BOOK REVIEW SENTIMENT ANALYSIS BY NLP AND MACHINE LEARNING IN BANGLA LANGUAGE

BY
Mosiour Rahman Sourav

ID: 181-15-11105

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

Supervised By

Ms. Nazmun Nessa Moon
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Sanzidul Islam
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

10 SEPTEMBER 2022

APPROVAL

This Project titled **"Book Review Sentiment Analysis By NLP And Machine Learning in Bangla Language"**, submitted by Mosiour Rahman Sourav, ID No: 181-15-11105 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 Bachelor of Science in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 14th September , 2022.

BOARD OF EXAMINERS

18/09/22

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

Al Amin Biswas — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

14.09.22

Mushfiqur Rahman (MUR) — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

14.9.22

Dr. Md Sazzadur Rahman — **External Examiner**
Associate Professor
Institute of Information Technology
Jahangirnagar University

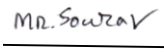
DECLARATION

We hereby declare that this thesis has been done by us under the supervision of **Ms. Nazmun Nessa Moon, Associate Professor,** Department of CSE, and co-supervision of **Md. Sanzidul Islam**, Lecturer, **Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Ms. Nazmun Nessa Moon
Associate Professor
Department of CSE
Daffodil International University

Submitted by:



Mosiour Rahman Sourav
ID: 181-15-11105
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. Nazmun Nessa Moon, Associate Professor, 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

Due to good consumer comments and reviews throughout the web, sentiment polarity detection has lately piqued the interest of NLP experts. The continued growth of e-commerce sites raises the purchasing rate of diverse items. People's interest in literature, for example, is fast increasing. Bangladesh already has a strong online marketing and e-commerce sector in this age of internet technology. Online product reviews, for example, have become a vital source of information for buyers making purchase decisions. It is believed that a person's best friend is a book. Books are essential to every person's existence because they provide knowledge about the outside world, help improve reading, writing and speaking abilities, and strengthen memory and intelligence. Our purpose is to rank Bangladeshi reviews and give accurate information about books and online bookstores in order to assist book lovers in purchasing the proper books and locating better online retailers. Using machine learning and natural language processing, this article demonstrates how to extract the sentiment polarity (positive or negative) from Bengali book reviews (NLP). We used five classification algorithm like: Multinomial Naïve Bayes(MNB), K-Nearest Neighbor (KNN), Random Forest Tree (RFT), Support Vector Classifier (SVC), and Stochastic Gradient Descent(SGD).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

CHAPTER 1

INTRODUCTION

1.1 Introduction

Sentiment analysis, also known as information extraction, is a field of study that collects people's emotions, opinions, and sentiments in order to forecast the polarity of text data in public opinion or microblogging sites. [2] on well-known subjects. It is now relatively simple to connect to the internet in Bangladesh. Before making a purchase decision, new buyers want to read previously published product reviews. Sentiment detection is a technique for determining a user's viewpoint on a certain issue. Textual information in the form of tweets, reviews, comments, postings, or bulletins can be assigned a positive, neutral, or negative polarity. Sentiment detection is a technique for determining a user's point of view on a specific issue. Assign a favorable, neutral, or negative polarity to textual information tweets, reviews, comments. People's buying preferences have shifted considerably in recent years, with books being one of the most popular things available online.

In Bangladesh, we may buy Rokomari, Boibazar, Boikhata, eBoighar, Daraz, Bookshopbd, Boi-BoiBoi, and backpacks online. When advancements in technology bring people closer together, it becomes easier to deceive or lie on the Internet. People are cautious to buy books after viewing bookshop advertisements on the internet. We also pay attention to previous consumer comments and ratings regarding these items and service providers. Bangladeshi people prefer Bangladesh as a language to learn more often. I don't care because it's my mother's guidance. Reviews are very important in the broader Internet economy. It aims to use machine learning algorithms to score and identify whether a particular book received negative or positive comments on a ratio basis. Sentiment analysis is a key feature of NLP, so more researchers than ever before are paying attention to sentiment analysis. Create a data set of 1600 Bengali book reviews categorized into positive and negative attitudes. Then We use Multinomial Naïve Bayes(MNB), K-Nearest Neighbor (KNN), Random Forest Tree (RFT), Support Vector Classifier (SVC), and Stochastic Gradient Descent(SGD).

1.2 Motivation

The phrase "internet bookshop" has grown popular in Bangladesh. As the number of Internet users grows, so does the number of clients for online bookshops. Technology has brought the entire world closer together. E-commerce has underpinned the rapid expansion of online bookstores. Today, people don't want to waste time going to physical bookstores. Instead, they seek convenience in their lives and want to live as simply as possible. Additionally, online bookstores have made it easier for customers to obtain both e-books and physical books. Customers merely require a few clicks to place a purchase. Book reviews are another option. It is an evaluation of a document, case, thing, or phenomena. Books, articles, any genre or industry, architecture, sculpture, design, restaurants, politics, exhibits, performances, and many more can be reviewed. This lecture will concentrate on book reviews. Before purchasing the book, the majority of consumers read the review. A book review is a good way to get a broad sense of a book. It is, however, a time-consuming technique. It may also be tiresome for a consumer to read each and every comment. The previous explanation shows that something must be done to handle this issue in a way that saves individuals time while also allowing them to finish their assignment. Finally, we've chosen to use NLP and machine learning to solve this challenge. Algorithms, as we all know, do not understand strings directly. First, we must transform the string to numerical representation. We utilized the TFIDF approach in this example. A Machine Learning technique was used to classify each remark. We utilized a different set of parameters for each method. And we picked these parameters since they produced the best outcomes.

1.3 Problem Definition

For more than a decade, people have used the internet in their homes. The internet revolution has affected individuals of all ages, from elderly to children, veterans to trainees; everyone has their own way of learning the method and applying it to their own requirements. In terms of entertainment, envisioning, purchasing, researching, educating, and gaming, the internet excels all previous types of media. The Internet has grown into the most convenient and cost-effective way to connect to the global network. Attractive advertising, live videos, streamlined operations, and other elements have been introduced. The internet has grown into an excellent marketing and sales tool. The internet has become the new product sales catalog for retail firms. People are becoming more at comfortable with purchasing books from online retailers. This research is being carried out in order to save time and provide the finest book for consumers. We employed Natural Language Processing and Machine Learning in this project. As a result of this endeavor, a number of challenges have occurred. Data collection is a particularly difficult effort for our study since we are dealing with human emotions. We got information by visiting several book-selling websites. And I collected every single comment from each and every book. We received both negative and positive responses. This is the information we utilize as a feature. Approximately 1600 comments in Bangla were collected. This is our raw data, which includes a lot of noise like double words, extra punctuation marks, and emoji. During the preprocessing stage, we removed all of this noise so that our algorithm could learn properly. Following preprocessing, we used the TFIDF approach to convert the string to numerical representation. We used numerous Classification Machine Learning algorithms following the competition of establishing numeric formats because our job is classification-based. Each sentence is separated into two groups: positive and negative. When the training state is completed, we evaluate our work by obtaining original data that has not been trained. In the evaluation process, our technique outperforms others. Each level was represented by its own graph.

