CLASSIFICATION OF BOOK REVIEW SENTIMENT IN BANGLA LANGUAGE USING NLP AND MACHINE LEARNING

BY
MD. HAMIDUR RAHMAN

ID: 172-15-9583

MD. SAIFUL ISLAM

ID: 172- 15-9920

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

Supervised By

Ms. Subhenur Latif
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Zerin Nasrin Tupma
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

1 JUNE 2021

APPROVAL

This project titled “**Classification of Book Review Sentiment in Bangla Language Using NLP and Machine Learning** ", submitted by **Md. Hamidur Rahman** and **Md. Saiful Islam** 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 1st June 2021.

BOARD OF EXAMINERS

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

ফিজার আহমেদ

Dr. Fizar Ahmed **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

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

External Examiner

Associate 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 **Ms. Subhenur Latif**, Assistant Professor, Department of CSE, and co-supervision of **teacher**, Senior Lecturer, and **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:

Subhenur Latif
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Hamidure Rahman

Md. Hamidur Rahman
ID: 172-15-9583
Department of CSE
Daffodil International University

Saiful Islam

Md. Saiful Islam
ID: 172-15-9920
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 **Ms. Subhenur Latif**, Senior Lecturer, Department of CSE Daffodil International University, Dhaka, Bangladesh. Profound Knowledge & intense interest of our supervisor in the field of "Machine Learning & Deep Learning" make our way very smooth to carry out this thesis. Her remarkable patience and dedication, 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

Books are said to be a person's best friend. Learning books is essential to acquire expertise. Community for book reading has been around for over a few thousand years. Since ancient civilization learns to write, knowledge is said to be the ancestors of books in tablets or walls or stones. The most up-to-date type of books is e-books, digital print or print on paper. In Bangladesh, even a few years ago, the people had to go to the library in person to gather books. Many of the advantages are simple and the internet bookshop has a penalty, i.e. the reader does not know the books or the book store itself. Book readers prefer to rely on feedback and ratings in order to prevent this. Our aim is to evaluate Bangladesh language reviews and to provide correct input on the books and online shop so that book readers can purchase correct books in order to read and get better online bookshops. About 1500 raw data have been obtained for training the system. In order to distinguish negative and constructive review results of previous users, we use Natural Language (NLP) processing. The data analyzers include a random forest, logistical regression, decision tree, random forest and some common algorithms like Subject Vector Machine (SVM).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

CHAPTER 1
INTRODUCTION

1.1 Introduction

Literature has a particular position in Bangladesh. Literary traces have been written in nearly a few centuries [1]. While reading books at the time became a pleasure, it was normalized over time. Because books are the primary source of expertise, people have increasingly become books. In modern days, books are unfortunately not as common among the people of Bangladesh as they were a few decades ago. However, it retains its highest status for scholarly purposes. As the Internet is now a fixation for almost all ages, bookshops and libraries are online to keep the new generation present.

The most famous bookshop in Bangladesh are online shops like Rokomari, Boibazar, Boikhata, eBoighar, Bookshopbd, Boi-BoiBoi, Wafilife. When technological progress helps to bring people together, it's easier to cheat people or lie on the Internet. For the internet ads of bookstores in particular, people are afraid to purchase books. They also focus on feedback and ratings from previous customers on these products and service providers. In Bangladesh, more often than not, people prefer Bangladesh as the language of the study. They're fine with it, because it's their mom's guide. Reviews play an important part in the general online marketplace. We aim to evaluate and determine negative or positive feedback of a given book on a ratio basis via the Machine Learning Algorithm. Sentiment analyzing has drawn more re-searchers to it than ever since it is a main property of NLP (natural language processing). One of the main topics of feeling assessment is the classification of polarity, the process of understanding and the as-session of the document in positive and negative type [2]. We attempt to solve the problem in a special way. Question: Why do we need this review analysis, when the exact number of participations is positive or negative? The answer is, sometimes we see that, regardless of their assessment, people seem to rate any amount as scores. What's unclear. So we use NLP to evaluate feedback of positive or negative feelings to minimize such uncertainty. The data are analyzed using the Support Vector Machine (SVM), Decision Tree, KNN, Random Forest and Logistic Regression machine learning algorithms.

1.2 Motivation

The Book selling website is increasing day by day in Bangladesh. Most of people are comfortable by ordering book via online. Due to corona pandemic this incident is increasing day by day for book lover. But when they order book in online some of limitation is arrived. One of this is they cannot read overall book when they order book though some of website is allowed to read some page of a book but is very difficult to take idea about whole book by this way. Another way is book review, it is an analysis that refers key assessment of a document, case, entity, or phenomena. Books, posts, whole genres or sector, architecture, sculpture, design, restaurants, policy, exhibits, performances and many other types can be considered in reviews. This presentation concentrates on book reviews. Then most of buyer read review of book to by this. Book review is efficient way to get overall idea about any book. But it is very time consume task. And sometime buyer feels very boring to read each and every comment. From above discussion we can understand that something is needed to solve this problem in such way that save people time as well as they can get their wanted book easily. These two limitation is motivated us to do our work. And we want to decide we will make an intelligence system that can analyze the use comment and percentile them into positive and negative categories. Now the question is rating is another things and by the help of rating it is very easy to find out best product from a website. But actually rating is given based one different parameter. For example, book quality, delivery time, behavior of delivery man that means it represent overall service quality of this company. And for this reason it is not a perfect way to select a book. And finally we have decided to solve this problem using NLP and Machine learning. As we know algorithm does not understand string directly. At first we need to convert string to numeric form. In this case we used TFIDF algorithm. And to categorize each comment we used Machine Learning algorithm. For each algorithm we used different parameter. And we selected these parameters which is produced best performance.

1.3 Problem Definition

A book is a medium for documenting material in writing or in the form of photographs, usually consisting of several pages linked and covered. This physical structure is technically called Book. Bengali or Bangla Shahitto literature reflects the corpus of Bengali Language. Most of time in Bangladesh people go to store and then buy their needed book. As the era of internet and ecommerce online book shop is increases day bay day in Bangladesh. So now people are very comfortable with online book website for ordering book. For time saving and providing best book for customer this research is introduced. In this research we used Natural Language Processing and Machine Learning. There are many problems arrived to do this work. Data collection is very sensitive task for this research because we work with human sentiment. For data collection we visited different book selling website. And collected each and every comment of different book. We collected Negative and positive comment. These comment is our feature data. we collected around 1500 Bangla comment. This is our raw data that contains lots of noise like sometime we found double word extra punctuation mark and different emoji. In preprocessing stage, we removed all of this noise to learn our algorithm properly. When preprocessing is completed we used TFIDF algorithm to convert string to numeric format. After competition of making numeric format we used different Classification Machine Learning algorithm, as our work is a classification based. Every sentence is classified into two categories one is Positive another is Negative. After training state is completed we evaluate our work by fetching original data that is not trained. In evaluation stage our algorithm perform better. We illustrated each stage by different graph.

1.4 Research Questions

- How will the dataset be gathered and prepared?
- Can the positive and negative groups be correctly defined?
- How will positive and negative be classified?
- Can the machine learning process predict Positive and Negative class correctly?
- Is it possible to implement the on the Internet?
- How can the people be helped by this work?

