree. Trees are recognized to be incredibly adaptable, and their interactions are modest. The drawback is that the tone, dubbed "high variance," is aware of the outcomes. Overfitting is also a result of strong disparities, with tree projections being too optimistic. The Decision Tree produces excellent results and works with a complex dataset. [11] As a result of changing the division variables, a stronger tree can be formed. Low distortion is a term used to describe how versatile trees are in their interactions. The disadvantage is that they will learn the tone of the findings, which is referred to as high variance. High variances also lead to overfitting, since the tree makes too optimistic predictions. Figure 4.2.1 shows that the Decision tree algorithm's best accuracy is 90.06 percent.

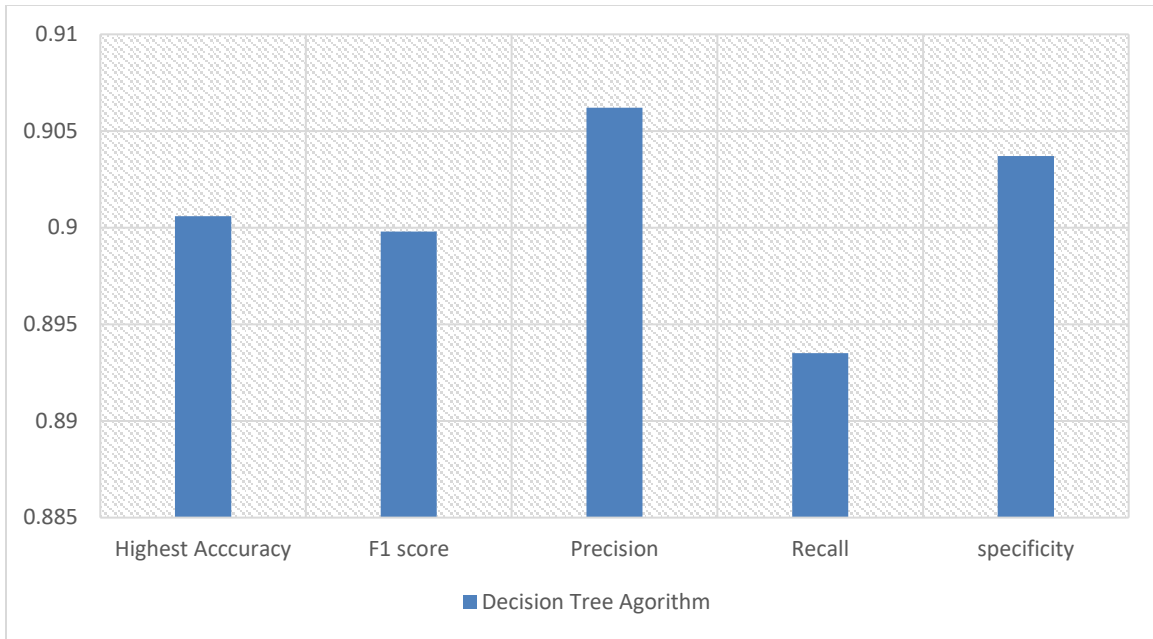


Figure 4.2: Different Score comparison graph of Decision Tree.

4.2.3 SVM

Support vector machines are a collection of supervised learning algorithms for classification, regression, and identification of outliers. These are all frequent machine learning tasks. You may use them to detect malignant cells based on millions of photos, or you can use a well-fitted regression model to forecast future travel routes. Support vector regression (SVR), which is an extension of support vector classification, is one sort of SVM you may employ for certain machine learning challenges (SVC). The most important thing to remember is that they are merely arithmetic equations that have been fine-tuned to provide you with the most accurate result possible as rapidly as feasible. SVMs vary from other classification algorithms in that they choose a decision boundary that optimizes the distance between all classes' closest data points. The maximum margin classifier or maximum margin hyper plane is the decision boundary established by SVMs. Figure 4.2.2 shows that the Decision tree algorithm's best accuracy is 98.55 percent, with a precision rate of 0.9961.

