A STUDY OF BANGLADESHI CUSTOMERS REVIEW SENTIMENT ANALYSIS USING BNLP AND MACHINE LEARNING APPROACHES

BY

MST. SWARNA SARKER

ID: 221-15-5162

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

Supervised By

Mr. Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

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

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

13 JANUARY 2025

APPROVAL

This Thesis titled **"A study of Bangladeshi customers review sentiment analysis using BNLP and machine learning approaches "**, submitted by Mst. Swarna Sarker, Student ID: 221-15-5162 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 13 January, 2025.

BOARD OF EXAMINERS

Dr. Md. Zahid Hasan — **Chairman**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mohammad Monirul Islam — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

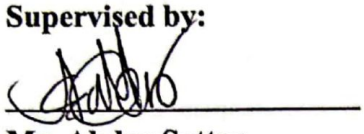
Mr. Afjal Hossan Sarower — **Internal Examiner**
Sr. Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

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

DECLARATION

I hereby declare that this project have been done by me under the supervision of **Mr. Abdus Sattar,** Assistant Professor, **Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

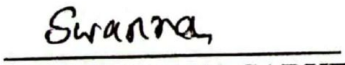


Mr. Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

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

Submitted by:



MST. SWARNA SARKER
ID: 221-15-5162
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, i express my heartiest thanks and gratefulness to almighty God for His divine blessing making me possible to complete the final year project successfully.

I am really grateful and wish my profound indebtedness to **Mr. Abdus Sattar**, **Assistant Professor**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of my 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 drafts, and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Sheak Rashed Haider Noori**, **Professor & Head**, Department of CSE, for his kind help to finish my project and also to other faculty members and the staff of CSE Department of Daffodil International University.

I would like to thank my entire course mates in Daffodil International University, who took part in this discussion while completing the course work.

Finally, i must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

This work concerns a sentiment analysis of the Bangladeshi customer reviews through bangla natural language processing and machine learning approaches. The main goal is to create a model that makes it possible to define whether customers have a positive or negative attitude toward a company. For training and testing the model 6,445 reviews of products or services have been collected and annotated for sentiment. Some of the preprocessing steps include cleaning the text data where noise such as stop words and any special characters were removed from the text data so as to allow for analysis. The analysis used text preprocessing methodologies such as TF-IDF to quantize the textual data into machine learning formats. Decision Tree, Random Forest, SVM, SGD, XGB Classifier used to identify the model with optimal performance in sentiment classification. The project also entailed designing the application programming interface using Streamlit and setting the function where users can enter the custom text and immediately get a sentiment analysis report. First outcomes indicate that the models attained high accuracy of sentiment prediction which proves the importance of using machine learning strategies in analyzing customer's opinions. This paper emphasizes the role of NLP in analyzing consumer behavior and offers a useful instrument – the Customer Sentiment Analyzer – for businesses to evaluate customer's opinions. It also contributes to the literature on processing regional language provides an understanding of the sentiment of Bangladeshi markets.

Keywords: BNLP, Machine learning, Decision Tree, Random Forest, SVM, SGD, XGB, dataset, review detected.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Often seen as a sub-discipline of natural language processing, sentiment analysis has proven to be an essential tool for business and other organizations to understand the written emotions that customers convey. This project of the study specifically highlights the sentiment of Bangladeshi customer feedbacks using the machine learning approach for developing a classifier system that can categorize the outcomes as positive, negative. Given the huge amount of online customer feedback generated daily the companies are beginning to realize the need to mine this "text data" to gain knowledge on customer satisfaction, preferences and expectations. Nevertheless, while there is a plethora of established English-based sentiment analysis systems, there is a lack of systems developed particularly for Bangladeshi language use which contains a combination of Bengali and English along with specific regional and colloquial language controls incorporating English and Bangladeshi terms. This complexity of language is something that and-general-purpose methods of sentiment analysis do not have to combat. Hence, the goal of this work is to develop the sentiment analysis model for Bangladeshi reviews which means to monitor and determine the customer sentiments considering the linguistic and cultural factors of Bangladesh. This project applies other methods including Decision Tree Classifier, Random Forest Classifier, SVM, SGD Classifier, and XGB Classifier to determine diverse methods of performance in classification of the reviews. The activity performed involve data cleaning, data extraction and dimensionality reduction as well as exhaustive testing of the models. In addition, a user interface of Streamlit which is an interface layer is used to create an interface that enables user of Bangladesh who are often not experienced to deal with models using codes to derive insights related to sentiment.

1.2 Motivation

The purpose of this research is to fill the current research gap that has emerged from the rapidly growing digital environment in Bangladesh and mounting number of customer's interactions with the brand through digital media. More specifically, the limitless streams of new information caused by ongoing developments in e-commerce, social media and

online service platforms present businesses with an immense volume of unstructured customer feedback that cannot be effectively assessed without the help of advanced technologies. It may also indicate what the customers were expecting what they wanted and what they expected from the service and where there is room for improvement. Prevailing sentiment analysis tools are mostly constructed in English and do not have the capability of understanding the regional flavor of Bangladeshi language and culture. Local businesses therefore, do not have at their disposal an effective sentiment analysis tool that captures the manner in which the customers from Bangladesh tend to couch their sentiments and feelings. To this end, this project aims to design a new sentiment analysis tool that would enable the analysis if customer reviews in not only Bengali but in English as well thus ensuring businesses get accurate and culturally relevant information. Besides aptly implementing and customizing a new sentiment analysis model for localized language worsens Bangladesh's business decision-making ability but also shows how BNLP and machine learning handle multilingualism.

1.3 Rationale of the Study

The main motivation for this research is to fill gaps observed when using state-of-the-art sentiment analysis techniques on the Bangladeshi customer's feedback. There are few global platforms that provide sentiment analysis models which are mainly trained on the data collected from the English writing world and do not take into account the regional languages and their usage are not complex. Reviews in Bengali language for instance may be written in mix of Bengali and English and contain phrases that are beyond the normal model's comprehension abilities may be due to slangs cultural practices and so on. Further, based on the proposed sentiment analysis for Bangladeshi customer reviews, company operations can be enhanced by connecting customer sentiment metrics to the trials of business making them transparent and comprehensible. This research proposal focuses on improving customer sentiment analysis for Bangladeshi businesses by feeding machine learning models with appropriate data and using accurate and efficient algorithms. Using Decision Trees, Random Forests, SVM, SGD and XGBoost this project tries different techniques to determine which one outperforms all other models in accurately classifying Bangladeshi customer reviews.

1.4 Expected Output

The main output that we create is an ML model that effectively identifies Bangladeshi customer reviews with positive and negative sentiments. What is important in the sentiment classification is the high level of accuracy achieved by this model which takes into consideration the language features of Bangladeshi usage. The outcome of the study will offer a comprehensive comparison of and between Decision Tree Classifier, Random Forest Classifier, SVM, SGD Classifier, XGB Classifier concerning their efficiency, accuracy, efficiency and the F1-score. This comparison also shows some pros and cons of each model before selecting the one that corresponds to the best approach to analyze the given data set. Another significant output of this project includes another crucial output is a user interface designed using Streamlit to make the sentiment analysis model understandable and operable without the requirement for nominal programming knowledge. It will allow the user the input the reviews of the customers and get the sentiment prediction of that review This way, the businesses will be able to gather such knowledge in an easy and technical way. The template of work will also consist of a comparison of other sentiment analysis tools the plan highlighting improvements made with the personal model. This comprises reporting an assessment of performance indicators in a tabular form and a consideration of the appropriateness of the models for mirroring the sentiment of Bangladeshi users. The project will also recognize and present the areas for future work which may include the following open problems, and limitations or what other algorithms one could employ or other local languages that could be considered.

