Sentiment Analysis on Bangla and Phonetic Bangla Reviews: A Product Rating Procedure using NLP and Machine Learning

BY

Mohammad Mukit
ID: 191-15-12302

Md. Moshiul Huq Rabbi
ID: 191-15-12574

Mohammad Masud Ahmed
ID: 191-15-12364

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

Abu Kaisar Mohammad Masum
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

25 JANUARY 2023

APPROVAL

This research titled "**Sentiment Analysis on Bangla and Phonetic Bangla Reviews: A Product Rating Procedure using NLP and Machine Learning**", submitted by **Mohammad Mukit, ID: 191-15-12302**, **Moshiul Huq Rabbi, ID: 191-15-12574** and **Mohammad Masud Ahmed, ID: 191-15-12364** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 25th **January 2023**.

BOARD OF EXAMINERS

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

Dr. Md. Monzur Morshed **Internal Examiner**
Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dewan Mamun Raza **Internal Examiner**
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

25.1.2023

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

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Ms Subhenur Latif, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Ms Subhenur Latif
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Abu Kaisar Mohammad Masum
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Mukit

Mohammad Mukit
ID: -191-15-12302
Department of CSE
Daffodil International University

Rabbi

Md. Moshiul Huq Rabbi
ID: -191-15-12574
Department of CSE
Daffodil International University

Ahmad

Mohammad Masud Ahmed
ID: -191-15-12364
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Ms Subhenur Latif, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of Machine Learning to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

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

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

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

ABSTRACT

The e-commerce industry has expanded quickly in recent years. Online commerce become more well-liked during the reign of Covid. However, maintaining product quality and customer pleasure are significant obstacles for internet businesses. Even when a product has a five-star rating, customers frequently complain about its poor quality. This is so that the quality of the goods can't be specifically described by the star ranking system. The star rating and the actual product reviews don't always agree. Our research study tries to find a solution to this issue. People are more aware than ever before and they try to express their opinion after buying a product by posting reviews on the e-commerce websites. In this study, we attempted to derive rating from customer feedback. The language that is most commonly used in our nation, Bangla and Phonetic Bangla, was our main focus. In order to do this, we first built our own dataset. we gathered over 4,000 Bangla and Phonetic Bangla product reviews from various e-commerce websites, online store pages, social media platforms, and YouTube videos. The data was then divided into two categories, 1 (Positive), and 0 (Negative), and the dataset was preprocessed, which involved cleaning the data and feature extraction. In order to extract features, we employed TF-IDF. Finally, in order to determine the polarity of the reviews, we trained our model using five different supervised machine learning algorithms namely Logistic Regression, Multinomial Naïve Bayes (MNB), Decision Tree, Random Forest and Support Vector Machine (SVM). SVM had the highest accuracy in the Bangla and Phonetic Bangla datasets when the model was tested using test data, and it attained 82% accuracy in Bangla and 94% in Phonetic Bangla dataset.

TABLE OF CONTENTS

CHAPTERS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

E-commerce business refers to the buying and selling of goods or services over the internet. This type of business has grown significantly in recent years, thanks to the increasing use of the internet and the convenience it offers to consumers.

One important aspect of e-commerce business is the rating system, which allows customers to rate and review products and services they have purchased. This helps other potential customers make informed decisions about what to buy, as they can see what others thought of the product.

The rating system can also be a valuable tool for e-commerce businesses, as it provides feedback on their products and services and can help them improve and meet the needs of their customers. However, it's important for businesses to be transparent and honest in their rating system, as fake or misleading reviews can harm their reputation and trust with customers. It's also important for businesses to actively manage their ratings and reviews, by responding to feedback and addressing any issues raised by customers. Overall, the rating system plays a crucial role in the e-commerce industry, helping businesses to improve and customers to make informed purchasing decisions.

We proposed a review based rating system in which we have used 2114 Bangla and 2000 Phonetic Bangla actual reviews from different e-commerce websites. After preprocessing we have used TF-IDF (Term Frequency-Inverse Document Frequency) to extract feature from the data. Then we followed supervised learning approach to train our model which will be described. When we tested our model, we found the outcome as expected. And we are hoping that the accuracy of our model could be better if we increase the size of the dataset.

1.2 Motivation

The e-commerce industry is expanding quickly and gaining popularity over the past year. Online purchasing is quite popular these days because it is quick and convenient. Additionally, internet stores frequently provide alluring discounts, substantially boosting their popularity. As a result online shopping has become a part and parcel of our daily life. As the popularity of online business has increased, so does the problem. Maintaining product quality and customer satisfaction are two major concern in online shopping. Nowadays people are more conscious than ever before. People try to justify the product before making any purchase. To ensure the product quality, everyone tries to check the rating of the product given by the user. Most of the e-commerce platform has star rating system. The star rating system seems inconvenient to us. In addition, many online shops offer a product review system where customers can share their thoughts on the item. But when it comes to purchase any product, most customers avoid reading the reviews as it takes too much time. This brings us to the primary reason for conducting this research investigation. Our goal is to create a platform that can evaluate any product based on publicly available user reviews. This platform might help to ensure consumers protection by providing the experiences and alert others to potential issues or problems with a product. This can help consumers make informed purchasing decisions and avoid products that may be defective or otherwise problematic. Another key motivation behind this study is it might increase the efficiency of online business. This product rating system can be a valuable marketing tool for companies. Positive ratings and reviews can increase the perceived value of a product and encourage more people to purchase it. Additionally, it might be able to regulate the product's quality and uphold the e-commerce company's reputation.

Therefore, as we can probably assume, this research study will benefit both consumers and the e-commerce industry by improving customer happiness and boosting organizational effectiveness.

1.3 Objectives

- To create a platform which can rate any product based on the reviews posted by the user.
- To create a model which can detect positive and negative reviews from Bangla and Phonetic Bangla text and helps to improve customer experience.
- To enrich the Bangla and Phonetic Bangla dataset for future study.

1.4 Rationale of the Study

As Bangla is a complex language, distinguish between positive and negative reviews from Bangla text is difficult. Another key factor is most people in Bangladesh want to express their opinion in Phonetic Bangla text which makes it more difficult. Because, not everyone uses the same letter to write the same word. On the other hand, Identifying sentiment from Bangla text is tough as it has large vocabulary.

A lot of research has been conducted based on sentiment analysis in English language over the recent year and obtained satisfactory results. Therefore tons of textual data on English language are available on the internet. However, there hasn't been much research done on product reviews for Bangla and Phonetic Bangla texts. So, we decided that it would be worthwhile to carry out research on this subject. We collected two distinct datasets for Bangla and Phonetic Bangla text consisting over four thousand sentences together and categorized into two polarities: positive noted by 1 and negative noted by 0. Then we applied five different classification algorithms to train the model.

1.5 Expected Output

Finding the sentiment of comments is a challenging task as natural language processing involves a lot of data. By continuing our project, we are exploring the concept of sentiment from users' comments by analyzing different supervised learning algorithms. Our expected outcome is to build a model that can accurately identify the overall satisfaction level of customers based on product reviews. We are examining models of different algorithms using a newly created dataset to see how precise the sentiment of the product can be.