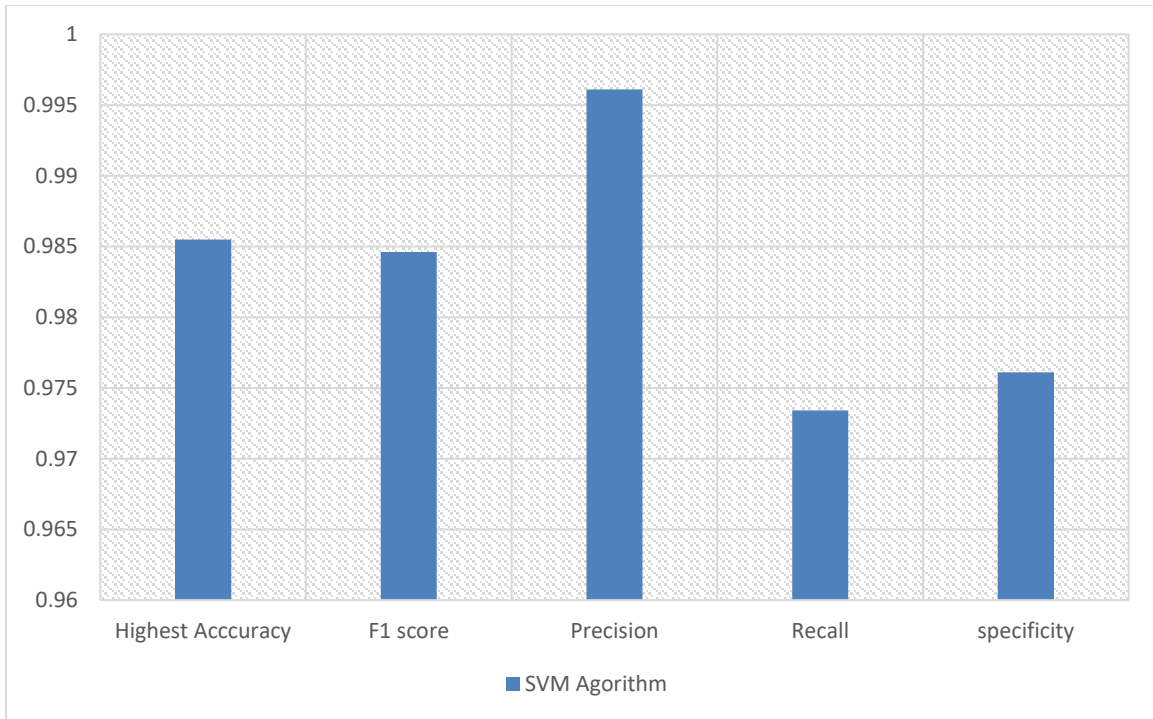


Figure 4.3: Different Score comparison graph of SVM

The SVM algorithm provided the best results. The greatest level of accuracy was 98.55 percent, and the other numbers were extremely close to it. Figure 4.2.3 depicts the entire score matrix visually.

4.2.4 Random Forest

Random forest is a flexible, easy-to-use algorithm that, in most cases, gives excellent results without the need of hyper parameters. It is also one of the most often utilized algorithms due to its simplicity and versatility (it can be used for both classification and regression tasks). In this post, we'll look at how the RFAl works, how it differs from other algorithms, and how it's employed. It creates a "forest" of decision-making trees that are often taught in "sacking." The core premise of the box method is that a combination of learning models improves the end output. Random forest may be used for classification as well as regression.

In our classification job, random forest provides an accuracy of 99.39 percent and a precision rate of 0.9987 percent, as shown in figure 4.2.4.