1.5 Research Methodology

Our workflow, involving data processing, information processing, data classification, algorithm implementation, will be covered in this section. Model training, Algorithm assessment.

1.6 Research Objectives

- To anatomize consumer analysis by using those classification algorithms or classifying them.
- To build a model that can reliably detect positive and negative comments.
- To imagine a certain scientific feeling research.
- Create a software application to view the price using engineering tools and machine learning

1.7 Research Layout

The substance of our study is as follows:

Chapter 1 This first segment is a critical step of the initial analysis. In addition, this chapter explains what inspired us to perform such analysis. The problem description is the most critical aspect of this chapter. The segment also contains the study issue, the challenge

Chapter 2 This consists of an input analysis which provides a concise overview of the work in this area. The related notable work with machine learning is described here.

Chapter 3 is a simple methodology or workflow summary. is given. How was the analysis conducted in this segment addressed?

Chapter 4 It's in the assessment of the results. It includes the results of the graphic analysis.

Chapter 5 It's the part of the study closing. The model output is discussed in this section. The exactness of the relation is also seen in this section. This section also contains the online implementation portion of the concept and performance. The chapter concludes by pointing at the work's limits. The potential study was also encoded.

1.8 Expected Outcome

- We will detect negative and positive sentiment of user comment.
- We will save customer time.
- We will try to show best book based on customer choice.
- We created a robust web application which shows the outcome of any sentiment of a book review.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

Machine learning for prediction has been studied in different ways. Prediction is one of Machine Learning's most widely used applications. A large number of sentiment analysis studies were carried out. These studies concentrated on issues and used various machine learning algorithms to solve the problem. This chapter offers a summary of the corresponding activities carried out efficiently by several experts in the previous region.

2.2 Related Works

In today's time almost everything is web-based. People on the internet share their views. The researcher magnet is often measuring people's feelings. In several sectors and languages, this issue was presented.

Aspect-based Sentiment Analysis is a method of sentimental analysis that selects an individual subject and analyzes the feelings around the subject. With this approach, Rahman et.al [3]. conducted Bangladesh research. In the Bengali language the sentiment analysis is advancing and is considered a primary research interest. As tools like a properly annotated data set are scarce, corporate language analysis, lexicon as part of the speech tagger etc, is difficult in Bengali. Their emphasis was on a restaurant review and the use of aspect-based research for cricket opinions. SVM gave 71% and 77%, which was the highest validity for extracting and finding polarity in crickets and restaurants, respectively.

Mittal et.al [4]. suggested a method for the analysis of Hindi that gives positive and negative validity to 82.89 and 76.59%. They decided to evaluate emotions and expand coverage of the data base in order to enhance the consistency of the database. This article presents an educational model which explores the emotions of the Roman Urdu people, including the following genres: sports; software; food & recettes; drama; and political. It includes 10,021 phrases from 566 online threads. The objectives of this work are twofold:

(1) creating a human-annotated corpus for Roman Urdu for sentimental analysis; and (2) testing feeling analysis techniques based on the 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. Feedback from online products is growing every day. Reviews and opinions therefore play a key role in recognizing satisfactions for people. These opinion mining has the ability to discover details

In many downstream natural language processing tasks, modeling built with this type of network and its variants recently demonstrated their performance, especially in resource-rich language, such as English. However, for Bangladesh's classification tasks, these models have not been explored entirely. They fine-tune the multilingual text classification transformer model in Bangladesh. Alam et.al [6]. proposed a model of Convolution Neural Network in order to explain the text-based feelings conveyed by analyzing in Bangla (CNN). With 850 data, 350 were negative and 500 were positive remarks, CNN achieves 99.87 percent accuracy.

Knowing consumer needs are critical in online shopping, but firms can not be as conscious as they need to be. C. Chauhan et.al [7]. used algorithms for machine learning to distinguish negative and optimistic feedback of a good for potential users to authenticate their reviews. They reviewed various papers and concluded that Naïve Bayes generated positive results but the results differed in the setup and methodology with different objectives.

Tuhin et al [8]. suggested two methods for the classification and identification of various kinds of emotion from all kinds of Bangladesh. These were joyful, furious, sad, frightened, enthusiastic, and tender. Topical solution and method of grouping in Naïve Bayes are strategies. 7400 Bangladesh phrases were used as a data set and 90% accuracy was provided by topical approach. They also compared their paper with two other papers which gave SVM 93 percent and document frequency 83 percent. The parameter of the emotions discussed in all three articles was distinct.

2.3 Comparison of related work

Table 2.1 Comparison Table of related work

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

From the above discussion, we found that there was no remarkable work in Bangladesh for book review. By comparing the work in question, we can see that our model has a larger dataset of very adequate precision and has performed in many categories. In a web-based framework, we can apply our content.

2.4 Research Summary

The study above is carried out in different research teams, which demonstrates what research has been undertaken in the area of sentimental analytics. By analysis, we have successful results. Although not enough resources are available, it is hoped that each sector can become more resourceful by adding details on the buying of various products after one day.

2.5 Challenges

During the job we face the biggest challenge is to plan the data sets for further handling. We also used some advanced usable ML software to make the data set accurate for our work or for further manipulation. Another problem we face in Bangladesh is not to find sufficient money or employment. One of the biggest problems in our work is when we attempt to apply the ML paradigm on 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.

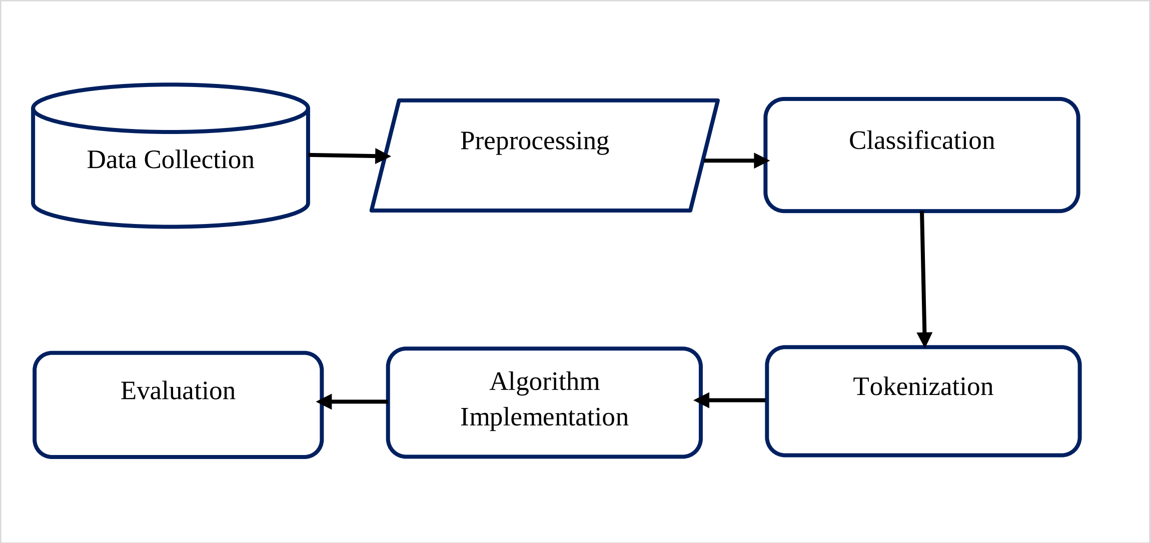


Figure 3.1: Methodology diagram

3.2 Data Collection

The core aspect of all research is data collection. Book review is a sensitive data and is the essence of every book. We must also obtain our knowledge from a credible source. Book reviewer comments are the data of our study. This has been compiled from various book seller websites and book review pages for the Facebook. We only collected comment for Bangla as our job for only Bangla.