1.6 Project Management and Finance

We followed a structured project management approach to ensure the project met our objectives and achieved the expected outcomes. By defining the scope and objectives of our project, we ensured the specific part we were working on. We developed a project timeline to keep track of our project and assign tasks to team members, which helped us stay on schedule and synchronized.

In terms of finance, there is no support required as of now. Implementing the web version of our model can be expensive, as deploying comes with server and storage costs. There will be a need for financial help if we intend to do that.

1.7 Report Layout

This paper is divided into six sections. Following the introduction, which includes the project's goals and motivation, Chapter 2 discusses background analysis, related work, and challenges. Chapter 3 explains the research methodology and the methods used to gather, preprocess, and extract features. These methods are crucial for accurate and reliable results. We also discuss the data sources used in the analysis and the steps taken to ensure high-quality data. Chapter 4 summarizes the analysis, including the experimental setup and results. Chapter 5 covers the social impact and ethical aspects of our work. Chapter 6 concludes with a summary of our research findings and implications for further study, as well as suggestions for future research. This paper provides a comprehensive overview of our research on sentiment analysis of product ratings.

CHAPTER 2

BACKGROUND

2.1 Preliminaries

These days, there is a lot of interest in textual data analysis. This is due to the fact that there are increasing amounts of textual data on the internet. As a result, textual data processing is required to improve user experience and company efficiency. Today, evaluating a product based on customer feedback has become essential. We aim to overcome this problem. Our project involves both natural language processing and machine learning. If we wish to work in the field of sentiment analysis and natural language processing in the future, this research will likely help us advance our technical knowledge in those areas.

2.2 Related Works

Researchers are drawn to examine problems faced by online business owners and customers because of the growing popularity of online e-commerce businesses and the amount of information available on e-commerce websites. Many studies have been conducted to identify sentiment in reviews and comments. Bhowmik, Nitish Ranjan, et. al. [1] proposed supervised machine learning approach along with their own LDD (lexicon data dictionary) and BTSC (Bangla text sentiment score) algorithm to detect sentiment from Bangla text. Chakraborty, Partha et. al. [2] applied seven machine learning algorithms including five classical approaches and two deep learning approaches to detect polarity of Facebook comments in Bangla. They have shown a comparative analysis between classical and deep learning approaches and they found deep learning algorithms perform far better than classical machine learning algorithms. A combined deep learning approach has been introduced by Hossain, Naimul, et. al. [3]. They have implemented a combined CNN-LSTM technique to identify the sentiment in Bangla restaurant reviews. They have created their own dataset with 1000 Bangla reviews and labeled them as positive and negative. Both CNN and LSTM layer is used to extract low level and high level features from data and their model has shown optimistic accuracy.

Khatun, Mst Eshita et. al. [4] used AdaBoost, Decision Tree, SVM, Random Forest, and LightGBM, five distinct machine learning techniques, to identify sentiment in Bangla book reviews. Their dataset, which consists of 5500 user-generated Bangla book reviews, and they achieved a remarkable accuracy of 98.39% in Random Forest. Akter, Mst Tuhin et al. [5] presented the KNN method to assess sentiment in Bangla product reviews on e-commerce websites in order to increase the performance of detecting sentiment. They have used oversampling technique to balance their dataset. Another significant approach is presented by Junaid, Mohd Istiaq Hossain, et al. [7] for Bangla food review sentiment analysis. They have followed supervised learning and deep learning methods. They have used Count vectorizer and TF-IDF along with unigram and bigram features for training supervised algorithms. In addition, they have also used Word2sequence and Glove techniques for training deep learning algorithms. Their experiment has shown the LSTM with word sequence model perform much better than other models. Munna et. al. [8] proposed two different Deep learning NLP models to identify the sentiment and classify review of the product. They used a traditional preprocessing method ,fastText pretrained model to extract the feature and Adam optimization algorithm with a learning rate 0.001. In [9] this study with the supervised approach they applied three different classifier algorithms to find out SVM perform the best with the N-gram techniques followed by HashingVectorizer, CountVectorizer and TF-IDF for vectorization. Jagdale et. el. [10], they proposed a method where they separated the product domain and used a dictionary based approach in lexicon-based to come up with the best accuracy with SVM algorithm. In [12] , a sentiment analysis approach is proposed to rate restaurants based on reviews of food. They analyzed emoticons and emphasized the identification of words to add value to the score. After adding score value manually for each categorized set, they gave the rating out of 10. Kiran et. el. [13] Came up with a recommendation system of product finding the sentiment of product reviews. They used POS tagger to divide the different sentences for form segmentations. Using lexical database WordNet they compared the word to detect sentiment and rate those products accordingly. Senti-Lexicon is a unique way to identify the sentiment of a given text. In [14] this study, Mahmud et. al. who explored the NLTK to tokenize and used the VADER LEXICON dictionary to check polarity. Then they scored the rating for a particular product. Shafin et. al. [15] applied sentiment analysis on product review using traditional methods and compared five different algorithms results. Using porterstemmer they tokenized the raw data and applied the algorithms to find out 88.81% accuracy for SVM.

2.3 Comparative Analysis

There has been a lot of work done on sentiment analysis in both Bangla and English text. But the number of work in Bangla is very little as there is no efficient way to collect the data. We focused on the use of both Bangla and Phonetic Bangla text for sentiment analysis. We created our own dataset and balanced it to ensure the accuracy of our model. We were able to identify the most accurate approach for sentiment analysis in these versions of text by comparing the performance of various models using both Bangla and Phonetic Bangla text. Overall, our study highlights the importance of considering multiple language versions in sentiment analysis and the potential benefits of combining them.

2.4 Scope of the Problem

Online businesses are expanding along with the global use of the internet. Additionally, a huge volume of textual data is produced by online businesses, primarily customer reviews. Both the customer and the business owner can benefit from these reviews. It can be used to determine the precise features of the product that are being addressed as well as the overall sentiment of the reviews: For instance, a review can touch on a product's dependability, usability, or material quality. Understanding the overall evaluation of the product can be improved by identifying these factors. Extracting specific ratings or scores from the review text, however, might assist customers in making crucial choices before making a purchase.

2.5 Challenges

Creating two unique datasets with a massive quantity of data was one of our first problems, as we previously indicated that we have focused on Bangla and Phonetic Bangla language. It is also more challenging to infer sentiment from Bangla reviews due to Bangla's extensive and diverse lexicon. On the other hand, Phonetic Bangla is a popular language in Bangladesh. Although English letters are used when writing, not every word contains the same letter. For example:

1. “amr kase onk valo lagse product ta”
2. “amar kache onek balo lagche product ta”

The following two example has the same meaning and sentiment but their spelling is different from each other. This was another challenge for us to overcome. Besides, ambiguous uses of language and mixed emotions in the reviews has made our job more difficult.