1.4 Research Questions

- What procedures are employed in the collection and processing of datasets?
- Can machine learning approaches predict positive and negative classifications correctly?
- Can you appropriately identify positive and negative groups?
- Are there any vacancies available online?
- What advantages does this position provide?

1.5 Research Methodology

This section will go through our workflow, which consists of data processing, information processing, data classification, and algorithm implementation. Model training and algorithm evaluation.

1.6 Research Objectives

- To develop a model that can distinguish between good and negative comments.
- Using or classifying such categorization procedures to anatomize consumer analysis.
- Create a pricing software application using engineering tools and machine learning.
- Conduct research to demonstrate a scientific notion.

1.7 Research Layout

Chapter 1: will cover the following topics: introduction, motivation, problem definition, research question, research methods, and our predicted study result. We also explain why we selected to perform this study in this chapter.

Chpater 2: The second chapter will look at the history of this study, as well as related studies and the present state of affairs in Bangladesh. It includes a contextual analysis as well as a quick overview of the work.
Chapter 3: will make the investigation plan clear This chapter delves deeply into the approach or procedure. This section will show how the thought process was carried out.

Chapter 4: This chapter will walk you through the process of putting on the suggested show using a precision table and an exploratory outcome report.

Chapter 5: This chapter is located at the end of the report. This section summarizes the demonstration's execution. This section also includes a precision comparison. This part also investigates the show's internet usage and yield.

1.8 Expected Outcome

• We will differentiate between negative and positive client comments

• We will save the client's time.

• We will strive to display the best book based on the client's choices.

• We developed a useful online application that displays the results of each book review remark.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

Several machine learning prediction methods have been studied. Prediction is one of the most common applications of machine learning. Several studies have been conducted on sentiment analysis. These studies focused on specific challenges and employed a range of machine learning approaches to solve them. This chapter highlights the steps that various experts in the field have successfully accomplished in the past.

2.2 Related Works

Nowadays, almost everything is web-based. People express their views over the internet. People's emotions are regularly detected with the researcher magnet. This subject was presented in a number of settings and languages.

Mittal et al. [4] proposed an evaluation procedure for Hindi that yields 82.89 percent positive and 76.59 percent negative legitimacy. They chose to degree feelings and broaden the database's scope in arrange to move forward its consistency. This article highlights a program that explores Roman Urdu people's feelings through sports, computer program, food and formulas, theater, and legislative issues. It contains 10,021 sentences separated from 566 web discourses. This program has two objectives: (1) creating a human-annotated corpus for Roman Urdu enthusiastic investigation, and (2) exploring feeling examination approaches based on Rule-based, N-gram (RCNN) models.

Chowdhury et al. [5] created a mechanism that automatically removed persons from the Bangla language network, whether they were favorable or negative. In its recommended technique, SVM performed 93% with unique features from 1300 col-selected data. Sentiment Analysis (SA) is a technique for merging feelings, ideas, and linguistic subjectivities. SA is now the most demanding natural language processing challenge. Social networking services like Facebook are commonly utilized to broadcast several points of view on a single live unit. A reader commented on facts published in a newspaper

about an occurrence. Every day, the volume of feedback from online transactions grows. As a result, people's happiness levels are highly impacted by their judgements and opinions.

Aspect-based Opinion Examination may be a sort of assumption examination that looks at how individuals feel approximately a certain issue. Rahman et al. [3] utilized this approach to do investigate in Bangladesh. Estimation examination is progressing in Bengali and is right now respected as a major inquire about issue. Information collecting, corporate dialect investigation, lexicon as portion of the voice tagger, and other Bengali work are troublesome due to a need of assets. Their essential objective was a eatery audit and the application of aspect-based investigate to get cricket perspectives. SVM has the most noteworthy legitimacy for extricating and recognizing extremity in creepy crawlies and eateries, with 71 and 77 percent, individually.

Understanding customer preferences is critical in online purchasing, but firms may not be as well-versed in this area as they may be. In order to validate their judgments, C. Chauhan et al. [7] used machine learning algorithms to distinguish between negative and positive comments from potential clients. They examined a number of publications and determined that Nave Bayes gave good results, although the results differed depending on the environment, strategy, and goals.

This network and its variants have recently exhibited remarkable performance in a range of downstream natural language processing applications, particularly in resource-rich languages such as English. However, when it comes to Bangladesh's categorization issues, these alternatives have not been well researched. Bangladesh is developing its multilingual text classification transformer model. To characterize text-based emotions resulting from Bangla analysis, Alam et al. [6] developed a Convolution Neural Network model (CNN). CNN achieves 99.87 percent accuracy with 850 data points, 350 of which were negative and 500 of which were positive.

Tuhin et al. [8] offered two ways for identifying and detecting different forms of emotion in Bangladesh. These people were thrilled, indignant, sad, afraid, enthusiastic, and sensitive. Both the topical solution and the method of grouping are approaches in Nave Bayes. A topical method with 90% accuracy was used on a 7400 Bangladesh phrase data

set. They then compared their paper to two others that had an SVM score of 93% and a document frequency score of 83%, respectively. Each of the three compositions possesses a distinct emotional aspect.

2.3 Comparison of related work

Table 2.1 Comparison Table of related work

RELATED WORK	ACCURACY RATE
Using sentiment analysis to analyze Bangla microblog posts[2]	93%
Aspect-Based Sentiment Analysis Datasets in Bangla and Their Baseline Evaluation. [3]	77%
Sentiment analysis using a convolutional neural network for Bangla utterances[6]	99.87%
An Automated Sentiment Analysis System from Bangla Text was created using Supervised Learning Techniques.[8]	75%
Using machine learning techniques to analyze sentiment in the Google Play store Reviews in Bangla[7]	76.48%
Analyzing the Sentiment of Movie Reviews in Bangla Using Machine Learning Techniques[1]	88.90%

Based on the discussion above table 2.1, we concluded that there was no significant book review activity in Bangladesh. When the two studies are compared, we can see that our model has a larger dataset, is more accurate, and has done well in a variety of fields. Our content might be used on a web-based platform.