3.3 Data Pre-Processing

Data preprocessing is a technology used for data mining that transforms raw data into a useful and efficient format. Preprocessing of information is very important for gaining knowledge. Our function is KDD-based. The key four pre-processing methods of data are removal, data massage, weighting, and Same poling, as defined by Kamiran et. al [9]. In our work, we primarily adopted the strategies of data messaging to produce usable data sets. In this level, we have removed words from Bangla stop and unnecessary points. We have chosen our updated feedback as features to execute all the measures.

3.4 Classification

Our data collection has been divided into two classes: positive and negative. The courses are designed according to the feeling of the user. If analysis is favorable in the novel, this phrase is graded as positive. Likewise, we have picked negative assessment groups

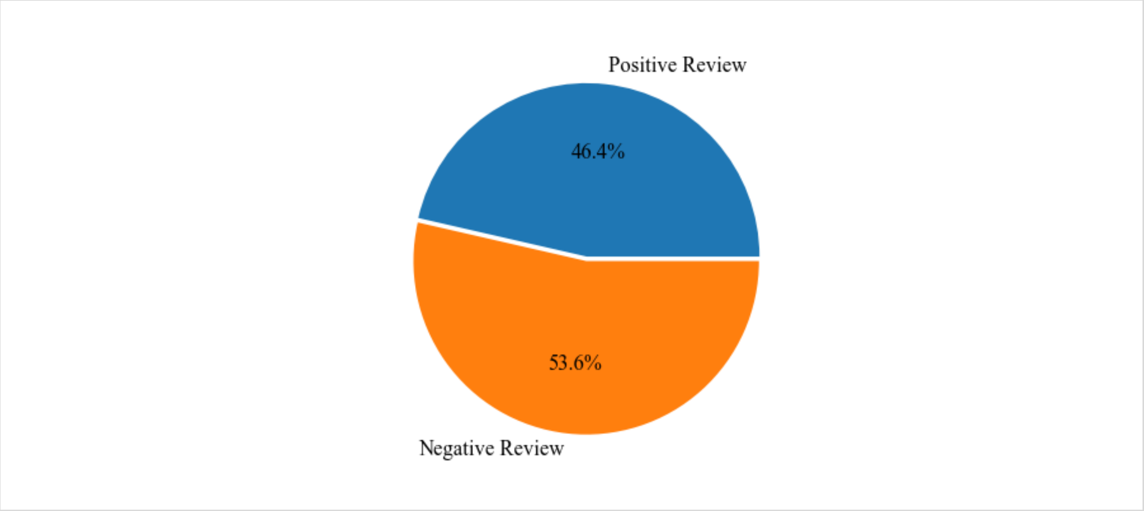


Figure 3.2: Classification

This is our data collection in Fig.2. Our data collection includes 46% favorable reviews and 53% negative reviews of the 1500 data.

3.5 Tokenization

In a Pinto et. al [10]. tokenization is defined as a way of separate flag phrases that may be words or signs. There is a number of phrases in our dataset. We did not do our job by phrase mark instead of by word label. Tokenization is also important. We divide our whole sentence into terms by tokenization. The tokenization method is seen in Table 3.1.

Table 3.1 Tokenization Table

Raw Data	Type	Tokenized data
অনেক সুন্দর বই	Positive	‘অনেক’ , ‘সুন্দর ‘, ‘বই’
বই এর লিখা খুব একটা ভালো না	Negative	’বই ‘ ,‘এর’, ‘লিখা’, ‘খুব’ ,‘একটা’, ‘ভালো’, ‘না’
বইয়ের পৃষ্ঠা ভালো না	Negative	‘বইয়ের’, ‘পৃষ্ঠা’ ‘ , ’ভালো ‘, ’না’

3.6 Algorithm Implementation

In this section we described the process of algorithm implementation. For doing this process we have to complete the previous process to make the required dataset. We have five different classification algorithms as our work is classification form. We use 5 algorithms for classifier such as KNN, Decision Tree, SVM, Logistic regression, Random Forest. Table 3.2 represents the most fitting parameter to generate the maximum accuracy for various algorithms.

Table 3.2 Parameter usages

Algorithms	Details
Logistic Regression	cv=5,max_iter=300
KNN	K=3,p=2,random_state=42
Decision Tree	random_state=42
SVM	kernel='linear', random_state = 0
Random Forest	n_estimators=100

Table 3.3 illustrates the parameters and the various things that we applied for implementing the chosen algorithms.

3.7 Evaluation

By using real-time data estimation and uncertainty matrix, we assessed our chosen SVM algorithm. At first, we gathered 37 real data which our model didn't learn. Different pages of on-line book sales websites and the Facebook Bangla book reviews were adopted for each of the classes selected