CHAPTER 3

RESEARCH METHODOLOGY

3.1.1 Research Subject

The purpose of this research is to build a platform to identify the positive and negative reviews of product from Bangla and phonetic Bangla reviews on ecommerce websites. One of the key challenges in sentiment analysis is that people often express their emotions and opinions in complex and nuanced ways, and the meaning of a text can depend on the context in which it is used. This can make it difficult to accurately identify and classify the sentiment of text, especially when the text is ambiguous or contains multiple conflicting sentiments. There are many different approaches to sentiment analysis, including rule-based approaches that rely on predefined lists of words and their associated sentiments, and machine learning approaches that use supervised or unsupervised techniques to learn patterns in the data and make predictions about the sentiment of new text.

In our research we have used supervised learning to mine the opinion from the textual data. We mainly focused on the product reviews that a person make after purchasing a product from a online store. So, we have chosen Sentiment Analysis on Bangla and Phonetic Bangla Reviews: A Product Rating Procedure using NLP and Machine Learning as our subject of study for this research.

3.1.2 Instrument

For the purpose of this study, we have gathered 2100 sentences in Bangla and 2000 sentences in phonetic Bangla from various e-commerce websites, direct speech, restaurant websites, and YouTube reviews. When it comes to opinion mining, supervised learning is a well-liked method. Several classification algorithms are frequently

employed to determine the polarity of a sentence. To determine the sentiment of a set of sentences, we used five different classification algorithms in our research.

3.2 Data collection procedure

To determine the polarity of any given sentences of Bangla and phonetic Bangla text, we used the supervised learning approach. A vast amount of data must be needed to train the model in supervised learning in order for it to be effective. Therefore, the first difficulty for us was to create a dataset with such a large number of user reviews. Additionally, the two languages used in our research study are phonetic Bangla and regular Bangla. We had to produce two distinct datasets on two different languages as a result. Although there are numerous datasets in various languages available in various databanks, we chose to develop our own datasets for our research. So we manually constructed the datasets used in this study. We manually gathered the reviews of various products while browsing numerous online e-commerce websites, including darazbd, chaldal.com, and online social networking sites like Facebook and YouTube, in order to create the datasets. In our work, two distinct datasets were employed, and we named them the Bangla dataset and the Phonetic Bangla dataset.

A sample of each dataset is given below for clear illustration. Table 3.2.1 illustrate the Bangla Dataset and table 3.2.2 illustrate the Phonetic Bangla Dataset.

Table 3.2.1 Sample dataset of Bangla text.

comments	remarks
দাম অনুযায়ী বেশ ভাল	1
বিল্ড কোয়ালিটি নিয়ে বলতে গেলে অতটা মজবুত নয়	0
বেশ ভালোই যে উদ্দ্যেশ্যে নেয়া তা সফল হয়েছে।	1
প্রোডাক্ট ভালই সন্তোষ জনক	1
বিল্ড কোয়ালিটি ভালো না	0
জিনিসটি একবারে মন মত হয়েছে।	1
প্যাকটিং টা ভালো হয় নি।	0
প্যাকেট টা ছিড়ে গেছে।	0
নষ্টা একটু হালকা	0

Table 3.2.2 Sample dataset of Phonetic Bangla text.

comments	remarks
besh valoi finishing valo	1
ek kothay oshadaron	1
dekhte joss	1
product valoi shontosjonok	1
build quality temon akta valo na	0
jinish akbare moner moto hoise	1
besh smoothly kaj kore sobai nite paren	1
packeting valo hoyni	0
packet chire gese aro valo hote parto	0

3.3 Statistical Analysis

The number of data that we used for our research will be shown in the tables 3.3.1 and 3.3.2 below. Table 3.3.1 describe the total number of Bangla reviews and the amount of positive and negative Bangla reviews.

Table 3.3.1 Statistics of Bangla Dataset

Total Reviews	2114
Positive Reviews	1110
Negative Reviews	1004

Table 3.3.2 describe the number of total Phonetic Bangla reviews and the amount of positive and negative Phonetic Bangla reviews.

Table 3.3.2 Statistics of Phonetic Bangla Dataset

Total Reviews	2000
Positive Reviews	999
Negative Reviews	1001

Positive and negative reviews were treated equally because both have an impact on the research's anticipated outcome. The primary goal was to construct a well-maintained dataset with texts in both Bangla and Phonetic Bangla.

3.4 Proposed Methodology

Making a proper dataset is necessary before using a classification technique. There are some specific methods and strategies that must be employed to finish a research project, including data preprocessing techniques such as eliminating special characters, punctuation, and stop words, tools for gathering data from various e-commerce websites, and data analysis techniques. Additionally, it includes the feature extraction and algorithms used to train the dataset. For Bangla text specifically, preprocessing must be required because Bangla is a complex and highly inflected language with a rich and varied vocabulary. Preprocessing can help to reduce the complexity of the data and make it more amenable to machine learning techniques.

The entire procedure of classification is shown in Figure 3.4.1 below.