2.4 Research Summary

The aforementioned study was conducted by a number of research firms, highlighting the scope of emotional analytics research. As a result of our analysis, we have effective results. Despite a limitation of resources, each sector aims to become more resourceful by providing information on how to obtain numerous products in a single day.

2.5 Challenges

The most difficult aspect of the endeavor is planning the data sets for subsequent processing. We adjusted the data set for our work or future processing using highly useable ML tools. Another issue in Bangladesh is the struggle 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

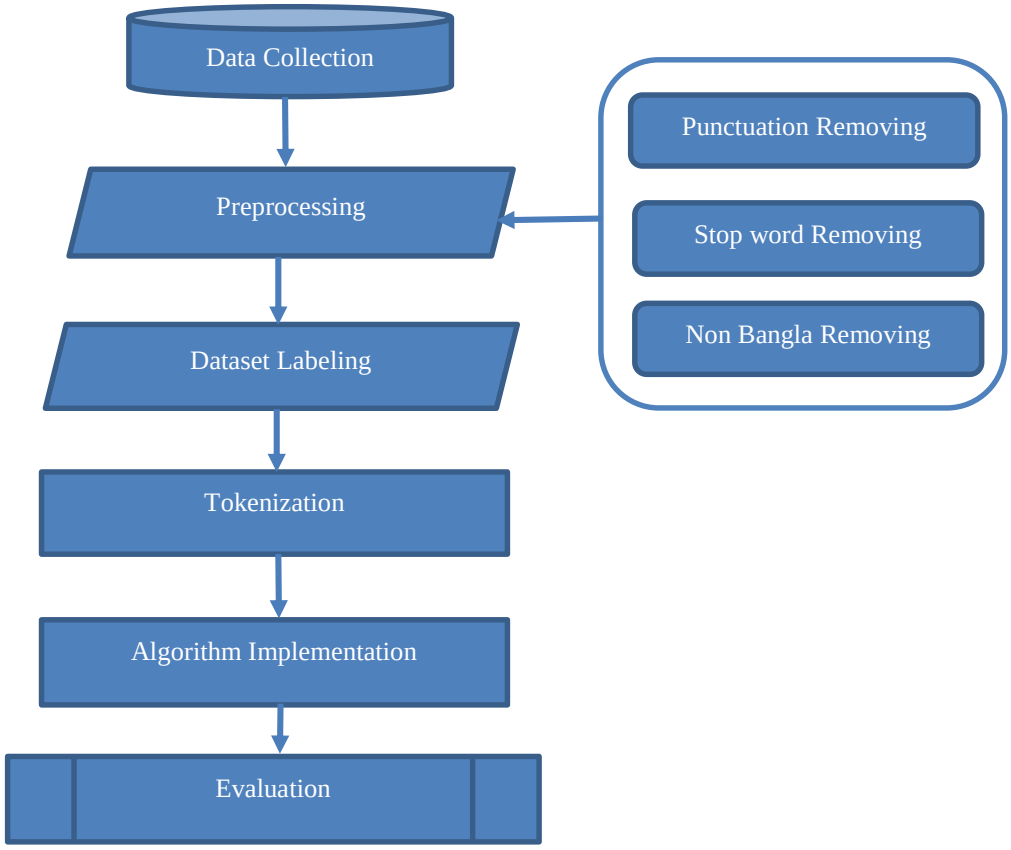


Figure 3.1: Methodology diagram

3.2 Data Collection

Each study begins with data collection. The success of any book depends on the delicate bit of information known as a book review. Additionally, we need to receive our knowledge from a reliable source. Our study's data was derived from comments made by book

reviewers. This was gathered from a variety of book retailer websites and book review pages and produced for Facebook. Only comments in Bangla were gathered since it was required of us to do so. 1600 data are collected for this study.

3.3 Data Pre-Processing

Data preprocessing is a data mining technique that converts raw data into an efficient and practical structure. Information preparation is crucial for knowledge acquisition. Our function is created on KDD. According to Kamiran et al. [9], the four most important data pre-processing methods are elimination, data massaging, weighting, and Same poling. In our effort, we created accessible data sets primarily using data messaging techniques. And prunctuation removing stop word removing and non bangla removeing is the main part of our preprocessing. We eliminated unnecessary words and points from the Bangla stop for this level. We've decided to use our revised input as a feature while putting all of the procedures into action. Figure 3.2 represents out preprocessing steps.

Figure 3.2: Data preprocessing steps

3.4 Dataset Labeling

Information was divided into two categories for us, positive and negative. When designing the courses, the user's emotions are taken into account. If the novel's analysis is solid, this line will score well. For unfavorable ratings, there are several categories. The information in figure 3.3 was acquired in this manner. We collected 1600 reviews, and 56.5% of them were positive, while 43.5% were negative.

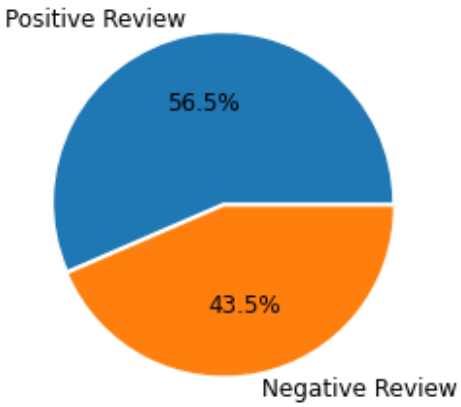


Figure 3.3: Classification

3.5 Tokenization

Tokenization is a method for segmenting flag phrases, which may consist of words or signals, according to Pinto et al. [10]. Our database includes a sizable quantity of phrases. To achieve our goal, we decide to employ a phrase mark rather than a word label. Additionally crucial is tokenization. Tokenization separates our statement into its component words. Table 3.1 depicts the tokenization process.