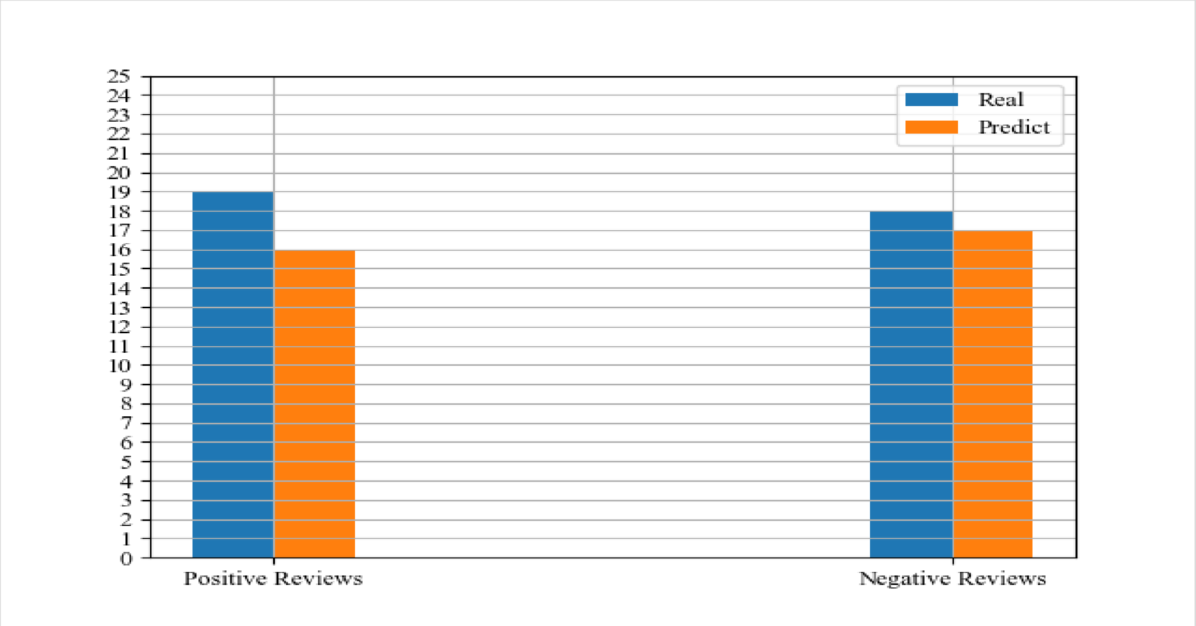


Figure 3.3: Comparison Between Real and Predicted

The actual and expected comparison as seen in Figure 3.3 Our dataset comprises 19 favorable reviews and 18 critical reviews which can be found in green bars. The color bar Orange reflects the value expected. Our model expects 3 fewer ratings with a favorable score. And 1 less review is predicted for the unfavorable review model. This is our model's tiny failure. We can therefore assume that for real-life data, our model was also very successful. This forecast can also be tested by confusion matrix.

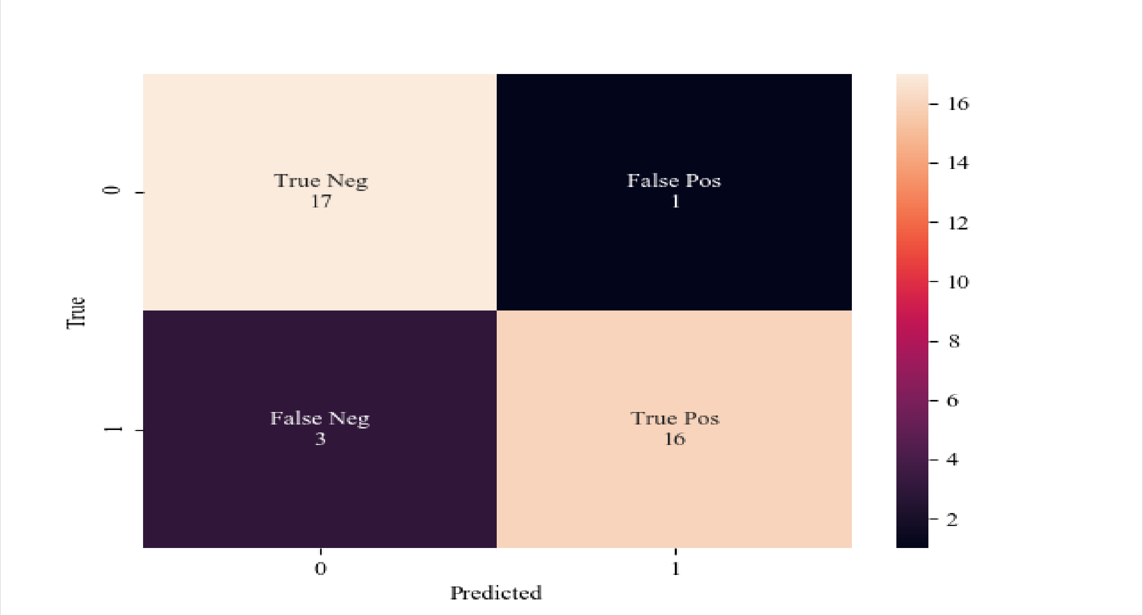


Figure 3.4: Confusion Matrix

$$\text{Accuracy} = \frac{16 + 17}{16 + 17 + 1 + 3} = 0.91 * 100$$

$$= 91\%$$

$$\text{Error} = 1 - 0.91 = 0.09*100 = 9\%$$

Recall rate for positive:

$$\frac{16}{16+3} = .842*100 = 84.2\%$$

$$\text{Recall rate for Negative}: \frac{17}{17+1} = 0.944*100 = 94.4\%$$

We used Confusion matrix to detect overall results. Figure 3.4. displays the validation dataset uncertainty matrix. We have 91 percent precision in the assessment process. This also means that our model is suitable for actual and invisible data. The positive recall rate is 84.2% and the negative recall rate is 94.4%. It's a good example for our model, rather than positive for poor reviews.

CHAPTER 4

RESULT ANALYSIS

4.1 Introduction

This section relies mostly on empirical evidence and test findings in the analytical study. What is the result analysis first when we evaluate the subject? The segment on consequences should be planned to say the results without interpretation or review. The guide is also available in the section on academic papers. The findings are announced and the test is shown. We also saw different algorithms and will clarify what algorithms are better in 5 algorithms. We also chosen precision, accuracy, reminder and f1 as a parameter for calculating the data.

4.2 Experimental Result

Table 4.1 Accuracy table

Test data usage rate		**30%**	**40%**	**50%**	**60%**	**70%**
Algorithms Accuracy	***KNN***	76.17	77.47	75.52	76.41	76.11
	Logistic	83.29	84.29	83.78	80.96	81.05
	DT	77.15	76.42	77.43	73.10	69.79
	SVM	85.01	84.84	84.66	81.57	81.47
	RF	82.56	83.16	82.45	79.73	78.21

The precision table in Table 4.1 is shown. In order to find out which anything best performs, we used 30 to 70 percent test results. Yellow boxes show the test percentage for each algorithm gives the highest accuracy. As seen in this table, most algorithms are best performed below 40% of the test results, apart from two, SVM and Decision Tree algorithms. Decision Tree gave 77,43 ppm most precision under 50 ppm, while the SVM gave 85,01 ppm precision using just 30 ppm. Of all the algorithms that have red boxes on the table, SVM still has the highest accuracy.

Table 4.2 Different Score Matrix

Score Matrix	**Algorithms**				
	KNN	***Logistic***	***Decision tree***	***SVM***	***Random Forest***
F1 Score	0.7868	0.8373	0.7518	0.8565	0.8020
Recall	0.8443	0.8255	0.7214	0.8585	0.7642
Precision	0.7366	0.8495	0.7846	0.8545	0.8438
Specficity	0.7988	0.8159	0.7217	0.8454	0.7674

The Score Matrix is shown in Table 4.2. We have just 30% analyzed the score matrix. Since the precision table indicates only precision dependent on true positive and true negatives, we have attempted to validate the accuracy with more attributes such as true negative, false positive, true positive, false negative. In all dimensions, i.e. F1 score, recall, precision and specificity of the SVM performed the best verification of exactness table. For this analysis, we have therefore chosen SVM algorithm as the predictive algorithm.

4.2.1 KNN

This easy to use algorithm and suboptimal is the most common machine study algorithm [11]. The most popular machine learning algorithm is KNN algorithm, which is simple to use and a suboptimal algorithm. Figure 4.1 Represents the all parameter scores. We had a preview rate of 77.47% for KNN model, while the data rate is 40%. There is no blueprint for the KNN algorithm except to store the whole data collection known as the training dataset. Data can be preserved by different means, but k-d trees are the most widely used KNN algorithm data structure. It helps to look and balance the new trends instantly and more efficiently. As long as new data are generated, the training data is curated and updated.