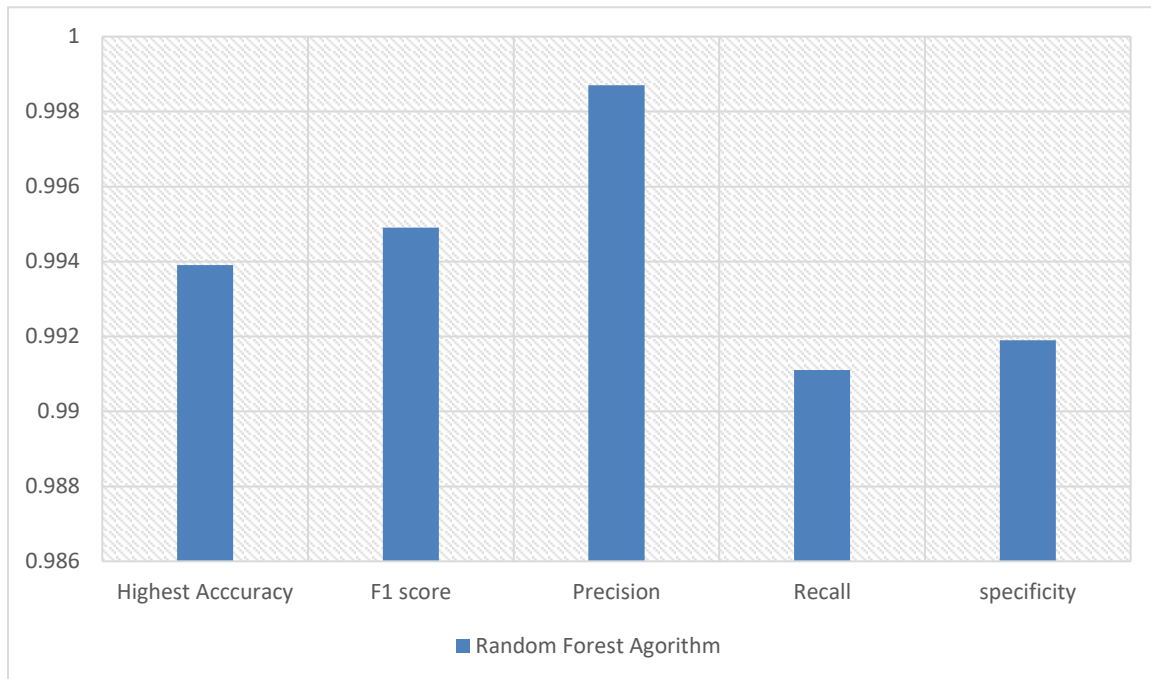


Figure 4.4: Random Forest Score Comparison

4.2.4 LightGBM

It's a gradient boosting framework that uses tree-based learning methods, which are regarded to be a very strong processing technique. It is thought to be a quick-processing algorithm. While the trees of other algorithms develop horizontally, the LightGBM method grows vertically, which means it grows leaf-wise while other algorithms grow level-wise. To grow, LightGBM selects the leaf with the greatest loss. When expanding the same leaf, it can reduce loss more than a level wise method. LightGBM is a highly optimized histogram-based decision tree learning technique that provides good results and memory savings. The highest accuracy is 93.09 percent and a precision rate of 0.9808 percent that is graphically represented in figure 4.2.5.

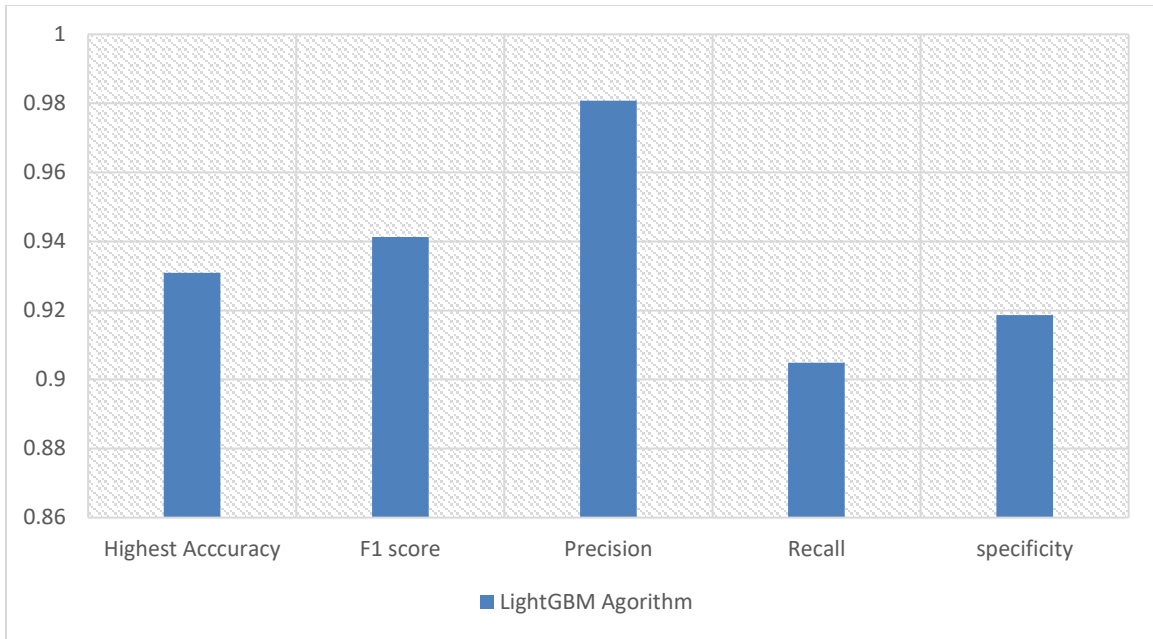


Figure 4.5: Different Score comparison graph of LightGBM Algorithm

CHAPTER 5

SUMMARY, CONCLUSION AND FUTURE WORK

5.1 Summary of the Study

Customers of online bookstores are increasing in lockstep with the number of people using the internet. Science and technology have brought people from all around the world closer together. The rapid rise of an online bookstore has been facilitated by e-commerce. People nowadays do not want to waste time going to a physical bookstore; instead, they want convenience and to live as simply as possible. Machine learning has gotten a lot of attention, but there hasn't been much study done in Bangladesh. Despite the fact that work in predictive styles is a common word for computer education, Bangladeshi Book is unaware of this. This type of research has recently been carried out as a result of those tasks causing a significant alteration in our machine life. However, there is little study being done in the subject of Bangladeshi economics. However, we expect that a lot of scholars in this subject have conducted study in a variety of nations.