Table 3.1 Tokenization Table

Raw Data	**Type**	**Tokenized data**
যতসব অশালীন ভাষা	**Negative**	‘যতসব’ ,‘অশালীন’ ‘ভাষা’
সবাই পড়ে দেখতে পারেন অতি মাত্রায় শিক্ষণীয়	**Positive**	‘সবাই’, ‘পড়ে’, ‘দেখতে’, ‘পারেন’, ‘অতি’, ‘মাত্রায়’ ‘শিক্ষণীয়’
অবাক করার মতো গল্পের প্রেক্ষাপট অনেক সুন্দর	**Positive**	‘অবাক, ‘করার, ‘মতো’ ,‘গল্পের’ ‘প্রেক্ষাপট’ ‘অনেক’ ,‘সুন্দর

3.6 Algorithm Implementation

In this section, we covered the implementation process for the algorithm. We must finish the previous step in order to produce the necessary dataset before we can finish this one. We have five different categorization techniques since our work is in the classification form. We use the Adaboost, Decision Tree, SVM, KNN, and Random Forest algorithms as our classifiers. Table 3.2 displays the value that will allow for the greatest accuracy for each approach.

Table 3.2 Parameter usages

Algorithms	Details
Multinomial Naive Bayes	n_informative=3, n_redundant=0, random_state=1, shuffle=True
Decision Tree	random_state=46
SVM	ernel='rbf’
Random Forest	n_estimators=80
KNN	random_state = 42

3.7 Evaluation

We assessed our favored RF procedure utilizing real-time information estimation and an instability framework. We at first obtained 80 honest to goodness information focuses from which we demonstrated fizzled to memorize. For each of the classes chosen, diverse pages of online book deals websites and Facebook Bangla book surveys were utilized.

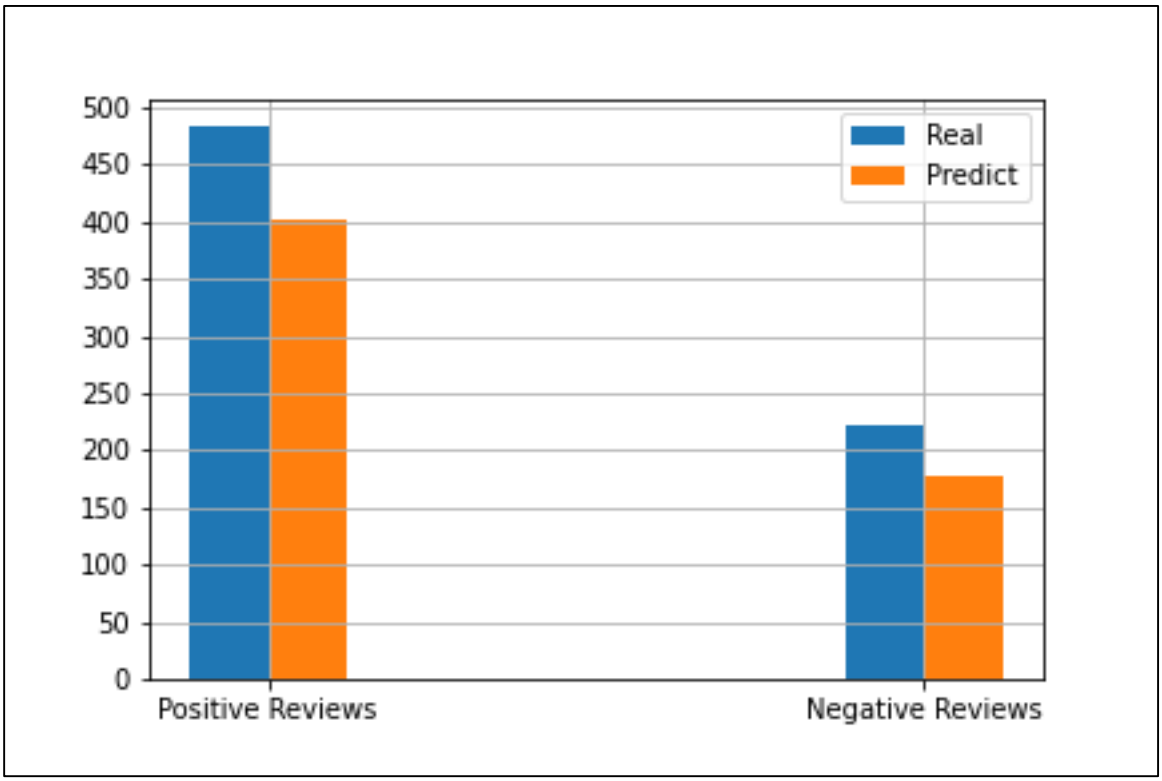


Figure 3.4: Evaluation

A comparison of the actual and expected outcomes is shown in Figure 3.4. In our dataset, there are 225 bad ratings and 480 favorable reviews, which are shown by blue bars. The expected value is shown by the Orange color bar. Our algorithm correctly predicted 178 out of 225 reviews and 402 out of 480 favorable reviews. We may assume that our model performed well with data from the actual world as a consequence. This prediction can also be tested using the confusion matrix.

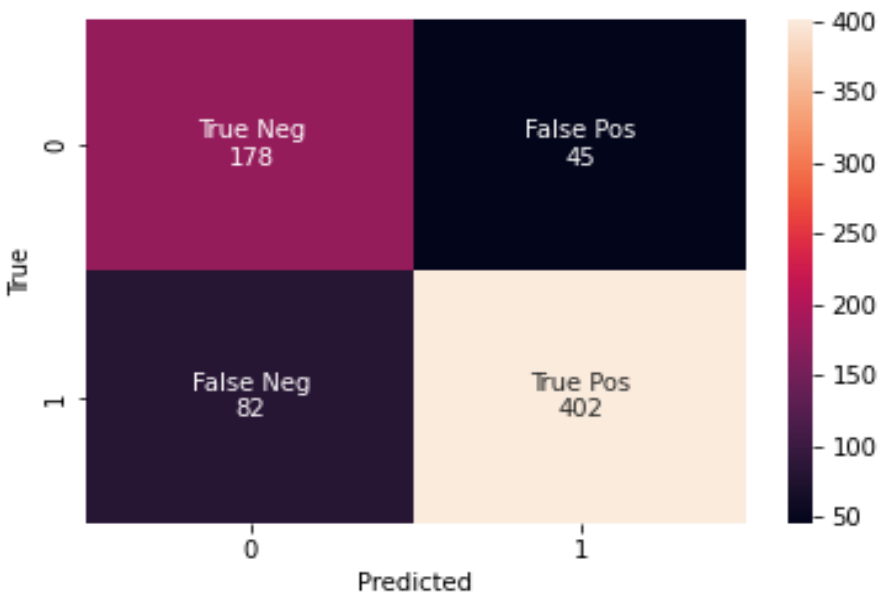


Figure 3.5: Confusion Matrix

$$\text{Accuracy} = \frac{178+402}{148+402+45+82} = 0.8203 * 100$$

$$= 82.03\%$$

Error $= 1 - 0.8203 = 0.17 * 100 = 17\%$

Recall rate for positive:

$$\frac{402}{402+82} = .830 * 100 = 83.0\%$$

Recall rate for Negative: $\frac{178}{178+45} = 790 * 100 = 79.0\%$

We used the Confusion Matrix to find overall results. Figure 3.5 displays the uncertainty matrix for the validation dataset. We found that 82.03% of the test data were accurate during the evaluation process. This also implies that both visible and concealed data may be used with our approach. Positive memory is 83% percent of the total, whereas negative recall is 79% percent. Based on this computation, we may conclude that our algorithm is more effective in predicting good reviews.