1.5 Report Layout

There are six full chapter in this report each of which covers and develops one aspect of the project systematically. The structure complied with the research background down to the methodology of the investigation, the results and the ramifications of the study. Explained individually each chapter offers detailed information that contributes to the overall understanding of the process of developing the sentiment analysis system its use and the effects it has on the Bangladeshi customer's reviews in this dissertation. The following is an endeavor to explain each chapter in detail.

Chapter 1: The first chapter of the work outlines the background to the entire project facilitating understanding of the importance of sentiment analysis in the evaluation of customer feedback. Sentiment analysis is briefly described at the start of this chapter because the text under investigation the Bangladeshi reviews, poses specific linguistic challenges. This chapter introduces the rationale for the research concerning the incapability of most current sentiment analysis tools to capture regional language variations and where the opportunity for business organizations in Bangladesh lies. Also this chapter defines the research objectives and gives details of the importance of having a specific sentiment analysis model for the Bangladeshi market.

Chapter 2: The background study reviews existing work done on sentiment analysis as well as the application of machine learning and natural language processing in text categorization. This chapter provides a review of the literature with focus on current research done on customer review sentiment analysis with focus on machine learning models such as SVM, Random Forest and Decision Trees. An understanding of the sentiment analysis methodologies is gained through this chapter by comparing the different approaches, data processing techniques and challenges in prior works in the literature. Moreover this section highlights the limitations and shortcomings of prior works specifically in Bangladesh context and language analysis highlighting the relevance of this research.

Chapter 3: This chapter summarize and explain in detail the procedures used in developing the sentiment analysis system. This paper outlines each step the data gathering and cleaning process, the feature engineering process and the selection of the best model. Details about the techniques used in the study including oversampling, feature extraction like TF_IDF is described in addition to the rationale for selecting the ML models that are Decision Tree Classifier, Random Forest Classifier, SVM, SGD Classifier, and XGB Classifier. Furthermore the chapter presents how to train the model and validate the results and methods. Last but not least, this chapter introduces an interface design using Streamlit and thus enables ordinary users to use the model.

Chapter 4: This chapter gives the test results of the different algorithms employed when testing the theory in this study. These after training performance measures include accuracy, precision, recall, and F1-score by each model. For the purpose of comparing the merits and demerits of each algorithm this chapter contains comparative tables and graphical analysis for instance confusion matrices. Following that the discussion is carried out to elucidate why some models turned out to be superior to others with regards to prediction of Bangladeshi customer reviews. In this section the results of the assessment of the model and possible shortcomings in the classification of sentiment are described.

Chapter 5: This chapter reviews the real-world social impact of the developed sentiment analysis system on Bangladesh society and environment and sustainability aspects. It talks about the promotion of sentiment analysis to local firms since it helps increase customer satisfaction and director efficient business strategies. Also this chapter provides an understanding of the environmental cost of training machine learning models to that argument for efficient computing strategies. The chapter also discussed about the sustainability of project as the model would require some periodic updating and hence require constant maintenance.

Chapter 6: In the last chapter of the study conclusions are drawn relative to the project findings made and the limitations observed anywhere in the research process. It synthesizes ideas from each chapter describing how the project effectively managed the specific difficulties of reviewing Bangladeshi customers. This chapter also views open issues that could not be elucidated in the existing study and proposes recommendations for future research including incorporating enhanced NLP models. This chapter is concluded by an acknowledgment of the impact of the project and research direction for its further evolution.

CHAPTER 2

Background Study

2.1 Terminologies

To avoid confusion some basic terms involving BNLP sentiment analysis and ML are described in this section as a background to the subsequent sections of the paper. The said terminologies provide the premise for the comprehension of the technical approaches employed in the project. Opinion mining is defined as the process of using computational methods to analyze opinions, sentiments and emotions carried in a text. It is applied in BNLP mostly to establish whether the language used in a certain text fragment, e.g. review or a comment, has a positive, negative sentiment. Usually it categorizes an input into positive, negative. NLP is therefore defined as a sub field of Artificial intelligence that deals with human languages and interactions between people and computers. It enables text comprehension and synthesis comprehension and translation of human language text and language analysis specificities like sentiment analysis among other purposes. Machine learning refers to the process of using algorithms that are designed to learn from data then try and predict or make some decisions. For the purpose of sentiment analysis the ML algorithms such as the Support Vector Machines, Decision Trees used to classify the text data with regard to the sentiment type. This is the act of segmenting text into smaller parts otherwise known as tokens which then are analyzed individually. The tokenization phase is a very important phase of NLP preprocessing. TF-IDF on the other hand is a method which is used to measure the importance of a given word in that particular document within the context of a given set of documents. It is often employed for feature extraction in sentiment analysis problems. Such terminologies and many others assist in structuring the report's technical discourses to offer readers a conceptual anchor to apprehend the adopted methods and strategies.

2.2 Related Work

This section discusses previous work contributions in the field of Sentimental analysis NLP and ML especially in other languages than English and overemphasizes methods that cope with similar issues. Literature review of using machine learning algorithms for sentiment

analysis and further studies done on regional language with references to the new improvements in natural language processing are reviewed.

Rahman et al. (2020), used TF-IDF vectorization to apply Support Vector Machine (SVM) to categorize sentiment in Bangla movie reviews. Obtained an accuracy of 80% which basically reflected the importance of using classical machine learning in simple sentiment analysis [1] Hossain et al. (2021), For analyzing the customer reviews, the LSTM has been used with an accuracy of 85%. This study emphasized that deep learning models' ability to learn the sequence of word dependencies and the context [2] Alam et al. (2019), Implemented Naive Bayes text classifier with results taken through tokenization and stemming to 76% accurate. This research focused on how the Naive Bayes classifier is one of the easiest and fastest methods to use in sentiment analysis [3].

Khan et al (2020), Created a Convolutional Neural Network (CNN) for Bangla text classification and successfully achieved 78% of accuracy. This approach was also helpful in extracting local features of text [4] Ahmed et al. (2022), Examined a usage of Word2Vec embeddings in parallel with the usage of SVM what led to the score of 82%. In responding to this essential research question, this study aims at enhancing feature representation [5] Chowdhury et al. (2018),Centered on Random Forest and other ensemble methods related to sentiment analysis with moderately high accuracy. The manuscript laid a great deal of stress on the ensemble models as the way to improve performance [6] Akter and Nasrin (2020), Leaning performed experiments using Logistic Regression & Decision Trees on Bangla product reviews. Discovered that although the logistic regression provided higher interpretability than RT, seven percent of accuracy were lost in comparison with the more complex algorithms [7].