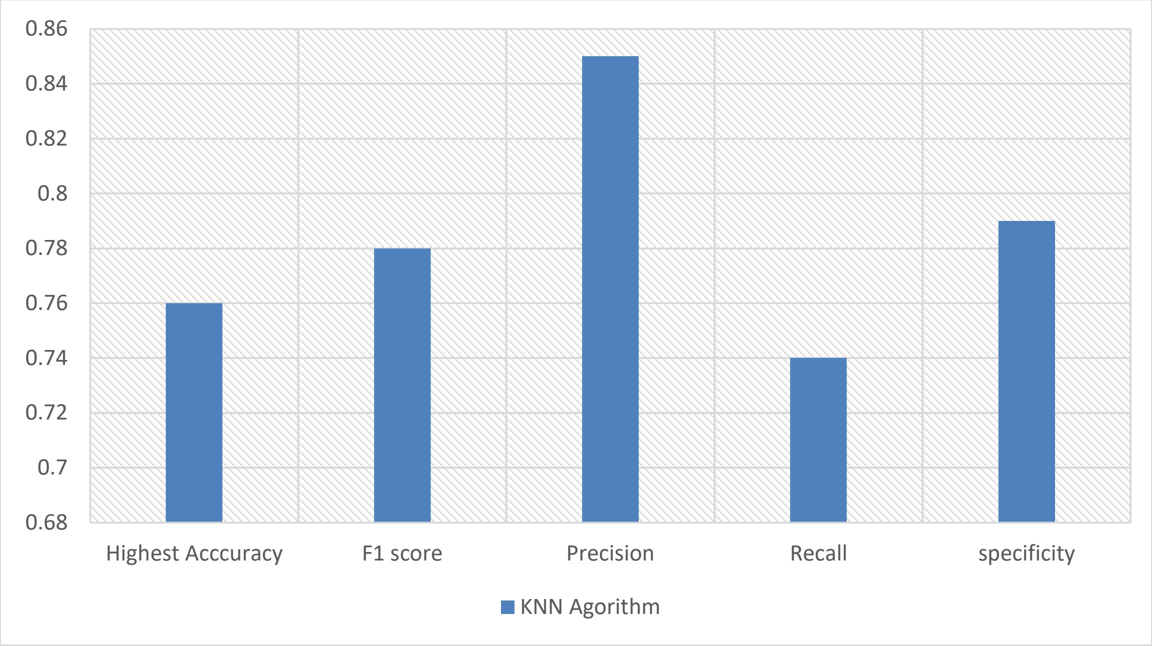


Figure 4.1 Different Score comparison graph of KNN

4.2.2 Decision Tree

Decision tree is a greedy algorithm that uses local information to divide each node correctly. One of the implications is that a stronger tree can be modified with the divisional variables. Trees are very flexible and their interactions are known to be minimal. The downside is that the tone, the named high variance, knows the results. Strong differences also contribute to overfitting, with tree forecasts being too optimistic. The Decision tree has impressive results and a complicated dataset [11]. One result is that a stronger tree can be generated by modifying the division variables. Trees have a high level of versatility in their interactions, known as low distortion. The downside is that they will learn the tone, called high variance, in the results. High variances also contribute to overfitting, with the tree making assumptions that are too optimistic. Figure 4.2 represents that the highest accuracy of Decision tree algorithm is 77% and the precision rate was 0.78.

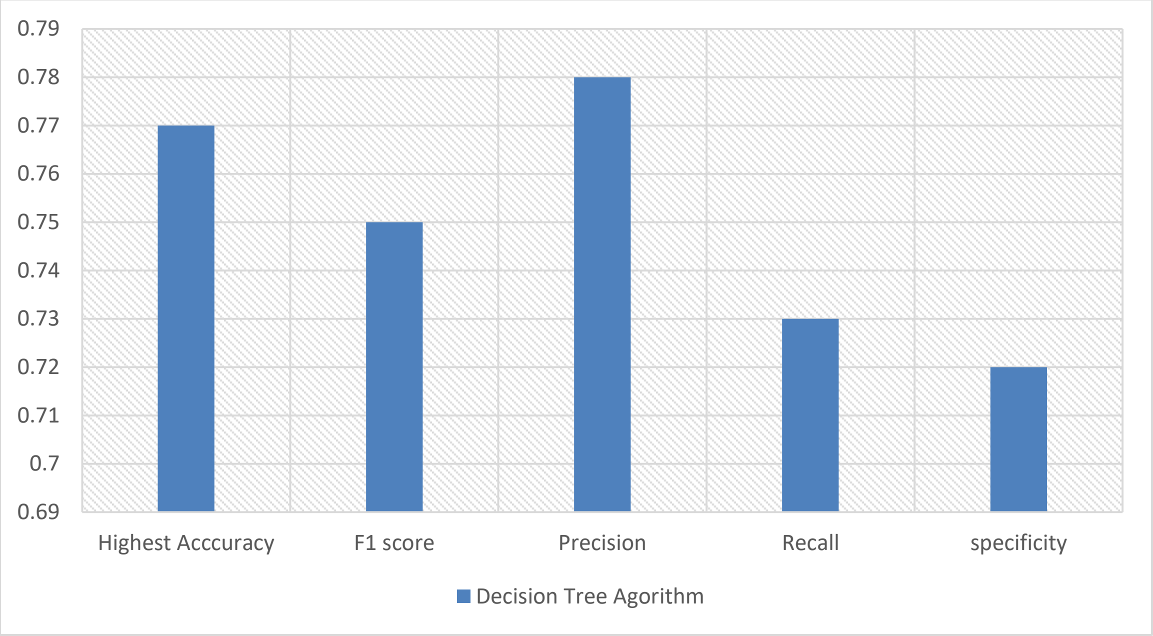


Figure 4.2: Different Score comparison graph of Decision Tree.

4.2.3 SVM

"Support Vector Machine" (SVM) is an algorithm of master training that can be used both for classification and regression. It is found primarily in classification questions, though. In the SVM algorithm, each data object is labelled with the value of a basic teamwork as a point in the n-dimensional space Then we classify the plane that differentiates very well between the two groups.

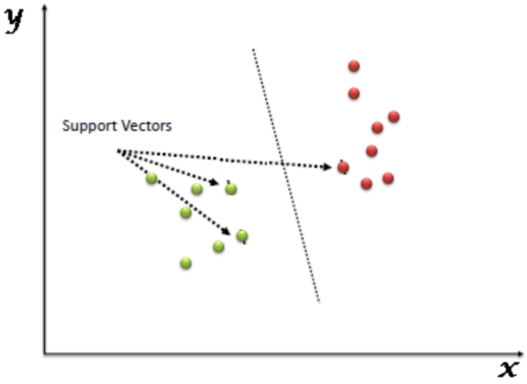


Figure 4.3 Support Vector

Help vectors are essentially independent observation coordinates Figure4.3. The SVM classification is a boundary that divides the two groups most.