CHAPTER 4

RESULT ANALYSIS

4.1 Introduction

This component of analytical research focuses on empirical evidence and test results in general. What is the preliminary end result analysis when we look at a topic? The consequences section should be dependent such that the outcomes are stated without any interpretation or evaluation. The instructions are also available in the instructive papers section. The results are presented, and the test is validated. We also examined a number of algorithms and determined which one was the best out of a group of five. Precision, accuracy, reminder, and f1 were also chosen as criteria for calculating the data.

4.2 Experimental Result

Table 4.1 Accuracy table

Test data usage rate		**20%**	**25%**	**30%**	**35%**
Algorithms Accuracy	MNB	75.31	74.75	71.88	73.04
	KNN	63.44	62.75	59.58	59.29
	SVC	75.94	75.75	74.17	73.75
	RF	79.19	74.75	72.71	73.93
	SGD	75.94	75.00	72.71	72.68

Table 4.1 shows the precision table. for data usage rate we used 20% to 35% test data. when we used 20% test data then training size is 80% same way when test data is 35% then training is 65%. We apply this technique for machine learning algorithm to find which percentage got best accuracy. And we can see from this table the highest accuracy is achieved by Random Forest algorithm. So we decided to use Random Forest for final prediction.

Table 4.2 Different Score Matrix

Score Matrix	**Algorithms**				
	MNB	**KNN**	**SVC**	**Random Forest**	**SGD**
F1 Score	78.57	61.89	79.02	79.56	78.71
Recall	78.14	51.91	79.23	78.69	79.78
Precision	79.01	76.61	78.80	80.00	77.66
Specificity	71.22	55.1	72.06	72.34	71.97

The rating Matrix is shown in table 4.2. We've only gone through 20% of the scoring matrix. Because the precision desk best reflects precision, which is based on genuine positives and true negatives, numerous criteria have been used to analyze the accuracy, including real terrible, fake great, real high quality, and false awful. The RF generated an exceptional exactness verification table in all dimensions, as well as an F1 score, recall, accuracy, and specificity. As a consequence, the RF algorithm was selected as the prediction approach for this study.

4.2.1 Multinomial Naïve Bayes Algorithm

The Multinomial Credulous Bayes run the show set is an critical approach to Bayesian dominance in characteristic dialect handling (NLP). Utilizing Bayes' hypothesis, the computer program gauges the tag of a printed component, such as an mail or daily paper article. Calculates the likelihood of each tag for a given test and returns the tag with the most prominent threat. The Credulous Bayes classification comprises of a few calculations that all have one thing in common: each characteristic that's classified is isolated from each other characteristic. The nearness or nonattendance of a include does not influence whether or not another include is included. Figure 4.1 shows that the leading exactness of the MNB calculation is 75.00 percent, with an exactness of 0.9165.

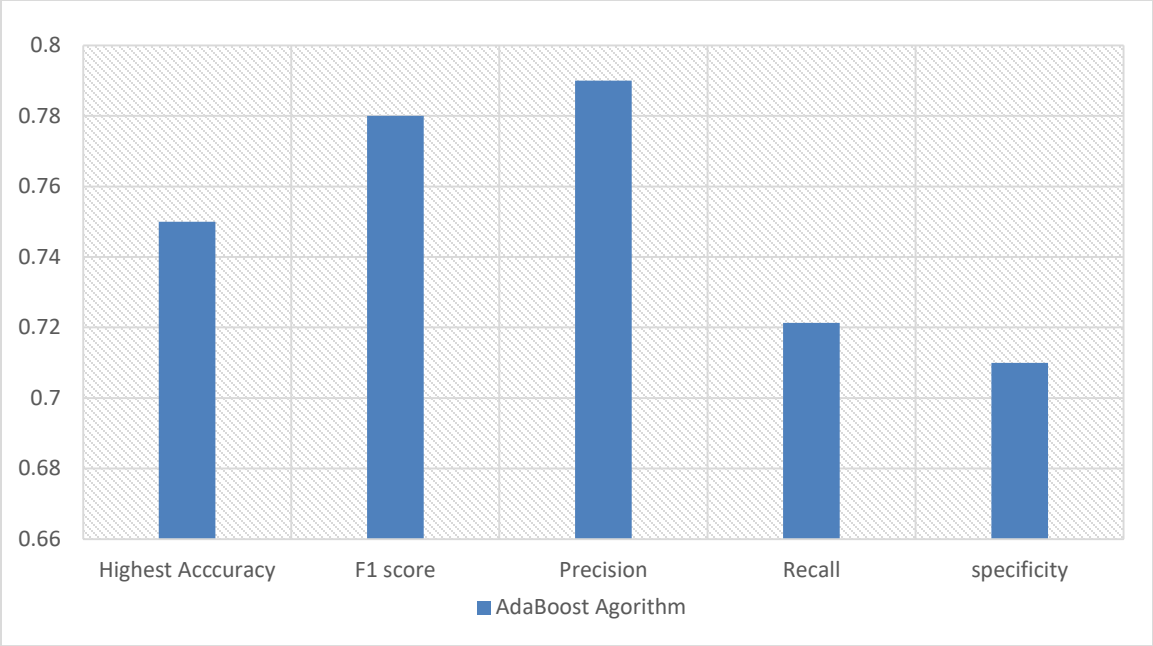


Figure 4.1 Different Score comparison graph of MNB

4.2.2 Stochastic Gradient Descent

A SGD is a comprehension algorithm that accurately separates each node based on neighbor information represented by figure 4.2. One result is that variable splitting can be used to drive more powerful trees. Trees have been shown to be fairly malleable, with only minor interactions. The downside is that "high dispersion" tones are perceived effects. Overfitting is also caused by large discrepancies when the tree's predictions are overly optimistic. Selection trees provide excellent results and work with large data sets. [11] By adjusting the department factors, a more powerful tree can be formed. The term "low distortion" refers to how flexible wood is in its interactions. Manage the spanned words anyway you see fit. decision Tree is a Supervised mastery technique that can be used for both class and regression issues, but it is significantly more commonly employed for type problems. It's a tree-based classifier, with core nodes representing dataset functions, branches representing selection criteria, and each leaf node representing the conclusion. There are nodes in a decision tree, which can be the selection Node or the Leaf Node. Selection nodes are used to make any decision and have a few branches, whereas Leaf nodes represent the result of these decisions and no longer include any

comparable branches. The options or the check are performed based on the characteristics of the provided dataset.