Sarker et al. (2021),Pre-trained Integrated Bidirectional LSTM (WBILSTM) addresses the long tail of tweets with 85% accuracy in sentiment analysis. This work demonstrated the indispensability of bi-directional context capture into sentiment models [8] Mollah et al. (2019), Predicted sentiment on customer feedback data using XGBoost for generalized boosting and shown how boosting techniques can improve accuracy if fine-tuned [9] Hasan et al. (2017), Carried out a comparative analysis with K-Nearest Neighbors (KNN) and SVM claiming that the general performance of the latter model is superior to KNN in terms of precision and time complexity [10] Rahim et al. (2020), Used pre-training embeddings

particularly Glove in deep learning for sentiment analysis in Bangla language. Discovered that all these embeddings enhanced the classification measures than the regular word vectorization method [11] Fahim and Ahsan (2021), Implemented the ensemble techniques together with deep learning categories for better outcome to find the 84% accuracy. Stressed that the contiguous multilevel hybrid systems still have a great potential in development [12] Islam et al. (2022), Implemented transfer learning using BERT for Bangla, which was best in the all the existing techniques for sentiment classification. Emphasized the importance of pre-trained language models in handling of specialized language problem [13].

Akash et al. (2018), Performed stomping using LSTM with specially designed pipelines for Bangla since the performance varies with different languages pre [14]. Sultana et al. (2019), Made a comparison between various algorithms among them including Decision Trees, Random Forest, and Neural Networks. Observed that the accuracy of the neural network was better, although the network needed more computational time [15] Kamal et al. (2021), Investigated the effectiveness of Recurrent Neural Networks (RNNs) for the sentiment analysis and concluded that, adding attention leads to an overall better sentiment prediction for long sentences [16] Rashid and Parvin (2020), Investigated the ensemble of Naive Bayes, given in the first part of the paper, with SVM to achieve enhanced results. Therefore, the interaction of these features produced lower total error and greater overall precision [17] Nahar et al. (2018), Supplemented textual sentiment lexicons to machine learning models and corresponding improvements in accuracy when lexicon-based sentiment scores are incorporated [18] Mahmud et al. (2020), Specifically aimed at developing a new annotated Bangla dataset exclusively for sentiment analysis thereby solving the data deficiency problem that has slowed down Bangla NLP research in the past [19]

Mitu and Alam (2022), A combination LSTM-CNN model was used to incorporate LSTM's sequential nature and the feature extraction effectiveness of CNN's [20] These works give general picture about how different models, preprocessing techniques and combined techniques are utilized for sentiment analysis particularly focused on Bangla text and customer reviews [21]

2.3 Comparative Analysis and Summary

The Comparative Analysis and Summary section describes and compares different approaches shown in the previous studies on sentiment analysis especially for the multiple languages and the specific features of the languages in particular such as Bangla. This comparison is important in evaluating the performance of various algorithms, pre-processing methods, variations of datasets, and language–dependent issues. Cross-lingual sentiment analysis studies reported that basic machine learning classification methods such as SVM, Naïve Bayes and Decision Trees are effective in achieving high levels of accuracy when applied to well preprocessed datasets. For example, previous work that has been done on sentiment analysis in Spanish and French has shown great deals of achievement when used along with models such as stemming, stop-word removal and tokenization.

These methods are essential for controlling the morphological richness and syntactic fluctuations typical for non-English languages. For example, a study using SVM with French customer reviews achieved a high level of accuracy because in addition to developing stop-word lists they used TF-IDF to extract features which do not change the language's semantic details.

Recently the Random Forest model has been extensively used in studies due to its stability and ability to address issues of big data by building a multitude of decision trees. A study conducted on languages like Hindi and Tamil to implement Random Forest proves that it is more tolerant of the linguistic variation, compared with single classifier models due to non-overfitting. For example, one work on Hindi sentiment analysis pointed out that Random Forest is slightly both more accurate and more stable than Decision Trees due to the characteristics of its ensemble decision techniques which integrates many Decision Tree's predictions. Regarding Bangla language particularly previous researchers have encountered some problem because of the language features namely script, syntax, phonology etc.

Bangla is used in informal manner in many cases with use of mixed texts Bangla as well as English which makes it even a little more challenging for performing NLP tasks. Due to this preprocessing has been modified for Bangla as well and new stop word list along with suitable method of tokenizing the mixed language documents have been incorporated. SVM and Random Forest were observed familiar algorithms in Bangla sentiment analysis

based on the previous work although the performances slightly fluctuated. For example, satisfactory accuracy was obtained while applying SVM to Bangla e-commerce reviews where tokenization was done considering the script of Bangla. But because of these shortcomings of these traditional models the accuracy levels have often been limited and particularly while considering the relatively informal and colloquial nature of text in Bangla which this proposal uses. SVM is very fast with the structured data and has been used effectively for sentiment analysis where very fine level of sentiment detection is required. However, due to the computationally intensive nature of machine learning models, its application is relatively rare in regional language analysis particularly where resources constraints are profound.

However these advance methods have been proved to offer possibility for high degree of accuracy and flexibility in allowing informal as well as computational practice in terms of availability of computational power. Comparing the results of the models, we can see that Naïve Bayes and XGBoost are the standard models with adequate accuracy, while SVM and SGD ensembles increase the model's stability and flexibility. However, the incongruities still remain in Bangla and other regional languages where raw text preprocessing needs individual attention and in some cases, may require powerful models capable of mastering complex language patterns. As such the aim of this study utilizing some of these insights to increase the Bangladeshi customer sentiment classification accuracy by employing methods such as traditional ones alongside ensemble methods with some extra steps of preprocessing techniques employed during the classification. To fill the existing gaps in previous research and the specific language features that are challenging within Bangla, this study aims to implement the findings and achievements of prior research of sentiment analysis for Bangladeshi users.

2.4 Scope of the Problem

This paper contributes the following open issues in sentiment analysis domain even after substantial research works and newly incorporating natural language processing with machine learning algorithms or techniques.

It is more common to find sentiment analysis datasets where one or more of the partitioned sentiments are dominant over the other. This skews model training and estimating as the classifier might rival the majority class while ignoring the minority ones. This issue is a common problem where oversampling techniques such as Random Oversampling can help reduce the problem but it creates redundancy or over fitting of the data in the model. The existing models of sentiment analysis often fail to compute the cultural and linguistic differences. Feeling and meanings of words or phrases might be entirely different from one language or culture to another. For example, the use of the figures of speech such as irony sarcasm or idioms is very difficult to unpack and especially from instances where Bangladeshi datasets have been collected from multi-lingual speakers where communication involves use of regional languages and code-switching between English and Bengali. Sentence level analyses for traditional NLP models are particularly challenging where complex and nested sentence structures are involved. Ambiguities, use of two negatives together contradictions or paradoxes are some of the challenges which can cause the predictor to get it wrong.

2.5 Challenges

Developing an effective sentiment analysis model for Bangladeshi customer reviews presents several challenges:

Linguistic Complexity: Bangla has a completely different syntax, script and vocabulary which makes a number of NLP processing tasks a lot harder. These challenges when translated to effective model performance involve a lot of preprocessing steps such as creating the tokenization and stemming specific to Bangla.

Data Scarcity: However, there are few Bangla sentiment datasets available unlike large sentiment analysis dataset in English. This scarcity demands more numerous

and pointwise efforts in data capturing and labeling, as well as more frequent oversampling or synthetic data creation.

Handling Informal Language: It may contain slang customer dialects and the text content often encompass Bangla mixed up with English language. This makes the tokenization and text processing difficult especially to machine learning algorithms that are hard put to decode informal language.

Computational Resources: Training a range of machine learning models to select the most efficient one takes computing resources, especially if using techniques like Random Forrest or deep learning.

Evaluation Metrics: To be able to identify the efficiency of the sentiment model more than simple accuracy is needed. Accuracy is useful but precision, recall and F1-scores help to understand how the model deals with false positives and negatives which might influence business decisions based on customer sentiment.