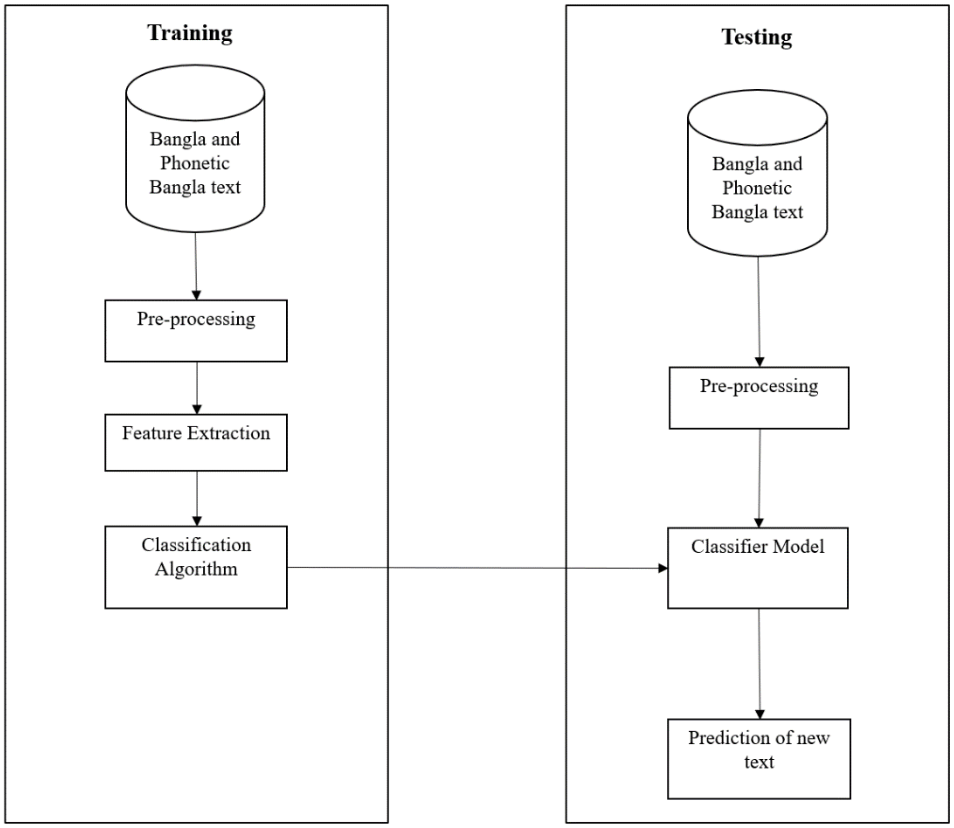
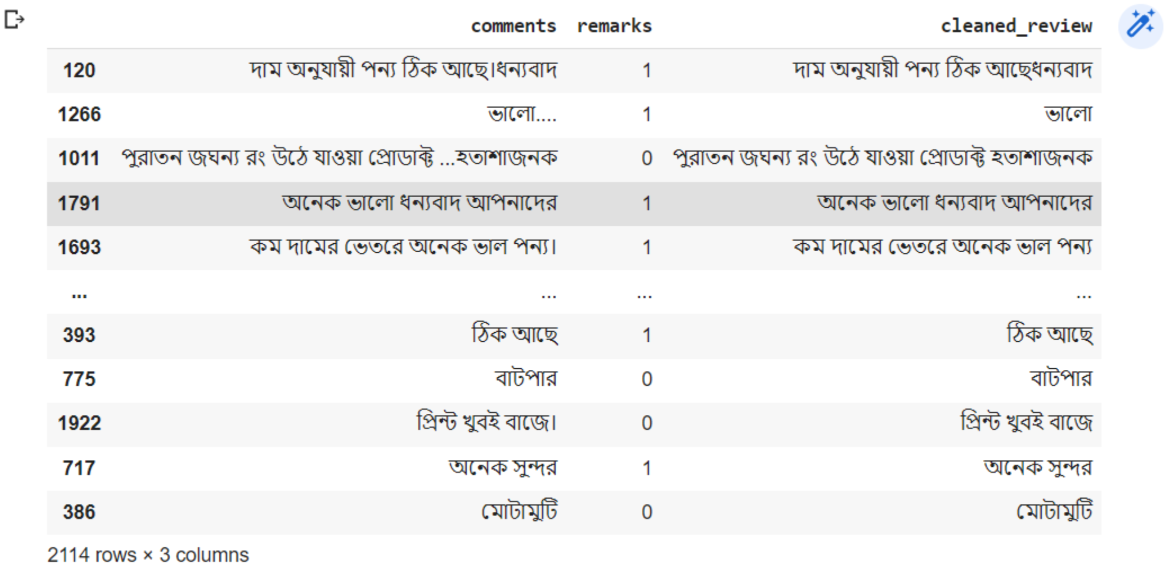


Figure 3.4.1 Text classification block diagram details procedure

3.4.1 Pre-processing

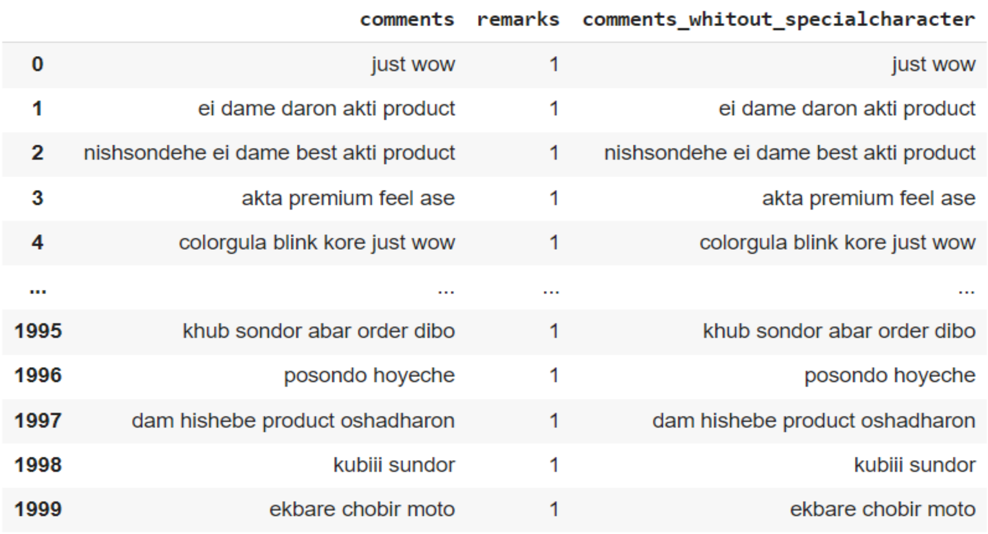
Data pre-processing is an important step in sentiment analysis, as it involves cleaning and preparing the text data for further analysis. At first, we stored the collected data on a google sheet file and labeled every single reviews as 1 (positive) and 0 (negative).This part is done manually. The Google Sheet file is then converted into .csv format after all the data has been labeled. The .csv file could now undergo additional pre-processing. Raw data contains many unnecessary special characters and symbols such as (<, >, \, @, &, ^, $, #, %, ~, +, -, \, . , \, ', ।, \, ",) and punctuation which had an impact on how well the categorization algorithm performed. Therefore, all the special characters, symbols and punctuations have been eliminated from the raw data. Then, we add a new column to our current datasets called "cleaned_reviews" and "comments_whitout_specialcharacter" for Bangla and Phonetic Bangla dataset respectively that contains the text data that has been cleaned. The figure 3.4.2 and 3.4.3 given below demonstrate the cleaned Bangla and Phonetic Bangla dataset respectively.



	comments	remarks	cleaned_review
120	দাম অনুযায়ী পন্য ঠিক আছে।ধন্যবাদ	1	দাম অনুযায়ী পন্য ঠিক আছেধন্যবাদ
1266	ভালো....	1	ভালো
1011	পুরাতন জঘন্য রং উঠে যাওয়া প্রোডাক্ট ...হতাশাজনক	0	পুরাতন জঘন্য রং উঠে যাওয়া প্রোডাক্ট হতাশাজনক
1791	অনেক ভালো ধন্যবাদ আপনাদের	1	অনেক ভালো ধন্যবাদ আপনাদের
1693	কম দামের ভেতরে অনেক ভাল পন্য।	1	কম দামের ভেতরে অনেক ভাল পন্য
...	...	...	...
393	ঠিক আছে	1	ঠিক আছে
775	বাটপার	0	বাটপার
1922	প্রিন্ট খুবই বাজে।	0	প্রিন্ট খুবই বাজে
717	অনেক সুন্দর	1	অনেক সুন্দর
386	মোটামুটি	0	মোটামুটি

2114 rows × 3 columns

Figure 3.4.2 Sample of cleaned Bangla dataset



	comments	remarks	comments_whitout_specialcharacter
0	just wow	1	just wow
1	ei dame daron akti product	1	ei dame daron akti product
2	nishsondehe ei dame best akti product	1	nishsondehe ei dame best akti product
3	akta premium feel ase	1	akta premium feel ase
4	colorgula blink kore just wow	1	colorgula blink kore just wow
...	...	...	...
1995	khub sondor abar order dibo	1	khub sondor abar order dibo
1996	posondo hoyeche	1	posondo hoyeche
1997	dam hishebe product oshadharon	1	dam hishebe product oshadharon
1998	kubiii sundor	1	kubiii sundor
1999	ekbare chobir moto	1	ekbare chobir moto

Figure 3.4.3 Sample of cleaned Phonetic Bangla dataset

3.4.2 Feature Extraction

After the pre-processing, all the data is now cleaned and ready for feature extraction. There are several methods that can be used for feature extraction in sentiment analysis, which involves converting the text data into numerical feature vectors that can be used as input to a machine learning model. Some common methods include Bag-of-words, Word embeddings, TF-IDF (term frequency-inverse document frequency) and N-grams. To extract features from text data, we used TF-IDF. because it gives a way to evaluate a word's significance based on how frequently it appears in different texts.