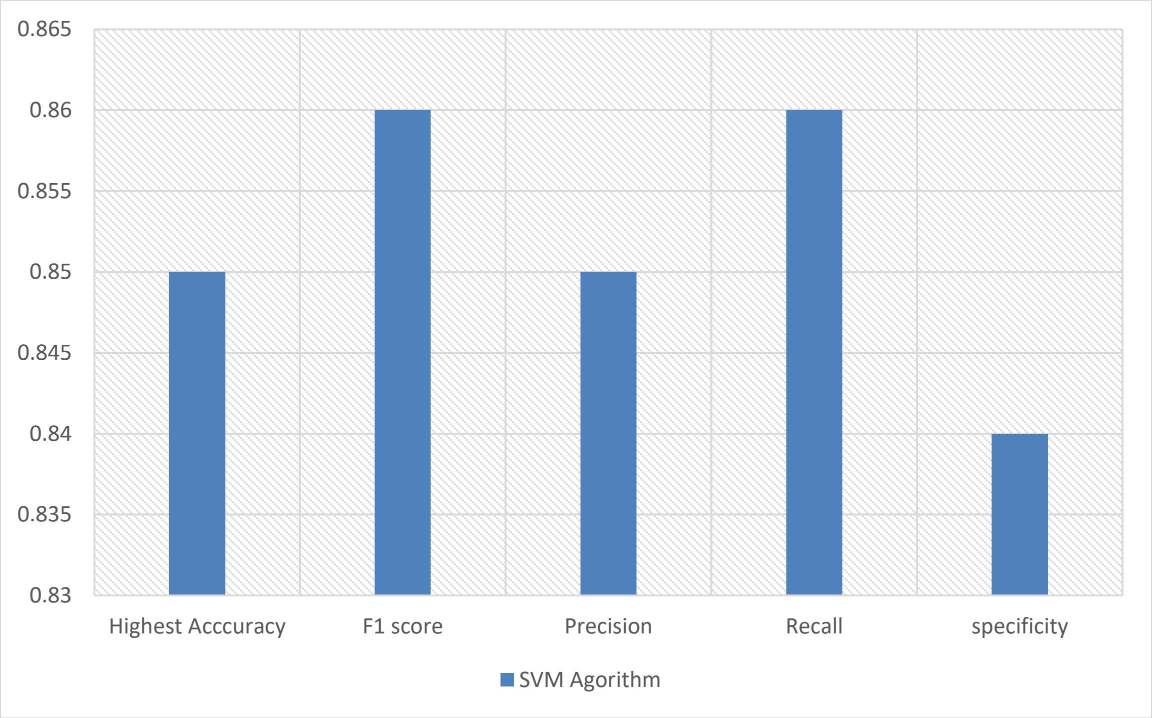


Figure 4.4: Different Score comparison graph of SVM

We found best performance from SVM algorithm. The highest accuracy was 85% and other scores are very close to accuracy score. All of score matrix graphically shown in Figure 4.4

4.2.4 Random Forest

Random forest is a versatile, easily used algorithm which produces a great result most of the time, even without the use of hyper parameters. It is also one of the most popular algorithms since it is simple and diverse (it can be used for both classification and regression tasks). In this article, we'll find out how the RFAl works, what varies from and how it is used by other algorithms. It constructs a "forest" with decision-making trees, normally trained in "sacking." The basic principle of the box approach is that the final result is improved by a mixture of learning models. Random forest can be used both classification and regression task. In our classification task random forest generates 80.20% accuracy and 0.84 precision rate represented in figure 4.5

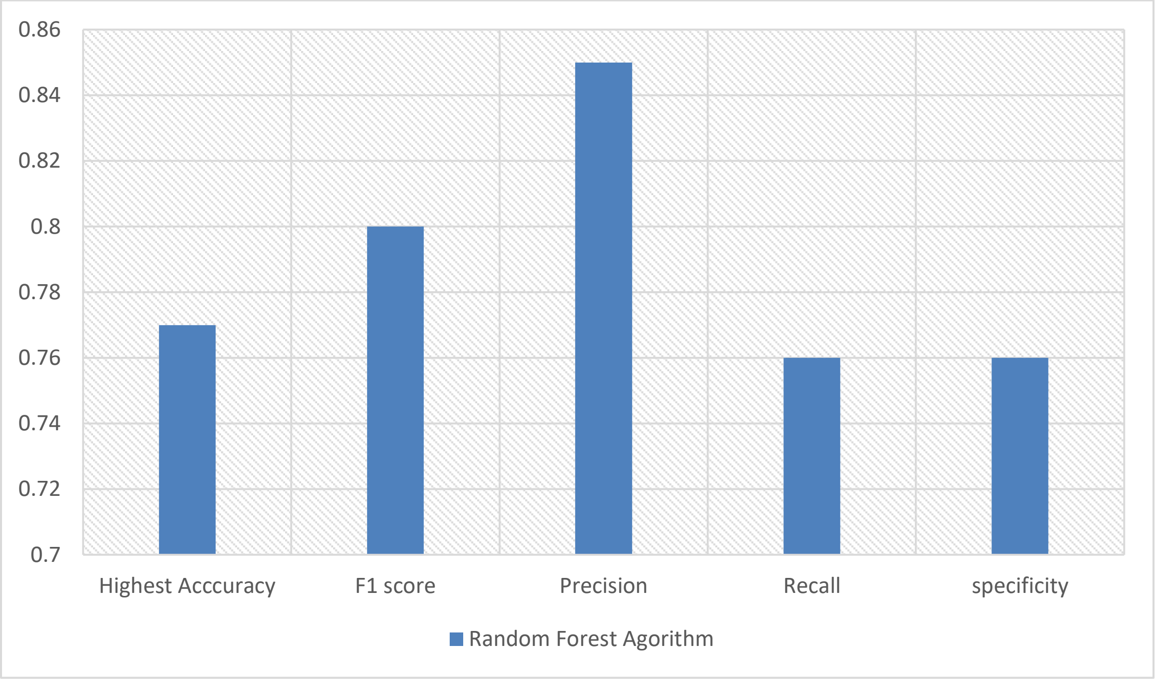


Figure 4.5: Random Forest Score Comparison

4.2.4 Logistic Regression

Logistic regression is one of the most common Machine Learning algorithms, which falls under the Supervised Learning methodology. It is used for estimating the categorical dependent variable using a given set of independent variables. Logistic regression forecast the output of a categorical variable dependent. The result must either be categorical or discreet. This may be Yes or No, 0 or 1, true or false, etc., but it has probabilistic values between 0 and 1. in place of providing exactly the value 0 and 1. Logistic regression is much like a linear regression, except how it is used. Linear regression is used to solve problems of regression, while logistic regression is used to solve problems of classification [12]. The highest accuracy and precision rate is same in our work and the ratio is 0.84 that is graphically represented in figure 4.6