To overcome these challenges, complex preprocessing, selection of appropriate machine learning techniques analysis is used.

CHAPTER 3
Research Methodology

3.1 Introduction

The objective of this project is to determine the sentiment of the Bangladeshi customer reviews through text analysis using Bangla natural language processing and machine learning. Opinion mining or sentiment analysis is a subfield of text mining that focuses on identification and categorization of text according to the flavor of the opinion expressed in this text. In the context of the Bangladeshi setting, the analysis of customer sentiment has immense worth, as it helps captive the understanding of the consumer his or her level of contentment and opportunities for changes. Nevertheless, it has to be noted that Sentimental analysis in requires an understanding of Bangla language sentiment which is not only lexically but also syntactically and structurally different from English and also lacks big pool of labelled data. Current state-of-the-art libraries aren't strongly in English balanced and therefore, it is challenging to employ approaches from the English language in Bangla text. To overcome these challenges this study involves a host of machine learning algorithms including Decision Trees, Random Forest, Support Vector Machine (SVM), and XGBoost etc. All the models used undergo assessments and tests and the model with the highest result in accuracy and reliability of the sentiment classification is selected. This study includes an interactive interface built using Streamlit, allowing users to analyze reviews in real-time. The platform aims to provide a comprehensive solution for Bangladeshi businesses to leverage customer feedback effectively. By focusing on Bangla sentiment analysis this project not only contributes to the field of NLP for underrepresented languages but also enables businesses to make data-driven decisions tailored to the Bangladeshi market. This introduction to sentiment analysis in Bangla paves the way for more localized applications in e-commerce, social media and customer support sectors.

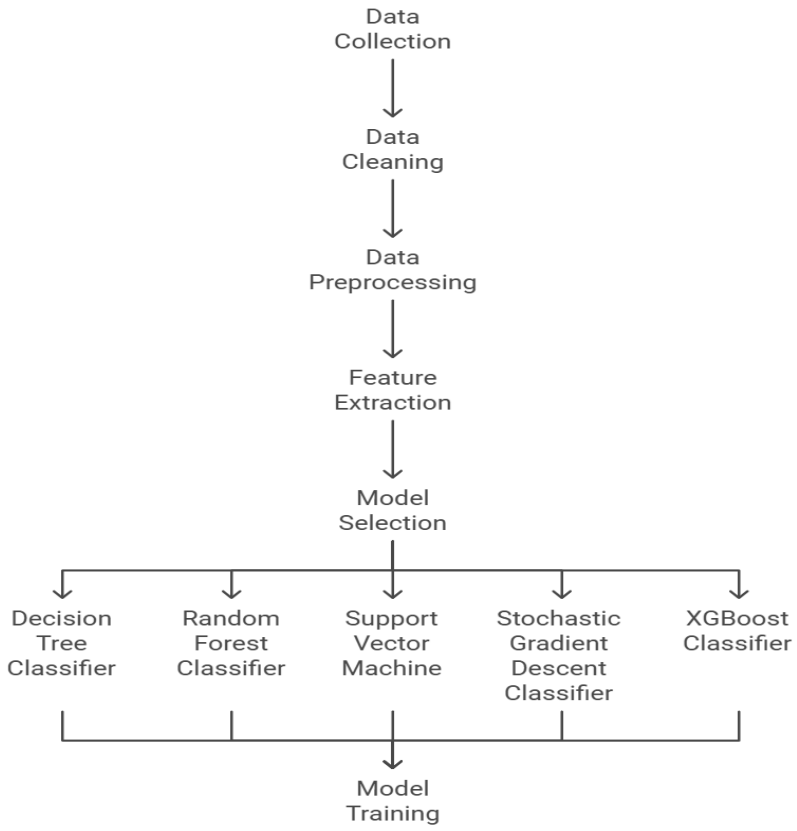


Figure 3.1: Proposed System Architecture

3.2 Data Collection Procedure

The first task in this project was to collect a proper dataset of Bangladeshi customer feedbacks. As this dataset is expected to reflect real word customer sentiment in Bangla reviews were gathered from different online sources and self-collected. My sources included e-commerce platform social media platforms and product review forums where Bangladeshi user give their insight on products and services. The collected dataset was stored in CSV format wherein every entry contains crucial metadata including the content of the review, its sentiment labels which can be either positive negative. While data gathering it was crucial to concentrate on material that reflects as much speakers and kind of Bangladeshi dialects as possible. This variation of words improves the generalization of the model to actual Bangla texts and makes the model versatile.

	comment	sentiment
0	খুব বাজে দামটাও অনেক বেশি	negative
1	চমৎকার একটি প্রোডাক্ট	positive
2	দামটা অনেক বেশি	negative
3	যে কালার অর্ডার করছি আসছে অন্য কালার	negative
4	কি আর বলতে পারি এই সার্ভিসের ব্যাপারে আর সময় য...	negative

Figure 3.2: Sample Dataset

3.3 Dataset Cleaning

Cleanliness is a crucial step in readiness for BNLP on the data that I am going to use hence the need to clean my data set. To start with customer reviews comprise a lot of noise in the form of peculiar characters, signs and mistyped words most of the time. This process included:

3.3.1 Removing special characters: To eliminate noise, symbols, emotions, and HTML tags were stripped from text.

3.3.2 Handling null values: In cases where there were missing or null data; these values were either deleted or filled with appropriate data.

3.3.3 Addressing duplicate entries: To avoid imbalance in data, there were duplicate values cleared from the data set.

Having cleaned data gives a standard and suitable foundation for the sentiment analysis and also the extraction of features.

TABLE 3.3: Positive Negative value of dataset

Type of sentiment	Count
Positive	3241
Negative	3147

3.4 Dataset Preprocessing

This data set is then cleaned in order to improve on the quality and to make them ready for analysis. The preprocessing pipeline includes:

3.4.1 Remove Punctuation:

	comment	sentiment	msg_clean
0	খুব বাজে দামটাও অনেক বেশি	negative	খুব বাজে দামটাও অনেক বেশি
1	চমৎকার একটি প্রোডাক্ট	positive	চমৎকার একটি প্রোডাক্ট
2	দামটা অনেক বেশি	negative	দামটা অনেক বেশি
3	যে কালার অর্ডার করছি আসছে অন্য কালার	negative	যে কালার অর্ডার করছি আসছে অন্য কালার
4	কি আর বলতে পারি এই সার্ভিসের ব্যাপারে আর সময় য...	negative	কি আর বলতে পারি এই সার্ভিসের ব্যাপারে আর সময় য...

Figure 3.4.1: Text Preprocessing to Remove Punctuation from Comments in a Data Frame

The output what such type of data brings includes the ‘comment’ column which exists in the first comment and the new ‘msg_clean’ column with the punctuations erased. For context the ‘sentiment’ column is also available here.

3.4.2 Tokenization:

	comment	sentiment	msg_clean	tokenized
0	খুব বাজে দামটাও অনেক বেশি	negative	খুব বাজে দামটাও অনেক বেশি	[খুব, বাজে, দামটাও, অনেক, বেশি]
1	চমৎকার একটি প্রোডাক্ট	positive	চমৎকার একটি প্রোডাক্ট	[চমৎকার, একটি, প্রোডাক্ট]
2	দামটা অনেক বেশি	negative	দামটা অনেক বেশি	[দামটা, অনেক, বেশি]
3	যে কালার অর্ডার করছি আসছে অন্য কালার	negative	যে কালার অর্ডার করছি আসছে অন্য কালার	[যে, কালার, অর্ডার, করছি, আসছে, অন্য, কালার]
4	কি আর বলতে পারি এই সার্ভিসের ব্যাপারে আর সময় য...	negative	কি আর বলতে পারি এই সার্ভিসের ব্যাপারে আর সময় য...	[কি, আর, বলতে, পারি, এই, সার্ভিসের, ব্যাপারে, ...