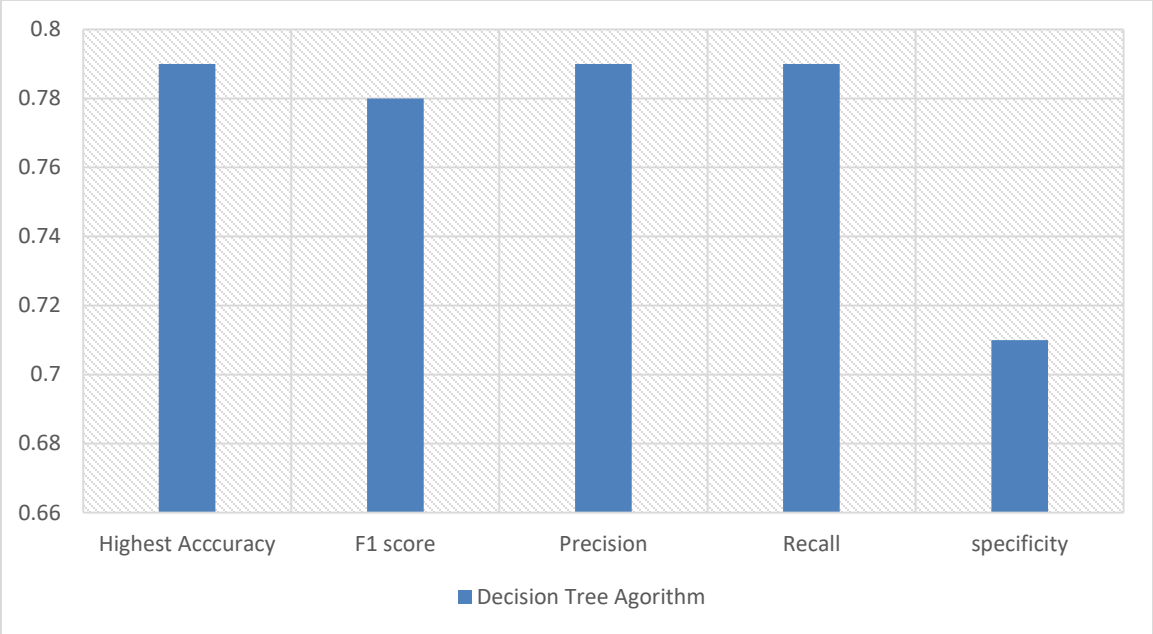


Figure 4.2: Different Score comparison graph of SGD

4.2.3 SVC

Support Vector Classifier are an own family of supervised studying algorithms used for type, regression, and outlier detection. All of those are common machine getting to know responsibilities. you may use them to discover cancerous cells based totally on hundreds of thousands of pictures, or you could estimate future tour patterns the use of a properly-outfitted regression version. One sort of SVM you may use for specific machine gaining knowledge of difficulties is help vector regression (SVR), that is an extension of guide vector classification (SVC). The maximum crucial issue to understand is that they're just arithmetic equations that have been excellent-tuned to offer you the high-quality correct solution as speedy as feasible. SVMs range from different classification algorithms in that they pick out a decision boundary that maximizes the gap among the closest records factors in all training. The choice boundary set up by using SVMs is the maximum margin classifier or most margin hyperplane. Figure 4.3 demonstrates that the best accuracy of the selection tree approach.

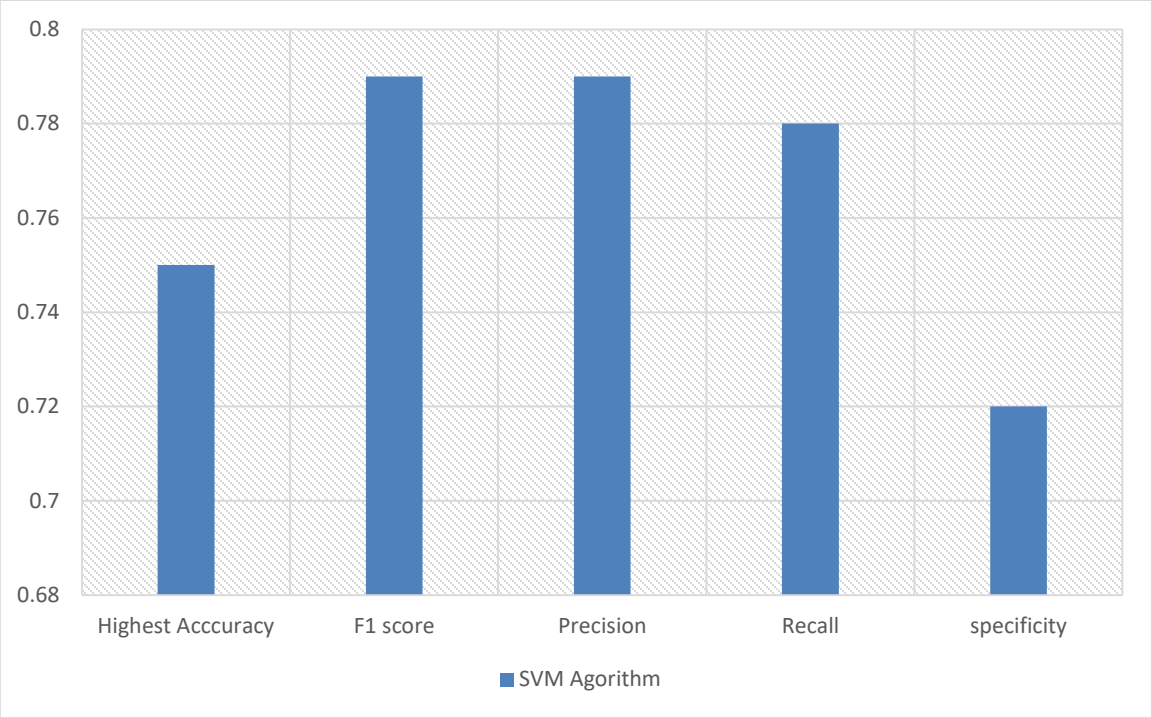


Figure 4.3: Different Score comparison graph of SVC

The SVC approach produced the most effective outcomes. The greatest level of accuracy was 78.80 percent, with the others being quite close. Figure 4.2.3 depicts the entire score matrix graphically.