3.4.3 Training The Model

Training a model in sentiment analysis involves using a machine learning algorithm to learn from a labeled dataset of text documents and their corresponding sentiment labels (e.g. positive, negative, or neutral). The goal of the training process is to learn a model that can accurately predict the sentiment of a given text document. In our research study, we split the whole dataset into 80/20. We used 80% of the data to train the model and the rest of the 20% is used for testing purpose. Algorithm selection is one of the important part of training dataset. As we said earlier, we followed supervised learning to detect the sentiment of a review, there are some common classification algorithms that are used to train the model. We used five different algorithms namely “Logistic Regression”, “Multinomial Naïve Bayes”, “Decision Tree”, “Random Forest” and SVM (Support Vector Machine) to train the model.

3.4.4 Algorithm

We used classification algorithm as the research we have conducted is sentiment analysis and the dataset used in this study is classification dataset. Classification algorithms performs well when it comes to identify the polarity of any textual data. As we mentioned in the [Chapter 3.4.3] that we have used “Logistic Regression”, “Multinomial Naïve Bayes”, “Decision Tree”, “Random Forest” and SVM to train the model. The use of those algorithms is motivated by a number of distinct factors. First of all, since they can foretell a class or category for a given input, classification algorithms are appropriate for sentiment analysis. In the context of our research, the input is typically some text, such as a review and the class or category is the sentiment of that text. Classification algorithms can handle large amounts of data efficiently, making them suitable for our research. Secondly, they can handle high dimensional data and are able to extract relevant features from the data that can be used to make predictions. And lastly classification algorithms can adjust hyperparameters and use appropriate evaluation metrics which makes them suitable for our research. For better understanding, the working principle and the basic architecture of some classification algorithms that we have used in our project are described below.

For example, The basic idea behind logistic regression is to find a mathematical equation that can be used to predict the probability that a given example belongs to a certain class. This equation is called the logistic function. To train a logistic regression model, we need a labeled dataset with input features and a binary output (e.g. 0 or 1). The model is trained to minimize the error between the predicted probability and the true output. This is typically done using an optimization algorithm, such as gradient descent. The basic architecture of Logistic Regression is given in figure 3.4.4.

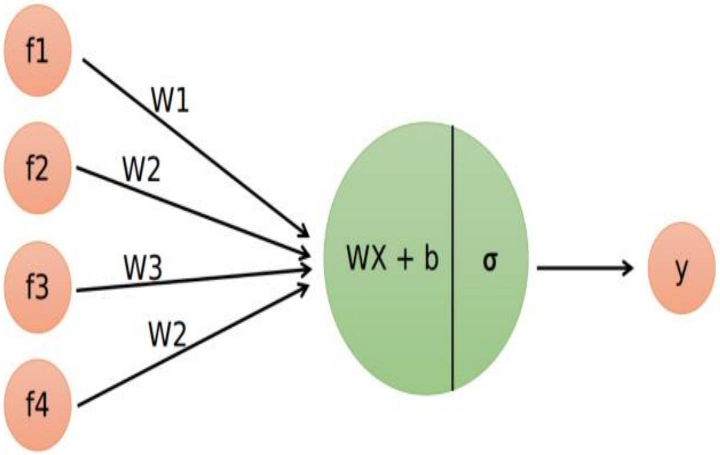


Figure 3.4.4 Logistic Regression Architecture

Now if we discuss about SVM (Support Vector Machine), The algorithm starts by finding the hyperplane that maximally separates the two classes in the training data. This hyperplane is called the decision boundary. The distance from the decision boundary to the closest data points from either class is called the margin. The SVM algorithm tries to maximize the margin, as this will result in a more robust model. In cases where the data is not linearly separable, the SVM algorithm can transform the data into a higher-dimensional space using a kernel function, allowing the decision boundary to be linear in this higher-dimensional space. Once the decision boundary has been determined, the SVM can make predictions for new data by determining which side of the decision boundary the new data falls on. Figure 3.4.5 illustrate the basic architecture of SVM.

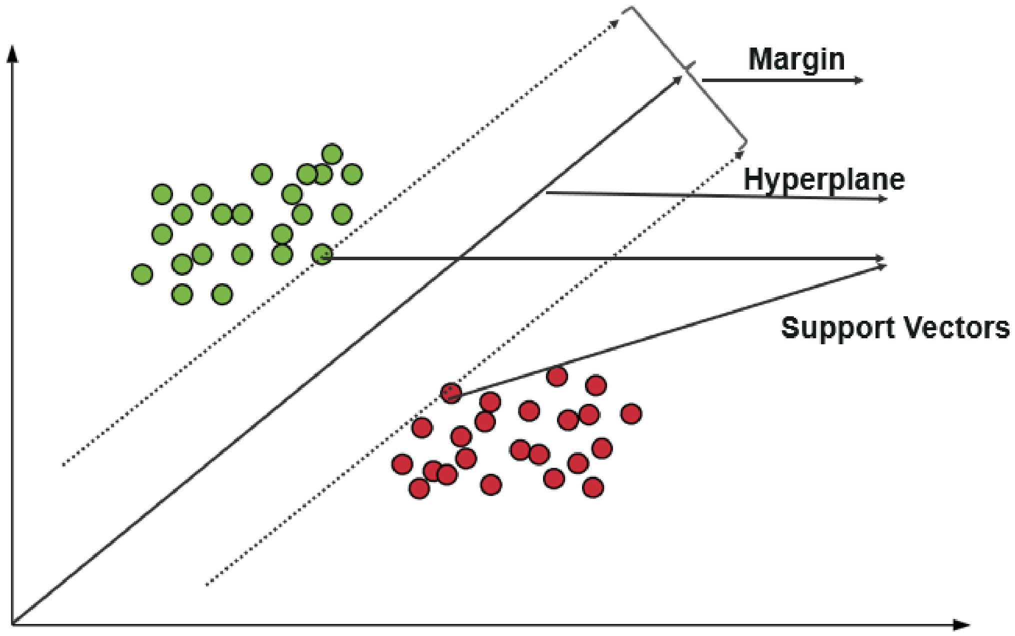


Figure 3.4.5 SVM Architecture

3.5 Implementation Requirements

We have used python programming language where the platform was google colab. The required tools and software are listed below.