Figure 3.4.2: Tokenization of Bengali Text Using Basic Tokenizer from the BNLP Library

The Data Frame as indicated below shows the original comment column sentiment msg_clean (this has had all the punctuations removed from the previous step and tokenized as shown in the newly created tokenized column) whereby tokenized contain lists of tokens that are the individual words in msg_clean.

3.4.3 Stop-Word Removal:

	comment	sentiment	msg_clean	tokenized	Removed Stopped word
0	খুব বাজে দামটাও অনেক বেশি	negative	খুব বাজে দামটাও অনেক বেশি	[খুব, বাজে, দামটাও, অনেক, বেশি]	[বাজে, দামটাও]
1	চমৎকার একটি প্রোডাক্ট	positive	চমৎকার একটি প্রোডাক্ট	[চমৎকার, একটি, প্রোডাক্ট]	[চমৎকার, প্রোডাক্ট]
2	দামটা অনেক বেশি	negative	দামটা অনেক বেশি	[দামটা, অনেক, বেশি]	[দামটা]
3	যে কালার অর্ডার করছি আসছে অন্য কালার	negative	যে কালার অর্ডার করছি আসছে অন্য কালার	[যে, কালার, অর্ডার, করছি, আসছে, অন্য, কালার]	[কালার, অর্ডার, করছি, আসছে, কালার]
4	কি আর বলতে পারি এই সার্ভিসের ব্যাপারে আর সময় য...	negative	কি আর বলতে পারি এই সার্ভিসের ব্যাপারে আর সময় য...	[কি, আর, বলতে, পারি, এই, সার্ভিসের, ব্যাপারে, ...	[সার্ভিসের, সময়, তারিখে, দাওয়ার, কথা, দিয়েছে, ...

Figure 3.4.3: Stop word Removal from Bengali Text Data

The python code employs the bnlp.corpus and bnlp.corpus.util libraries to have a column known as msg_clean in a Dataframe referred to as df deboned of any word familiarly referred to as a stop word.The data frame shown below is the data frames first five data points with Bengali comment, msg_clean, tokenized and Removed Stopped words column with sentiment labels.

3.4.4 Oversampling

This method of oversampling is used commonly in machine learning when in actual dataset a class significantly differs from other classes. This imbalance gives the model a tendency to predict the majority class and in most cases the model will have low accuracy on the minority classes. RandomOverSampler on the other hand takes instances from the minority class and mechanically reproduce them in order to bring about a balance of sample sizes of the classes. This helps the model get more instances of each class during training and thus improving the recall and F1-score of the minority class.

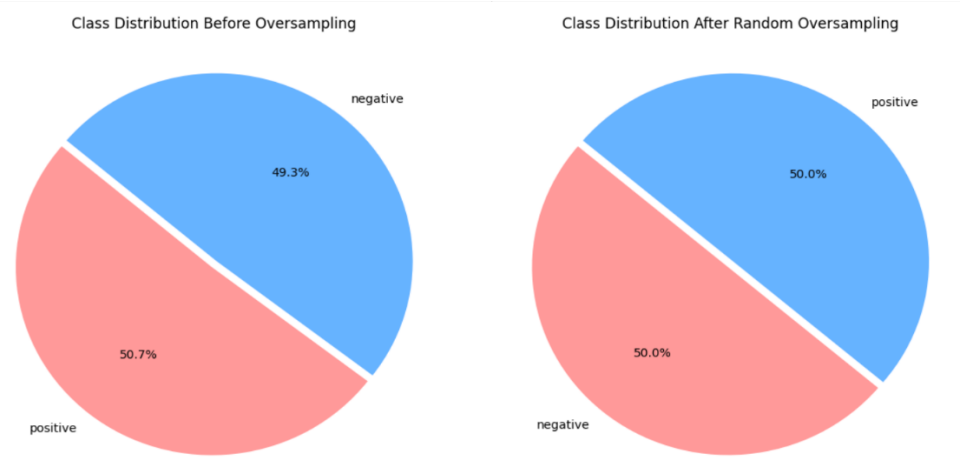


Figure 3.4.4: Class Distribution Before and After Oversampling

The pie charts depict the class distribution before and after class oversampling, and given that after oversampling the distributions are nearly equal they make sense for a stronger and more impartial model performance.

TABLE 3.4: Data Splitting

Split Percentage	Dataset Splitting	Number of Comment
80%	X_Train_tfidf	4537
	Y_Train	4537
20%	X_Test_tfidf	1945
	Y_Test	1945

3.5 Proposed Methodology

The approach used in this project proposal is to develop a model that would effectively classify Bangladeshi customer reviews in terms of sentiments as either positive, negative sentiments. This is a brief on several fundamental steps of this methodology including data preprocessing feature extraction, model selection and training model evaluation. The output of the model is presented to a user interface to make sentiment prediction for Bangla text in real time. The following are the detailed course of action.

3.5.1 Data Preprocessing

Balancing is performed as a data preprocessing step to get the text data ready for developing the machine learning models. It is crucial to decode Bangla language processing as it is not associated with many other standard languages, Bangla language script there are few Bangla NLP resources available and changes are required in preprocessing. Key steps include:

Tokenization: Segmenting each Bangla review into words or tokens so that each review can be further analyzed easily. This step helps to segment the whole sentences into smaller units so as to enable the model to derive meaning in the relation between the various words.

Stopword Removal: Excluding Bangla stop words which are more general and do not comprise sentiment information e.g. conjunctions and pronouns. This makes for less noise and enhances the model's efficiency whenever it is implemented. Stemming and Lemmatization: Removing inflections makes for a smaller

vocabulary in which the model can learn the meaning of these words without having to learn all the different forms of them. Generally for Bangla text stemming may require custom algorithms because there was scarcity of resources.

3.5.2 Feature Extraction

Feature extraction converts the text to a form that can be easily used by machine learning models, where text data is converted into real numbers. This kind of project applies the use of TF-IDF method, which calculates the relevance of a given term within the document against the whole text data set.

TF-IDF Vectorization: Hence, as a result of considering only the most important words in the document and weighting the distribution of words in a document it's frequency across the corpus, TF-IDF is chosen. This technique underlines those words which appear in specific reviews and which contain much sentiment load.

N-grams: Besides, unigrams it also uses bigram (pair of words) or trigram (set of three words) which helps in identify word sequence important in tracing the sentiment of Bangla text.

3.5.3 Model Selection

For accuracy there are multiple machine learning algorithms integrated each of which has its own approach to text classification. The models used include:

Decision Tree Classifier: The Decision Tree Classifier is taken as baseline model since it is comparatively easier for interpretation. In sentiment analysis we could make use of Decision Tree classifier for knowing whether which particular word or which combination of words (features) looks like deciding on which sentiment. The classifier chooses with its own criterion the subset of features (like words or phrases in Bangla reviews) that define the split at each node. For instance, a word like 'অসাধারণ'meaning excellent will be helpful for a positive sentiment classification. Decision Trees are useful in text classification if interpretability is of essence. This makes it possible to know how some words in Bangla impact sentiment estimation. Another disadvantage of Decision Trees is that they suffer overfitting, this happens more often when Decision Trees are exercised on small datasets or datasets containing noise or on some other datasets that have not been used in exercising the

Decision Trees. This limitation is controlled in the project by varying the tree pruning or by the number of nodes in the trees.