5.2 Conclusion

SA is dependent on a dataset of specific material due to the fast growth of Internet users. Using multiple feature extraction approaches, this research proposes a machine learning-based sentiment classification system that can classify sentiment into positive and negative categories from Bengali book reviews.

We analyzed 5500 consumer reviews from 55 different book categories. We gathered book information and descriptions, as well as screening book reviews for key feature phrases. As a consequence, we discovered five primary characteristics that exist in virtually all customer evaluations and have an impact on them: pricing, transportation, quality, design, and satisfaction. Following that, we created a machine learning model.

Our goal is to rate book reviews with a 99.39 percent accuracy rate. The RF algorithm's precision has been discovered. RF had the best perforation, which helped it outperform other common algorithms like AdaBoost, Decision Tree, SVM and LightGBM in terms of

accuracy. Both bookshop owners and customers may learn which books are worth looking at and which ones aren't, and potential purchasers can identify which books have good or awful characters. This method is beneficial to bookstore owners. The experiences of librarians and book users will undoubtedly change throughout this time.

5.3 Recommendations

There are a few excellent ideas for this:

- To increase the accuracy of data collecting in order to get better outcomes from this study.
- The amount of data in this paper is really limited.
- It would be preferable to utilize Deep Learning.

5.4 Future Work

The future guidance on the development of this work is given bellow:

- In Bangladesh, we aim to investigate the sensation of a caustic statement.
- We'll create a framework for implementing MLT.
- We want to working on a Web-based API to express analytical feelings in order to achieve this goal.
- We will develop an intelligence system based on deep learning techniques in the future.

REFERENCE

- [1] (2020). Literature of Bangladesh, culture. Available at << <https://www.bangladesh.com/culture/literature/>>> , last accessed on 2-03-2021 at 8 AM
- [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), Dha-ka, Bangladesh, 2014, pp. 1-6, doi: 10.1109/ICIEV.2014.6850712
- [6] M. H. Alam, M. Rahoman and M. A. K. Azad, "Sentiment analysis for Bangla sentences using convolutional neural network," 2017 20th International Conference of Computer and Information Technology (ICIT), 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), Singapore, 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. Gonalo 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.
- [14] 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.

APPENDIX

The first was to outline the procedures for the analysis, which presented a number of difficulties. Furthermore, no progress has been made in this area previously. Indeed. It wasn't your usual work. We couldn't find someone who could help us that much. Another stumbling block was data collection, which proved to be a huge issue for us. We created a data gathering corpus because we couldn't locate an open source Bangladesh text pre-processing program. We've begun manually collecting data. Furthermore, classifying the various postings is a difficult task.

PLAGIARISM REPORT

12/4/21, 9:38 AM Turnitin

Turnitin Originality Report

Document Viewer

Processed on: 04-Dec-2021 09:36 +06
ID: 1720086999
Word Count: 4709
Submitted: 1

Similarity Index 14%	Similarity by Source
	Internet Sources: 8%
	Publications: 7%
	Student Papers: 5%

BOOK REVIEW SENTIMENT IN BANGLA LANGUAGE USIN... By Mst. Eshita Khatun

[exclude quoted](#) [exclude bibliography](#) [exclude small matches](#) mode: quickview (classic) report

[Change mode](#) [print](#) [refresh](#) [download](#)

2% match (student papers from 11-Feb-2018) Submitted to Daffodil International University on 2018-02-11	■
2% match (publications) Md. Hamidur Rahman, Md. Saiful Islam, Md. Mine Uddin Jewel, Md. Mehedil Hasan, Ms. Subhenur 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	■
1% match (Internet from 17-Aug-2020) https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/	■
1% match (publications) "Progress in Advanced Computing and Intelligent Engineering", Springer Science and Business Media LLC, 2021	■
1% match () Alam, Firaj, Hasan, Arid, Alam, Tanvirul, Khan, Akib, Tajrin, Janntatul, Khan, Naira, Chowdhury, Shammur Absar, "A Review of Bangla Natural Language Processing Tasks and the Utility of Transformer Models", 2021	■
1% match (student papers from 27-May-2021) Submitted to Sheffield Hallam University on 2021-05-27	■
1% match (student papers from 18-Jun-2021) Submitted to University of Wales Institute, Cardiff on 2021-06-18	■
1% match (publications)	■

https://www.turnitin.com/newreport_classic.asp?lang=en_us&oid=1720086999&f=1&bypass_cv=1