1. Python

2. Google Colab

3. Google Sheet and MS Excel

4. Several e-commerce website such as darazbd, chaldal.com, social media like Facebook and YouTube reviews to collect the required data.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

We worked on Bangla and Phonetic Bangla dataset to find a wider range of aspects in sentiment analysis. In total, we have collected 2114 comments for Bangla and 2000 comments for Phonetic Bangla from various ecommerce websites. After collecting the dataset we labeled each sentence manually. We handled the storing part of our dataset using google drive and continued the computational work using google colab.

4.2 Experimental Result and Analysis

After creating the model with 80% of the training dataset, to test our accuracy we gave the 20% test set that we created earlier that is unknown to our model. while we tested each model with test dataset, SVM outperformed other four algorithms in both Bangla and Phonetic Bangla model. And we got the highest accuracy 82% and 94% for Bangla and Phonetic Bangla respectively. Although they both performed well, Logistic Regression and Random Forest were unable to surpass SVM. In the Bangla dataset, Logistic Regression attained 81%, which is second highest and Random Forest obtained 79% accuracy, but in the Phonetic Bangla dataset, they both achieved 92% accuracy. The lowest accuracy we achieved from Decision Tree in both case which is 77% for Bangla and 90% for Phonetic Bangla. Additionally, MNB failed to perform as expected in the Bangla dataset. It had a 77% accuracy rate. There are a few reasons why SVM performed well in our research. Firstly, SVM are particularly effective at handling high-dimensional data, which is common in our dataset. Secondly, SVM are able to find the "maximum margin" decision boundary, which means that they can create a boundary between different classes that is as wide as possible. This can help to improve the generalizability of the model and reduce overfitting. In contrast, Multinomial Naive Bayes (MNB), logistic regression and Decision trees do not explicitly try to maximize the margin between classes. This can lead to poor generalization performance on unseen data. Thirdly, SVM are able to handle non-linear decision boundaries, which means that they can capture complex relationships in the data. This is important in sentiment analysis because the meaning of words can change depending on the context in which they are used. MNB, decision tree and logistic regression, on the other hand, are limited to linear decision boundaries. This means that they may not perform as well on datasets with complex relationships between features. One of the major reason for getting lowest accuracy in Decision Tree is it requires large amount of data than SVM, Logistic Regression and MNB to make prediction accurately. And lastly, SVM are relatively robust to noise and less sensitive to outliers which makes it to perform better. The following 4.1 and 4.2 figures represents the accuracy measure of all five algorithms for Bangla and Phonetic Bangla respectively.

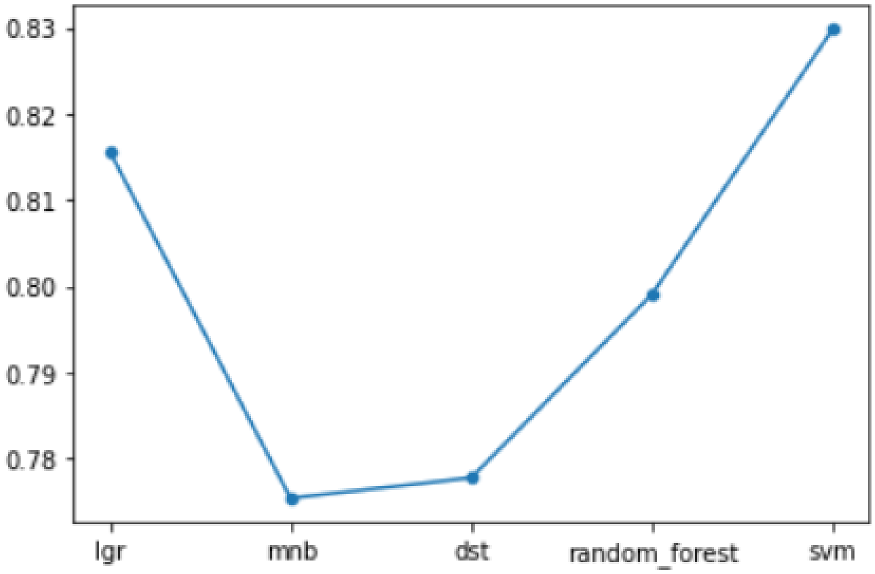


Figure 4.1 Bangla Test Dataset Accuracy Measure

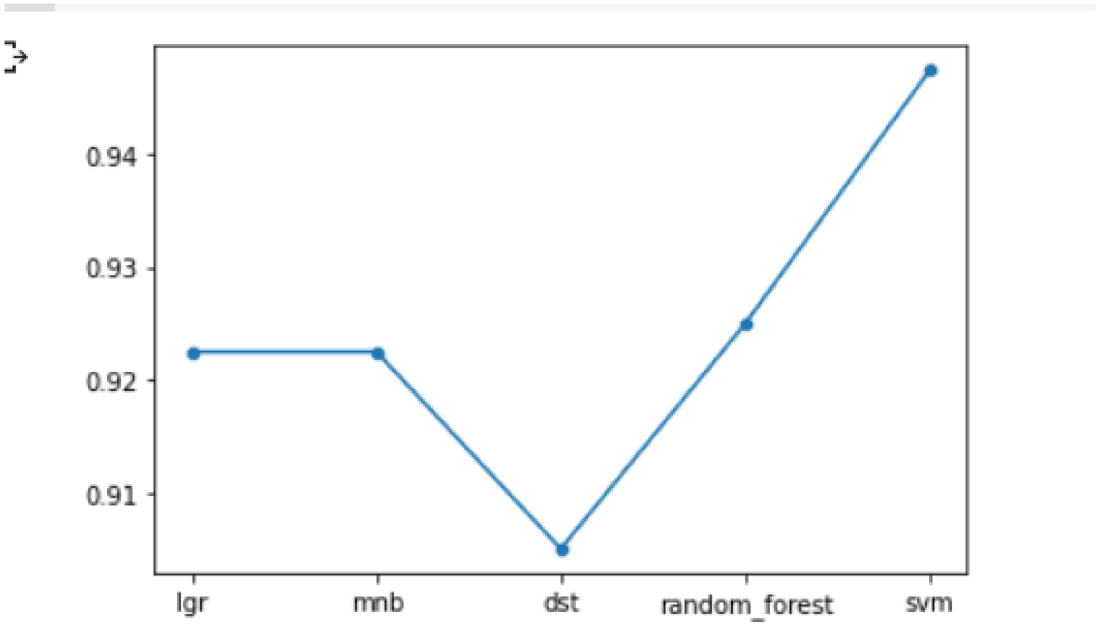


Figure 4.2 Phonetic Bangla Test Dataset Accuracy Measure

But in terms of evaluating classification algorithm, accuracy alone can not provide the standard evaluation. Due to biased datasets, ML model produces skewed result, which is why others measurement need to considered. Precision, Recall and F1 score along with accuracy is needed to asses a ML model perfectly. SVM performed well in the trial. Table 4.1 shows the precision, recall and F1 measure for Bangla and Phonetic Bangla datasets.

Table 4.1 Performance Measure

Dataset	Precision	Recall	F1
Bangla	0.87	0.76	0.81
Phonetic Bangla	0.95	0.94	0.94

Confusion Matrix for both Bangla and Phonetic Bangla dataset given in the following figure for better understanding.