Random Forest Classifier: Random Forest expands the concept of Decision Trees by forming a set of branched models trained in different sample data sets. This as well minimizes the chances of overfitting while improving accuracy during model building. Random Forest in the project uses bagging (Bootstrap Aggregation) that is the training of each tree is on a different subset of the training data. This technique increases the model's efficiency by decreasing the variance and thus preventing over fitting. In each split of any tree only a randomly selected subset of features (words) is used in the decision process. This is useful in order to prevent the model from stressing individual characteristics, thereby achieving a balanced model. For instance, specific words or n-grams relevant only to specific sentiment (e.g., "ভালি") for positive are highlighted without prejudice. Random Forest situation is advantageous in text classified tasks because it operates well with high-dimensional inputs such as TF-IDF and does not have a problem with classes imbalance. Because the proportions of some categories of sentiment classes may be skewed (for example, positive sentiment is more numerous than negative) Random Forest contributes to enhance classification accuracy.

Support Vector Machine (SVM): SVM is among the most preferable classification techniques in text classification and renders efficient results for sentiment analysis. SVM aims to find such a hyperplane which will separate the classes as far as possible from each other (positive, negative sentiments). Due to the fact that it creates the highest margin between classes, SVM can work with tough over-lying patterns in Bangla reviews. SVM is ideal for using with other features such as the TF-IDF along with n-gram features which result in the production of a large dataset of features. As for using Bangla text, the relatively small number of features in the form of TF-IDF data (most words are distinctive to the particular reviews) correspond to the strength of SVM. SVM does not tend to overfit the dataset, especially with large datasets containing high dimensions features, which is promising for using SVM on Bangla reviews with different word choice and structures. SVM is also used when sentiment-related features are not

clearly defined and which may take slightly different values in a slight way. When working with large data sets, SVM can be quite resource expensive because it goes through an optimization process. Nonetheless, the performance of SVM is satisfactory in this project provided the parameters of SVM are well tuned and has been regularized.

Stochastic Gradient Descent Classifier (SGD): SGD Classifier is used for its effectiveness and suitability for large scale data sets with sparse features as are derived from TF-IDF in text classification. Instead of computing the new model parameters from the scratch SGD adjust the current parameters a little at each training example by updating it, making it faster and more memory efficient. This is particularly helpful when all the Bangla reviews are grouped which are usually large in number. In this project the SGD Classifier is applied together with linear models including logistic regression, linear SVM for quick learning. This shows that SG does not suffer from slow convergence problems in high-dimensional spaces, as is often the case with TF-IDF features extracted from the Bangla text. Since customer's reviews come in different aspects including length and density of the work, the incremental nature of SGD is good for handling the variation. Many learners in SGD are dependent on factors such as learning rate control and wrong setting of the hyperparameters may lead to conjugate solution to the problem or slow convergence rates.

XGBoost Classifier: Extreme Gradient Boosting (XGB) is an advanced, highly efficient boosting algorithm, known for its excellent performance on structured data and text classification tasks. XGBoost works by training sequential trees where each new tree corrects the errors made by previous ones. This sequential learning enables the model to focus on hard-to-classify examples, which improves accuracy and generalization. XGBoost includes built-in regularization (both L1 and L2), which helps control overfitting. Given the risk of noise in customer reviews, especially with diverse Bangla vocabulary, regularization is crucial for creating a balanced model. Bangla customer reviews may contain uneven class distributions (e.g., more positive than negative reviews). XGBoost has several parameters, such as scale_pos_weight, that help balance class weights to improve model

performance in unbalanced datasets. XGBoost is robust against overfitting and performs exceptionally well on imbalanced datasets making it a strong candidate for sentiment classification tasks. It also allows the use of both numerical and categorical data making it versatile for a wide range of features extracted from Bangla text.

3.6 Model Training

Model training as mentioned above is an imperative stage that is involved in any machine learning project where algorithms determine probabilities on new unseen data using that which has already prevailed. For this project, multiple classification algorithms were used: These are Decision Tree Classifier, Random Forest Classifier, SVM (Support Vector Machine), SGD Classifier, XGB Classifier. Both models were tested using preprocessed and balanced data in gender and age to avoid bias and improve stability.

3.6.1 Training Pipeline Setup

The training pipeline is initiated by using train_test_split which divide the raw data into training and testing datasets so it can determine the performance of the model against the new data it has not been trained on. Oversampled data is used in this case to help balance the classes in order to get a more appropriate data experiments or data set to work on during training.

3.6.2 Training Individual Models

Each model follows a similar training structure, yet leverages unique properties:

Decision Tree Classifier: This algorithm entails creation of a tree based on division of data with an aim of finding feature values that define the decision-making rules. Linear regression is easy to understand easy to interpret and works well with non-linear data as well.

Random Forest Classifier: A set of decision trees of a group where the final prediction is performed after a set of trees delivers its results in order to avoid overfitting.

SVM (Support Vector Machine): It works with finding the best separating hyperplane between classes in a high-dimensional space. This one is suitable for classification duties where samples enjoy clear margins that separate them.

SGD Classifier (Stochastic Gradient Descent): Boasts of an iterative mechanism for enhancement of the model ideal for large scale learning.

XGB Classifier: An enhanced algorithm specifically used in gradient boosting to make virtual explicit predictions of the response variables by learning the residuals in sequence. It is described for its strong performance and dynamicity.

3.7 Implementation Requirements

- Different Machine Learning Frameworks and Libraries
- Windows 11
- Google Colab with runtime TPU
- Datasets
- Google Drive
- Streamlit

CHAPTER 4

Result Analysis and Discussion

4.1 Introduction

In this chapter we investigate how accurately different machine learning classifiers predict the sentiment of Bangla customer reviews. Sentiment analysis or opinion mining is a difficult task and this becomes more compounded when dealing with languages as such Bangla because of limited language resource and complexity as far as sentiment expressions are concerned. In this chapter the authors explain the results of five models which include Decision Tree, Random Forest, Support Vector Machine (SVM), Stochastic Gradient Descent (SGD) Classifier, and XGBoost Classifier based on the area of accuracy, precision, recall, F1-score, confusion matrix, and classification report. Furthermore, to check for overfitting or underfitting training and validation accuracy. Our main goal in this paper is to compare such models to determine which classifier will be most suitable for Bangla sentiment classification. It all supports the final step that is the selection of the best model which in turn improves the chances of gaining improved understanding of the customer sentiments.

4.2 Experiment Results and Analysis

This section explains the performances of all classifiers used in the project which are Decision Tree, Random Forest, SVM, SGD and XGBoost. Embedded below is a sample table comparing the accuracy and F1-score of each classifier discussing the performance of each model.

TABLE 4.2: The Experimental Result of the Evaluated Model

Machine Learning Model	Test Accuracy	recall	F1-Score	Precision
Decision Tree	81%	0.84	0.82	0.79
Random Forest	86%	0.91	0.87	0.84
SVM	89%	0.93	0.90	0.87
SGD	89%	0.92	0.90	0.87
XGBoost	85%	0.89	0.86	0.83

4.3 Generating Confusion Matrix

The confusion matrix gives the user a clear view of its model by outlining correct and wrong for each class of prediction.