4.2.4 Random Forest

Random Wooded Area is a flexible and easy-to-use set of rules that provide very good results on the largest instances without hyper parameters. It is also one of the most commonly used algorithms due to its simplicity and flexibility. This post explains how RFAI works, how it differs from other algorithms, and how to use it. Create a "forest" of trees to make choices, as is usually taught in belaying. The most important thing behind field methods is that merging deployment models improves the end result. In addition to categorical regression, random woodlands can be used. In the classification project, Random Woodland achieves 79% accuracy and 80.00% accuracy rate, as shown in Figure 4.4.

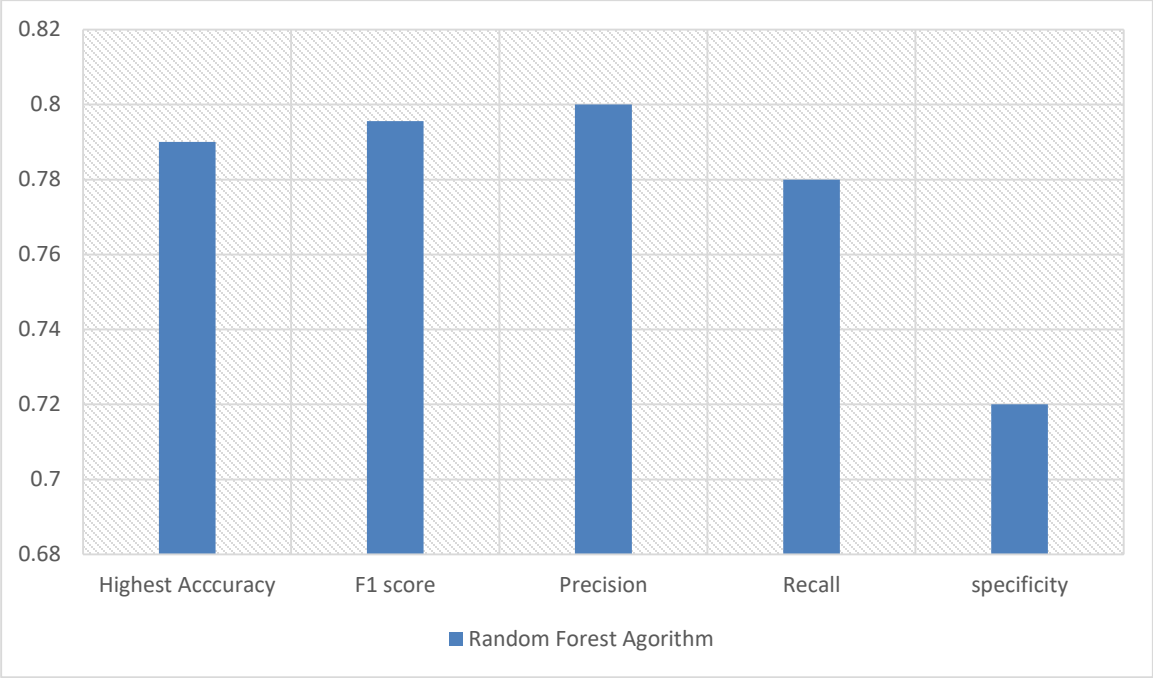


Figure 4.4: Random Forest Score Comparison

4.2.5 KNN

K-Nearest Neighbor is a fundamental Machine Learning method that employs the Supervised Learning method. It can make an assumption about the similarity between the new case/data and the existing cases and place the new case in the most comparable category to the existing categories. It can store all accessible data and identify fresh data points based on similarities. This means that when new data is created, it may be swiftly classified into a suitable category using the K- NN approach. This method may be used for both regression and classification; however, it is more typically utilized for classification problems. Because K-NN is a non-parametric approach, it makes no assumptions about the underlying data. KNN algorithm achieved highest accuracy about 63%. Figure 4.5 represents the knn score graph.

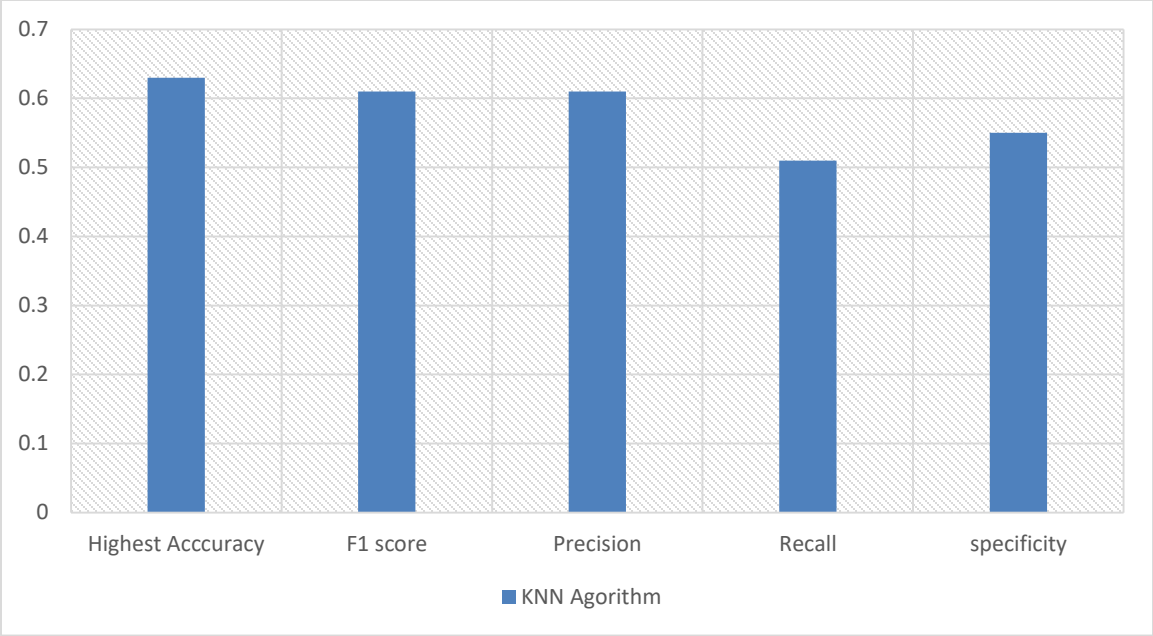


Figure 4.5: Different Score comparison graph of KNN Algorithm

4.3 Score Matrix of Test Dataset of Random Forest Algorithm

Table 4.3 represents the different score matrix of test dataset of random forest algorithm. From this graph we can see that the negative precision rete is 0.68 and positive precision is 0.90 and the testing accuracy of vgg16 got the accuracy of 0.82 for testing data. for macro average we got 0.79. The second highest value is 86 which is belongs to f1 score column.