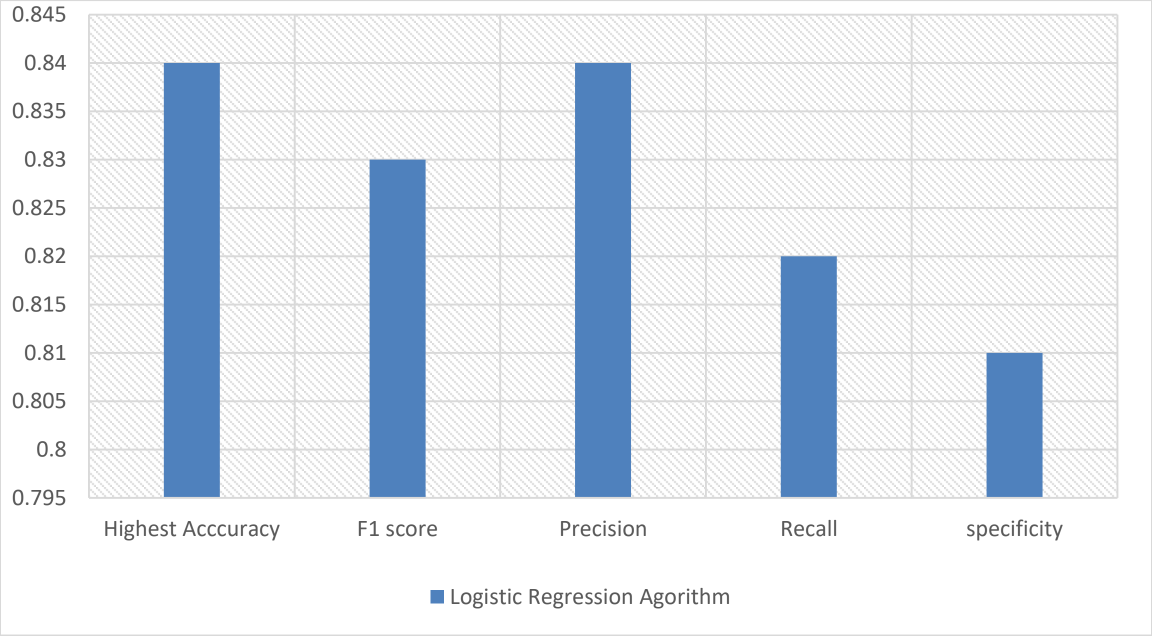


Figure 4.6: Different Score comparison graph of Logistic Regression

CHAPTER 5

SUMMARY, CONCLUSION AND FUTURE WORK

5.1 Summary of the Study

Much research has been carried out with regard to machine learning, but the amount of research carried out is very limited in Bangladesh. Though work in predictive styles are a popular term for computer education, Bangladeshi Book still doesn't know this. Recently, this kind of study is being carried out as the result of those jobs causes a drastic change in our machine life. We have some fascinating real-life applications to the advantage of this kind of research work. But in the field of the Bangladesh economy there is not much research being done. It is our hope, however, that a number of researchers in this field have done research from various countries.

5.2 Conclusion

Our aim is to evaluate 85.01% accuracy feelings in book reviews. The exactness of the SVM algorithm has been discovered. SVM stood out with the best perforation, which enhanced its accuracy against other common algorithms, including KNN, Logistic, Decision Tree, and Random Forest. We have been working with 1357 Bangladesh book reviews from famous online bookshops. We can verify with our proposed model whether a comment on an online bookstore is negative or optimistic, depending on the feeling of the comment itself in Bangladesh. Both the bookshop owners and shoppers can find out which book wants or not to be looked at, and potential buyers can see which book provides positive or bad character. This approach helps bookshop owners. This period will surely alter the experiences of librarians and book readers.

5.3 Recommendations

There are a few remarkable suggestions for this:

- To improve data collection reliability in order to achieve better results from this research.
- In this work the amount of data is very less. To achieved better result, Data must be at least 3 million.
- It would be better if Deep Learning is used.

5.4 Future Work

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

- We want to explore in Bangladesh the feeling of sarcastic comment.
- The emotion that is expressed in the summary segment seems to be typical in the presence of sarcasm. But sarcasm is not easy to foresee on a computer, so in a specific analysis such statement has a contrary feeling. So In future we will develop a system that can also detect sarcasm type sentences.
- We now work on a Web-based API to define the feelings of analysis to accomplish this achievement.
- We performed our work by machine learning algorithm. In near future we will create an intelligence system based on deep learning techniques.
- Our work is based on only Bangla sentences. But customer also post Banglish comment. So in Near future we also train Banglish Sentence that our system can detect Banglish comment.

REFERENCE

[1] (2020). Literature of Bangladesh, culture. Available: https://www.bangladesh.com/culture/literature/

[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 us-ing convolutional neural network," 2017 20th International Conference of Computer and In-formation 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, In-dia, 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 discrimina-tion. 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 dif-ferent NLP toolkits in formal and social media text," in 5th Symposium on Languages, Ap-plications and Technologies (SLATE'16), 2016: Schloss Dagstuhl-Leibniz-Zentrum fuer In-formatik.

[11] J. M. Keller, M. R. Gray, J. A. J. I. t. o. s. Givens, man,, and cybernetics, "A fuzzy k-nearest neighbor algorithm," no. 4, pp. 580-585, 1985.

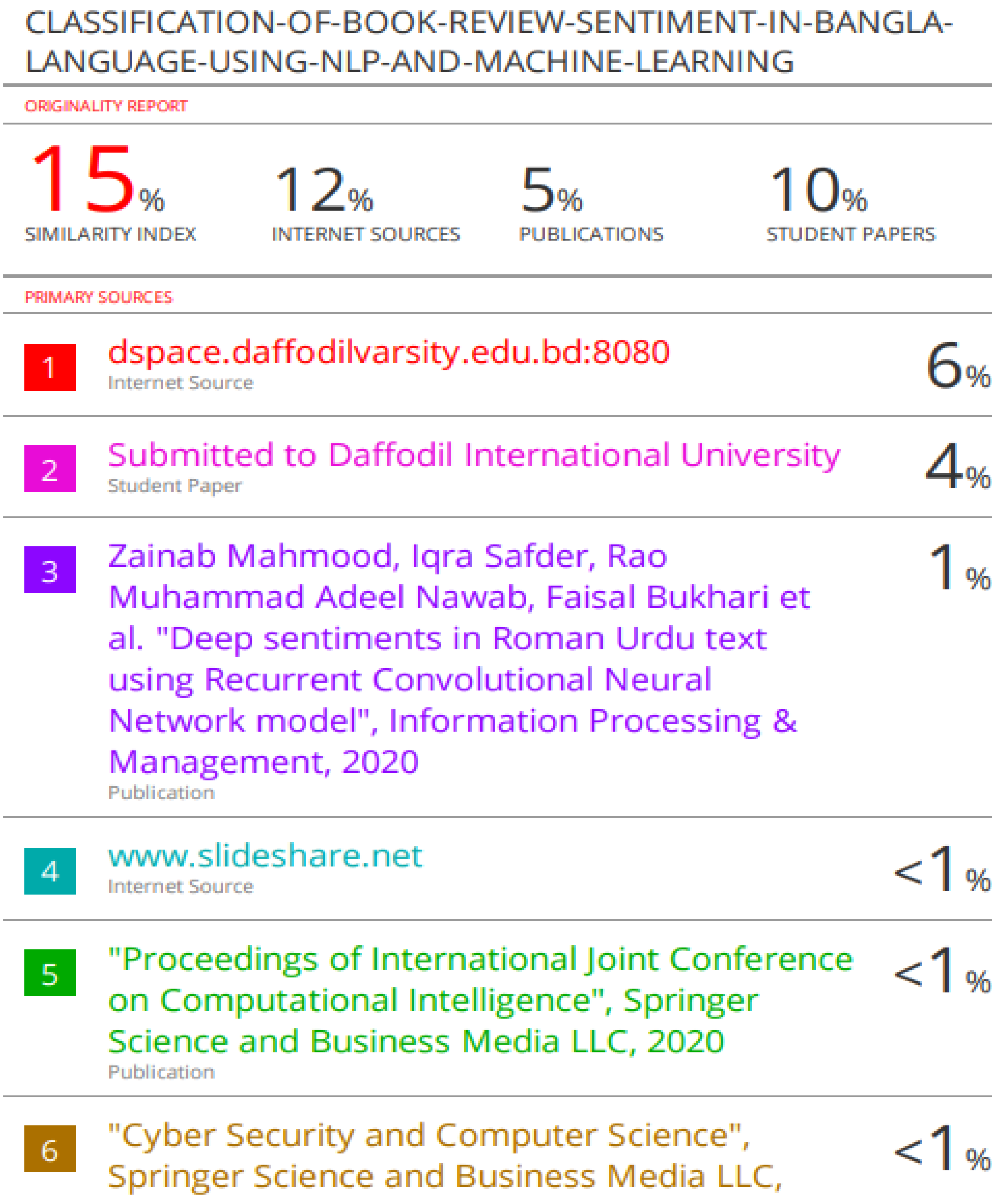
[12] S. R. Safavian, D. J. I. t. o. s. Landgrebe, man,, and cybernetics, "A survey of decision tree classifier methodology," vol. 21, no. 3, pp. 660-674, 1991.