4.3.1 Decision Tree

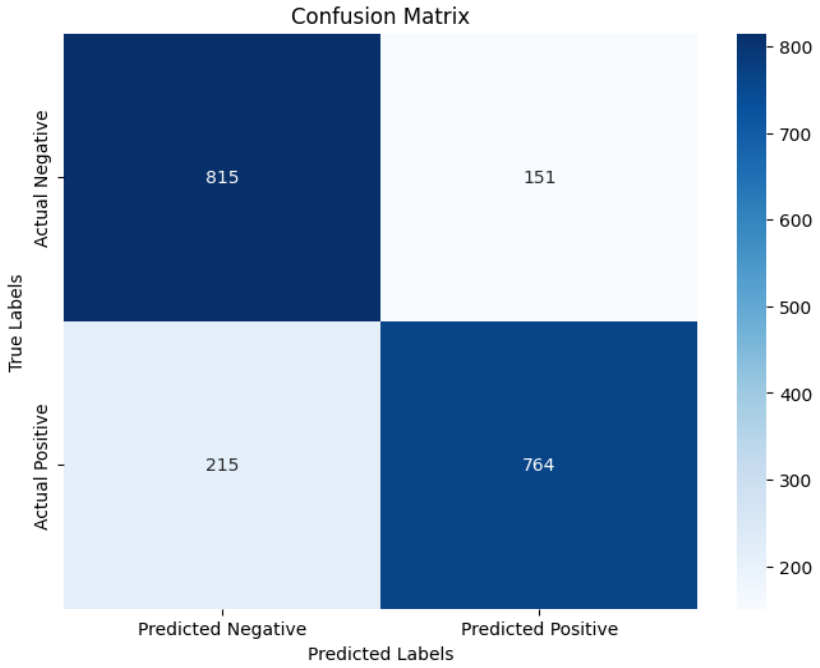


Figure 4.3.1: Confusion Matrix for Decision Tree Classifier

The confusion matrix is a 2x2 matrix that summarizes the performance of a classification model:

True Negative (TN): The quantity of negative instances that have been correctly classified by the model (815).

False Positive (FP): The number of wrongly classified samples with positive class (151).

False Negative (FN): The number of wrongly classified as negative samples (215).

True Positive (TP): The number of correctly predicted example of positive experiences (764).

To represent the matrix, it is widely used a heatmap, in which the color intensity in the cell indicates the number of cases. The x-axis usually holds the values of the predicted labels, on the other hand, the y-axis shows the values of the true labels.

4.3.2 Random Forest

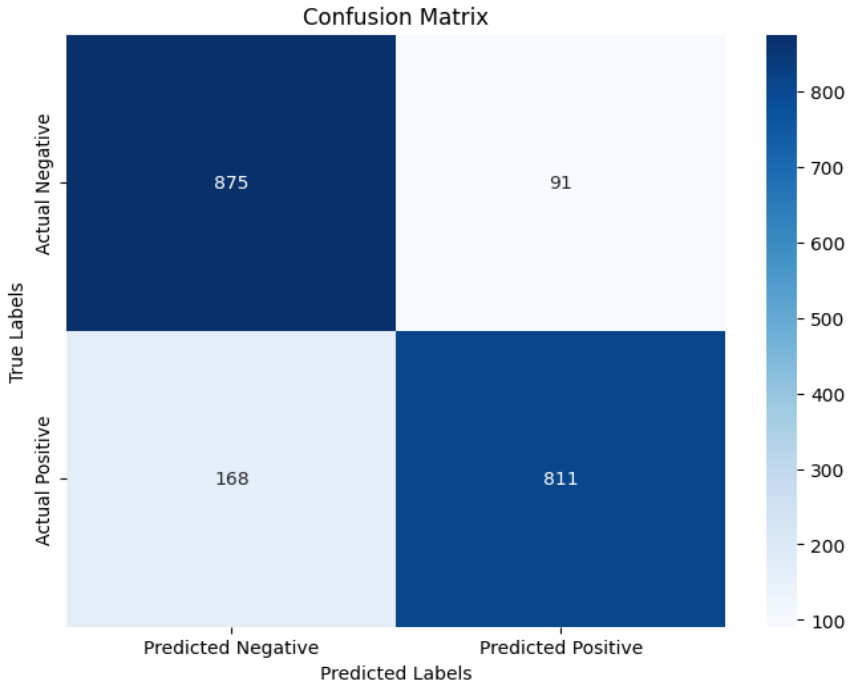


Figure 4.3.2: Confusion Matrix for Random Forest Classifier

There are many performance measurements for classification models and all of them are illustrated in the confusion matrix. It is a two-by-two matrix used in classification of two categories with binary outcomes that illustrates TP, TN, FP and FN. Here's a brief description of each element:

True Negative (TN): The number of times the instances are correctly predicted as negative (875).

False Positive (FP): The number of over-predictions of positives, that is, the instances that were predicted to be positive while actually they are negative (91).

False Negative (FN): The third type of error when only the test instances are used is the number of samples that were predicted to be negative but are actually positive (168).

True Positive (TP): The number of corresponding passed checks, i.e. the instances are positive in this case (811).

4.3.3 Support Vector Machine (SVM)

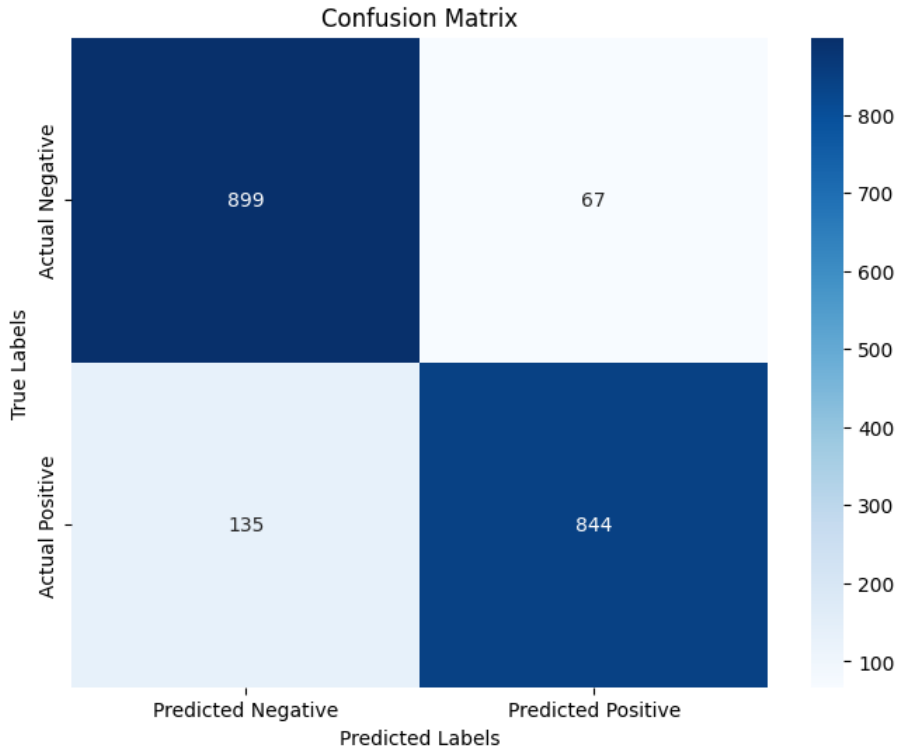


Figure 4.3.3: Confusion Matrix for SVM Classifier

Confusion matrix is a performance measure used on models that employ classifications. It consists of four key components:

True Negative (TN): The number of times the text under test is accurately classified as negative (899).