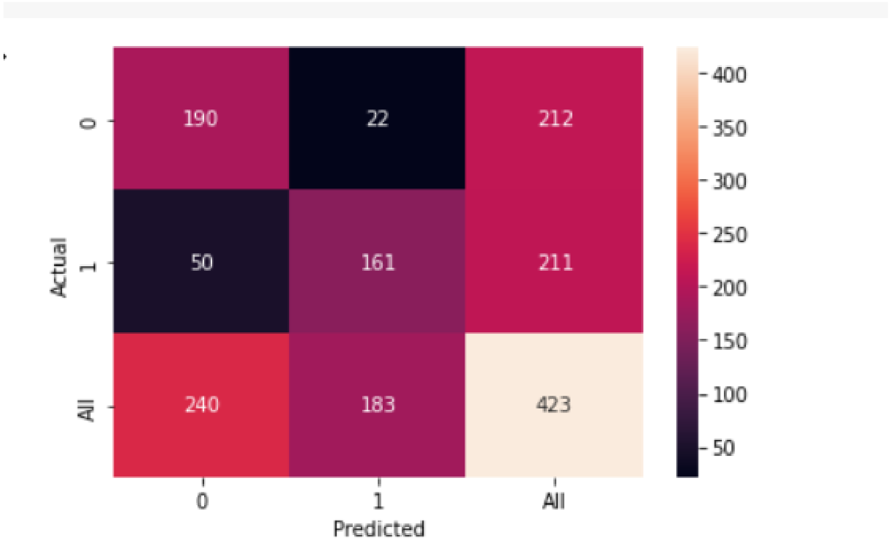


Figure 4.3 Confusion Matrix for Bangla Dataset

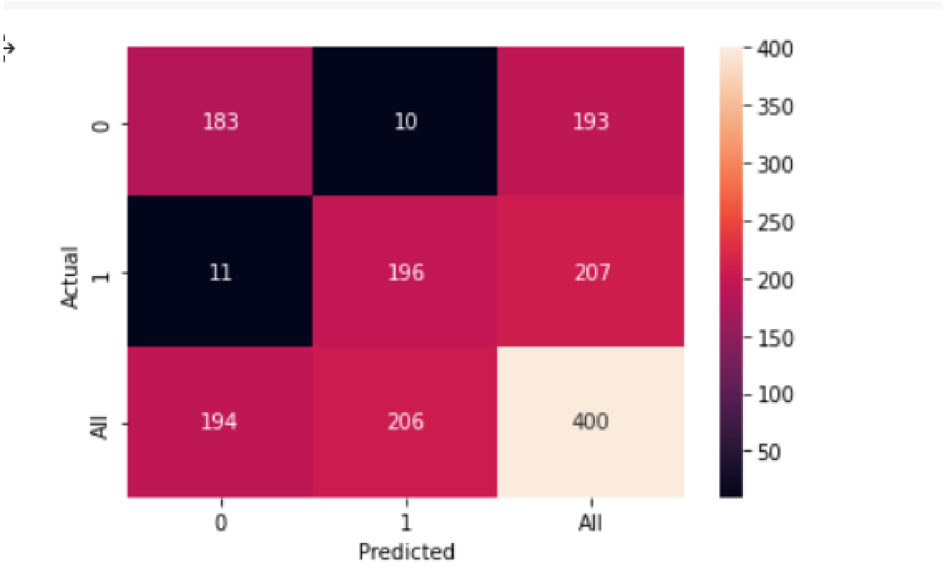


Figure 4.4 Confusion Matrix for Phonetic Bangla Dataset

4.3 Discussion

While implementing our model, we noticed that, even though the ratio of the dataset for Bangla and Phonetic Bangla is the same, the accuracy is not. As we have found out the result of precision and recall of Bangla vary with the size of the dataset, if we improve the size of the dataset, the result can improve significantly. Also the structure of Bangla words is complex and often changes the meaning of sentences with a few changes of stop words or spelling.

In Phonetic Bangla, it occurred to us that the result of accuracy is better because the word itself is in English letter and the algorithm we used is optimized for this language. Also proportion of the positive and negative data was also balanced.

After experimenting with the dataset and model, we found that a given comment can be positive or negative. With text extraction and classification, using the SVM algorithm , we have an accuracy of 94% and 82% for Phonetic Bangla and Bangla respectively.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Society can be significantly impacted by sentiment analysis of product reviews in a number of different ways. It can assist customers in making knowledgeable selections about the things they should buy. Consumers can have a better sense of the amount of general satisfaction among past buyers and determine whether a product is worthwhile to purchase by analyzing the sentiment of product ratings.

It may also affect a company's or a product's reputation. The company's reputation may suffer if the majority of the reviews are negative, which may affect prospective buyers from buying their items. On the other side, if the majority of the reviews are favorable, it may help the business become more well-known and draw in more customers.

5.2 Impact on Environment

Sentiment analysis of product ratings can have a potential impact on the environment in several ways. If a majority of the ratings express negative sentiments about the environmental impact of a product, it could discourage consumers from purchasing the product and potentially lead companies to adopt more environmentally-friendly practices. It can also help companies identify and address environmental concerns in their products and operations, leading to a decrease in their overall environmental impact.

This can help consumers make more environmentally-conscious purchasing decisions by identifying which products have a positive or negative impact on the environment.

5.3 Ethical Aspects

We are aware of the various ethical issues that must be kept in mind while gathering data to create a model. One of these aspects is privacy. All the data collected for our model is publicly available.

Another ethical aspect to consider is bias. We took steps to ensure that our sentiment analysis algorithms are trained on a diverse and representative dataset in order to minimize the risk of bias in the results.

By collecting only publicly available data and handling biased results, we are not violating any privacy standards.

5.4 Sustainability Plan

Our primary goal is to develop a universal platform for rating products based on given sentiment. By creating a web-based endpoint for our model, the product service provider can use it and give a diverse viewpoint of their product to the consumer. To maintain an accurate streamlining of provider-to-customer service, we aim to follow the transparency principle and perform continuous improvement on a regular basis.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

Classifying text to identify the sentiment is a vast field in natural language processing. A good amount of work has been done to find the best classifier for text classification using both supervised and unsupervised approaches. From the acquired result, this research highlights the effectiveness of SVM to identify sentiment in Bangla and Phonetic Bangla compared to other classifier algorithms. After collecting diverse datasets and preprocessing them, SVM comes out to be 93% and 83% effective in classifying positive and negative comments from Phonetic Bangla and Bangla, respectively.

6.2 Conclusions

Detecting sentiment can be an efficient technique to understand the overall thoughts of people about a particular product. In this project, we proposed a supervised machine learning technique for detecting sentiment in Bangla and Phonetic Bangla. Our diverse collection of comments made the dataset unique, which we labeled manually, and the balance of positive and negative comments helped the algorithm identify the sentiment more accurately. We used precision and recall to measure the performance of our algorithm. This research finds the best result from SVM after applying five machine learning algorithms with the TF-IDF vectorizer for both Bangla and Phonetic Bangla.