Table 4.3 Different Score Matrix

	Precision	**Recall**	**F1-Score**	**Support**
Negative	0.68	0.80	0.74	223
Positive	0.90	0.83	0.86	484
Accuracy			0.82	707
Macro Avg	0.79	0.81	0.80	707
Weighted Avg	0.83	0.82	0. 82	707

4.4 Model testing

For the final evaluation of model, we collected total 707 positive and negative sentences these are never seen by our model. For 485 of positive reviews 400 of them are accurately predicted, and for 222 negative reviews also 180 reviews are accurately predicted by our model. Figure 4.5 represents the model testing.

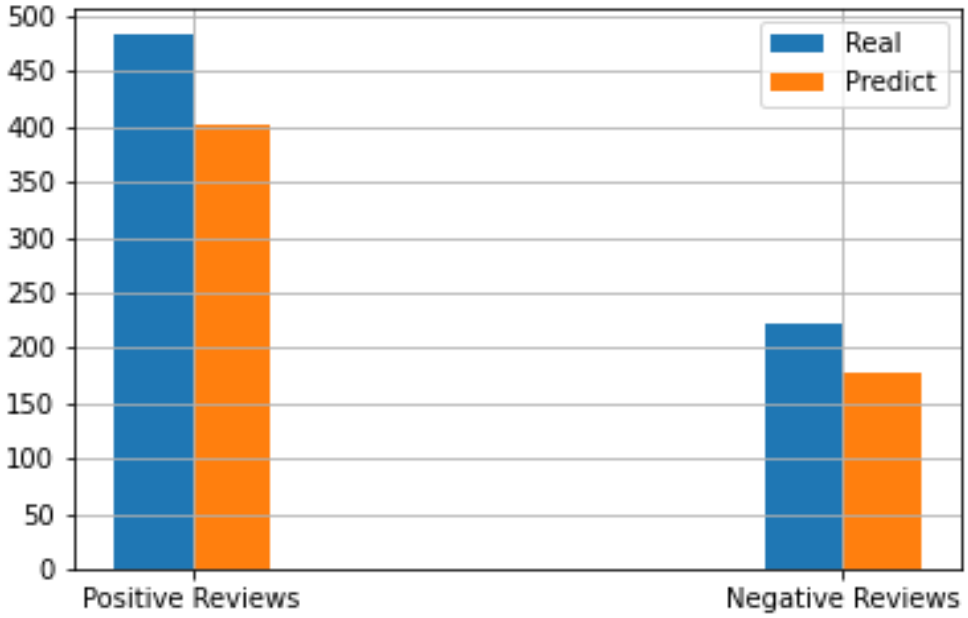


Figure 4.5: Testing model with real dataset.

CHAPTER 5

SUMMARY, CONCLUSION AND FUTURE WORK

5.1 Summary of the Study

Customers of online bookshops are growing in lockstep with the number of people who use the internet. Technology and innovation have brought people from all around the world closer together. E-trade has permitted the rapid growth of an internet bookshop. People nowadays do not want to waste time traveling to a real bookshop; instead, they want comfort and to live as honestly as possible. System learning has received some attention, but there hasn't been much in terms of exhausted Bangladesh. Regardless of the fact that working in vision styles is a frequent term for computer education, Bangladeshi literature is unaware of this. This type of investigation has recently been accomplished as a result of those assignments generating a significant change in our system existence. In any case, the difficulties of Bangladeshi financial topics may be a little concern. In any event, we anticipate that a number of scholars in this field have conducted study in a variety of nations.

5.2 Conclusion

Because of the rapid expansion of Internet users, SA is reliant on a dataset of specialized stuff. This study offers a sentiment classification system based on numerous feature extraction methodologies that can categorize sentiment from Bengali book reviews into positive and negative categories.

We studied 1600 customer reviews from 30 different book categories. We gathered book information and summaries, as well as screening book reviews for significant feature terms. As a result, we uncovered five fundamental characteristics that exist in all client views and influence them: price, transportation, first-rate, design, and pride. After that, we developed a system mastering version.

Our objective is to charge e-book critiques with a 79 percent accuracy rate. The precision of the RF algorithm has been discovered. In terms of accuracy, RF outperformed other typical algorithms like as KNN, decision Tree, SVM, and Random Forest because to its attractive perforation. Both bookshop owners and consumers can investigate which books are worth looking at and which aren't, and capable customers can discern which novels have accurate or bad characters. This strategy is advantageous to bookshop owners. The studies of librarians and e-book purchasers will undoubtedly swap at some point throughout this time period.

5.3 Recommendations

There are a few fantastic options for this:

- To improve data collection accuracy in order to acquire better results from this study.
- The quantity of information in this document is really minimal.
- It is advisable to employ Deep Learning.

5.4 Future Work

The following is the future direction for the development of this work: In Bangladesh, we intend to explore the feeling of a caustic declaration.

- We will develop an advanced structure for implementing this idea.
- We desire to run on an internet-primarily based API to certain analytical sensations in order to get this purpose.
- We can develop an intelligence device based on deep learning techniques in the future.

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), 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 available at <https://www.javatpoint.com/logistic-regression-in-machine-learning.> last accessed on 4-08-2021 at 11AM.

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

BOOK-REVIEW-SENTIMENT-IN-BANGLA-LANGUAGE

ORIGINALITY REPORT

29%	18%	19%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	14%
2	Rely Das, Md Forhad Hossain, Taufiq Ahmed, Ananyna Devanath, Shahnaz Akter, Abdus Sattar. "Classification of Product Review Sentiment by NLP and Machine Learning", 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2022 Publication	7%
3	Mst. Eshita Khatun, Tapasy Rabeya. "A Machine Learning Approach for Sentiment Analysis of Book Reviews in Bangla Language", 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), 2022 Publication	5%
4	Kajal Patil, Sakshee Pawar, Pramita Sandhyan, Jyoti Kundale. "Multiple Disease Prognostication Based On Symptoms Using	1%