[13] ‘Logistic Regression in Machine Learning’ online available: https://www.javatpoint.com/logistic-regression-in-machine-learning

APPENDIX

The first was to describe the methods for the analysis we faced with so many challenges. The first one was the report. In addition, nothing was achieved in this area before. In fact. It wasn't traditional work. We couldn't get that much assistance from anywhere. Another obstacle was the gathering of data, which was a massive challenge for us. We could not find an open source Bangladesh text pre-processing tool, so we developed a data collection corpus. We have started to collect data manually. In addition, it is another challenge to classify the different posts. After a long period of hard working, we could do it.

PLAGIARISM REPORT



CLASSIFICATION-OF-BOOK-REVIEW-SENTIMENT-IN-BANGLA-LANGUAGE-USING-NLP-AND-MACHINE-LEARNING

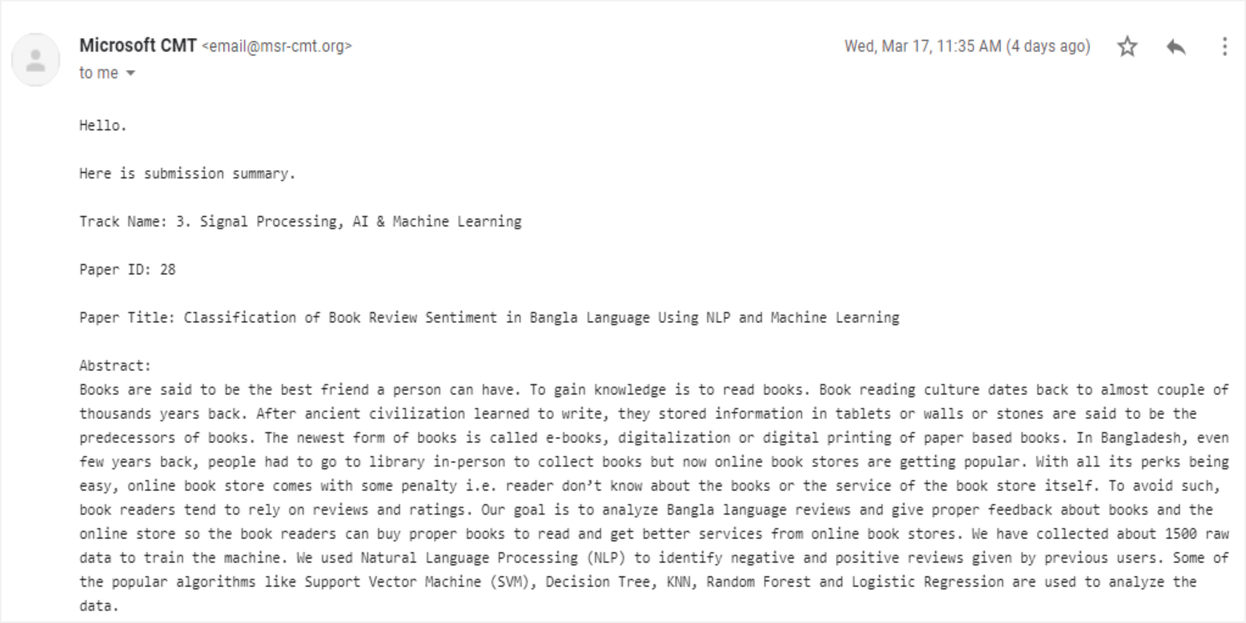
ORIGINALITY REPORT

15%	12%	5%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to Daffodil International University Student Paper	4%
3	Zainab Mahmood, Iqra Safder, Rao Muhammad Adeel Nawab, Faisal Bukhari et al. "Deep sentiments in Roman Urdu text using Recurrent Convolutional Neural Network model", Information Processing & Management, 2020 Publication	1%
4	www.slideshare.net Internet Source	<1%
5	"Proceedings of International Joint Conference on Computational Intelligence", Springer Science and Business Media LLC, 2020 Publication	<1%
6	"Cyber Security and Computer Science", Springer Science and Business Media LLC,	<1%

CONFERENCE SUBMISSION EMAIL



Microsoft CMT <email@msr-cmt.org>

to me

Wed, Mar 17, 11:35 AM (4 days ago)

Hello.

Here is submission summary.

Track Name: 3. Signal Processing, AI & Machine Learning

Paper ID: 28

Paper Title: Classification of Book Review Sentiment in Bangla Language Using NLP and Machine Learning

Abstract:

Books are said to be the best friend a person can have. To gain knowledge is to read books. Book reading culture dates back to almost couple of thousands years back. After ancient civilization learned to write, they stored information in tablets or walls or stones are said to be the predecessors of books. The newest form of books is called e-books, digitalization or digital printing of paper based books. In Bangladesh, even few years back, people had to go to library in-person to collect books but now online book stores are getting popular. With all its perks being easy, online book store comes with some penalty i.e. reader don’t know about the books or the service of the book store itself. To avoid such, book readers tend to rely on reviews and ratings. Our goal is to analyze Bangla language reviews and give proper feedback about books and the online store so the book readers can buy proper books to read and get better services from online book stores. We have collected about 1500 raw data to train the machine. We used Natural Language Processing (NLP) to identify negative and positive reviews given by previous users. Some of the popular algorithms like Support Vector Machine (SVM), Decision Tree, KNN, Random Forest and Logistic Regression are used to analyze the data.