6.3 Implication for Further Study

- Obtain higher accuracy using combined classifiers.
- Collect more data to enrich the dataset.
- Optimize dataset using stemmer and increase the model’s accuracy.
- Implement the rating system.
- Execute the model on real-time data.

REFERENCES

[1] Bhowmik, Nitish Ranjan, et al. "Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary." *Natural Language Processing Research* 1.3-4 (2021): 34-45.

[2] Chakraborty, Partha, Farah Nawar, and Humayra Afrin Chowdhury. "Sentiment Analysis of Bangla Facebook Data Using Classical and Deep Learning Approaches." *Innovation in Electrical Power Engineering, Communication, and Computing Technology*. Springer, Singapore, 2022. 209-218.

[3] Hossain, Naimul, et al. "Sentiment analysis of restaurant reviews using combined CNN-LSTM." *2020 11th International conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2020.

[4] Khatun, Mst Eshita, and 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)*. IEEE, 2022.

[5] Akter, Mst Tuhin, Manoara Begum, and Rashed Mustafa. "Bangla sentiment analysis of E-commerce product reviews using K-nearest neighbors." *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. IEEE, 2021.

[6] Gowri, S., R. Surendran, and J. Jabez. "Improved Sentimental Analysis to the Movie Reviews using Naive Bayes Classifier." *2022 International Conference on Electronics and Renewable Systems (ICEARS)*. IEEE, 2022.

[7] Junaid, Mohd Istiaq Hossain, et al. "Bangla Food Review Sentimental Analysis using Machine Learning." *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2022.

[8] Munna, Mahmud Hasan, Md Rifatul Islam Rifat, and A. S. M. Badrudduza. "Sentiment analysis and product review classification in e-commerce platform." *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2020.

[9] Haque, Fabliha, Md Motaleb Hossen Manik, and M. M. A. Hashem. "Opinion mining from bangla and phonetic bangla reviews using vectorization methods." *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE, 2019.

[10] Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh. "Sentiment analysis on product reviews using machine learning techniques." *Cognitive informatics and soft computing*. Springer, Singapore, 2019. 639-647.

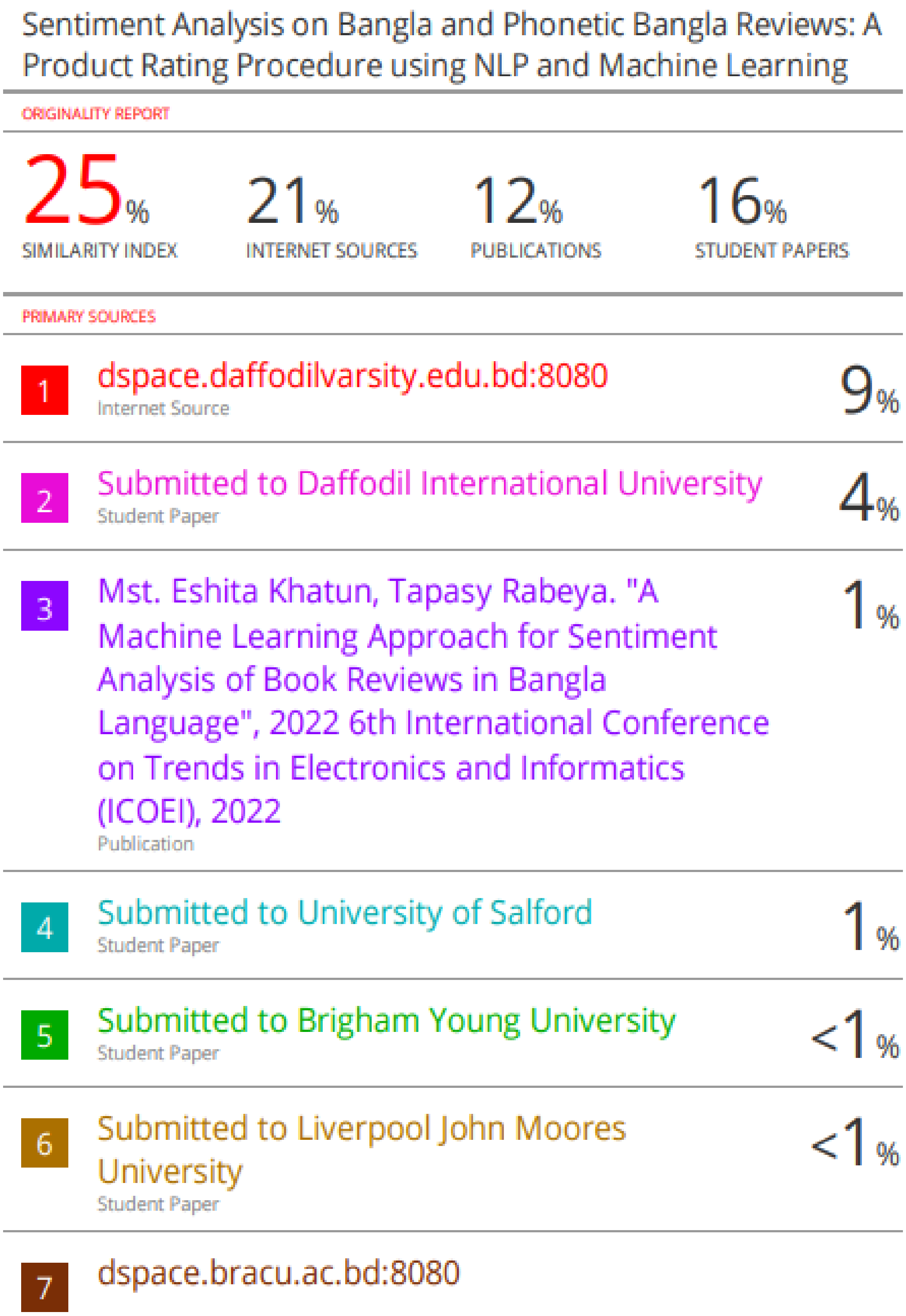
[11] Mai, Long, and Bac Le. "Joint sentence and aspect-level sentiment analysis of product comments." *Annals of Operations research* 300.2 (2021): 493-513

[12] Kaviya, K., et al. "Sentiment analysis for restaurant rating." *2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*. IEEE, 2017.

[13] Kiran, M. Vamsee Krishna, et al. "User specific product recommendation and rating system by performing sentiment analysis on product reviews." *2017 4th international conference on advanced computing and communication systems (ICACCS)*. IEEE, 2017.

[14] Mahmud, Bahar Uddin, et al. "Ecommerce Product Rating System Based on Senti-Lexicon Analysis."

[15] Shafin, Minhajul Abedin, et al. "Product review sentiment analysis by using NLP and machine learning in Bangla language." *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2020.



Sentiment Analysis on Bangla and Phonetic Bangla Reviews: A Product Rating Procedure using NLP and Machine Learning

ORIGINALITY REPORT

25%	21%	12%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	Submitted to Daffodil International University Student Paper	4%
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	1%
4	Submitted to University of Salford Student Paper	1%
5	Submitted to Brigham Young University Student Paper	<1%
6	Submitted to Liverpool John Moores University Student Paper	<1%
7	dspace.bracu.ac.bd:8080	