False Positive (FP): The total count of negative instances that where mistakenly flagged as positive (67).

False Negative (FN): The number of cases that are wrongly classified as negative (135).

True Positive (TP): They are the number of times that the algorithm accurately identified the instances as positive (844).

4.3.4 Stochastic Gradient Descent (SGD) Classifier

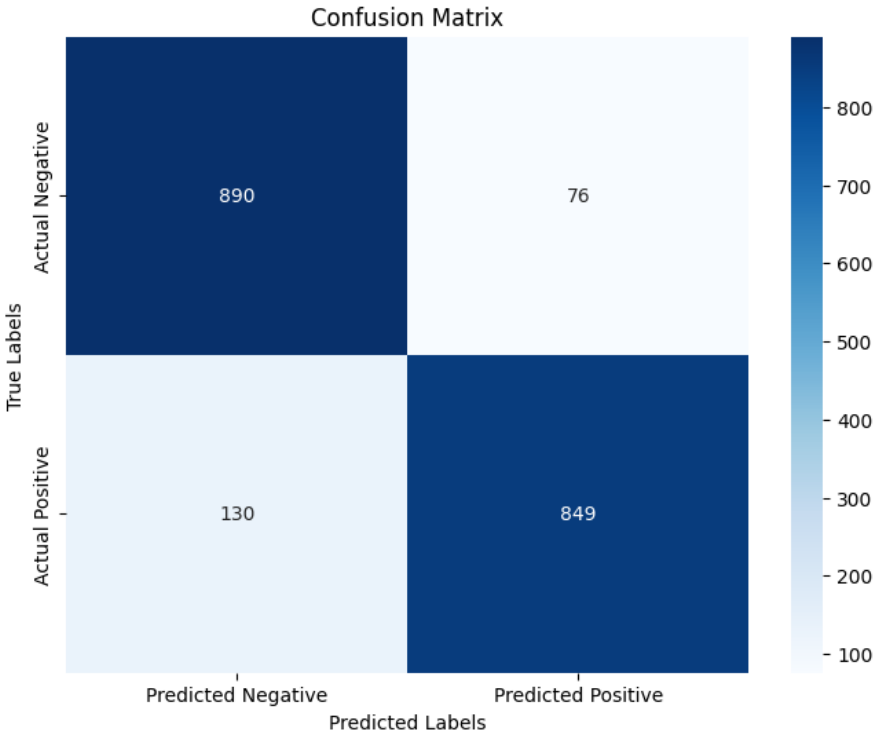


Figure 4.3.4: Confusion Matrix for SGD Classifier

A confusion matrix is a table that depicts the actual distribution of a model's classification. It useful in visualizing the operations of an algorithm more closely when the two classes are unbalanced.

True Positive (TP): The actual counts where about the number of cases where the model was right about the positive class (849).

False Negative (FN): The observations where model has wrongly classified the negative class (130).

False Positive (FP): Those for which the model assigned the positive class label while the actual label was negative (76).

True Negative (TN): Incorrect positive class, Instances of negative class correctly classified by the model (890).

sentiment analysis. It will be based on it to drive the identification of areas that require enhancements and choices of the right model for implementation.

4.3.5 XGBoost Classifier

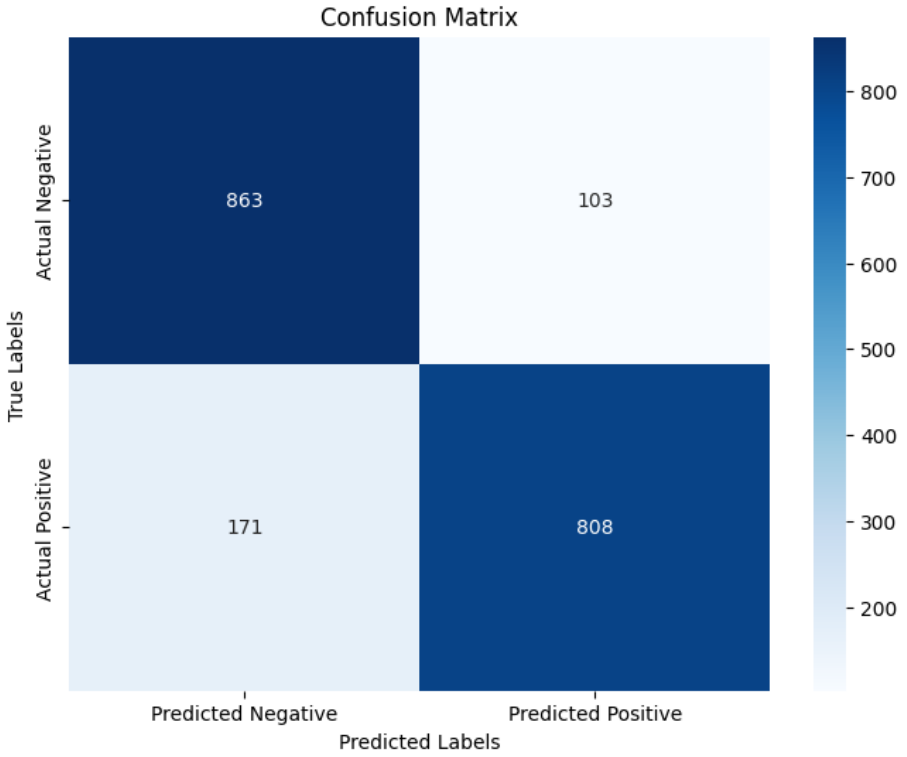


Figure 4.3.5: Confusion Matrix for XGB Classifier

Confusion matrix is a convenient tool which is used to assess the efficiency of the classification model together with the real and estimated classes. It is useful to determine the total of right and wrong predictions incorporated by the model. Here's the structure:

True Positive (TP): Positive cases as were accurately predicted (808).

False Negative (FN): Number of positive instances accurately classified as negative (171).

False Positive (FP): False positive which means the cases were negative but classified as positive (103).

True Negative (TN): Its negative cases specifications were positively predicted (863).

4.4 Generating Classification Report

The classification report is more informative as it displays the result of each model in terms of precision, recall and F1-score by class. All these metrics will provide information about how well the given model performs in case of different tendencies in the sentimental analysis.

TABLE 4.4.1: All Algorithm accuracy results

Algorithm	**Class**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
SVM	negative	0.84	0.91	0.87	0.89%
	positive	0.90	0.83	0.86	
SGD	negative	0.87	0.92	0.90	0.89%
	positive	0.92	0.87	0.89	
Random Forest	negative	0.84	0.91	0.87	0.86%
	positive	0.90	0.83	0.86	
XGB	negative	0.83	0.89	0.86	0.85%
	positive	0.89	0.83	0.86	
Decision Tree	negative	0.79	0.84	0.82	0.81%
	positive	0.83	0.78	0.81	

4.4.1 Performance/ Comparative Analysis

In the last part under the Comparative Analysis we present the outcome of each model with regard to accuracy, precision, recall and F1 score. This gives information on the effectiveness of each model and the areas of shortcomings hence suitable for choosing the optimal model for implementation. This code produces a bar chart of these values for all the models. The following code creates a bar chart of these for all the models.

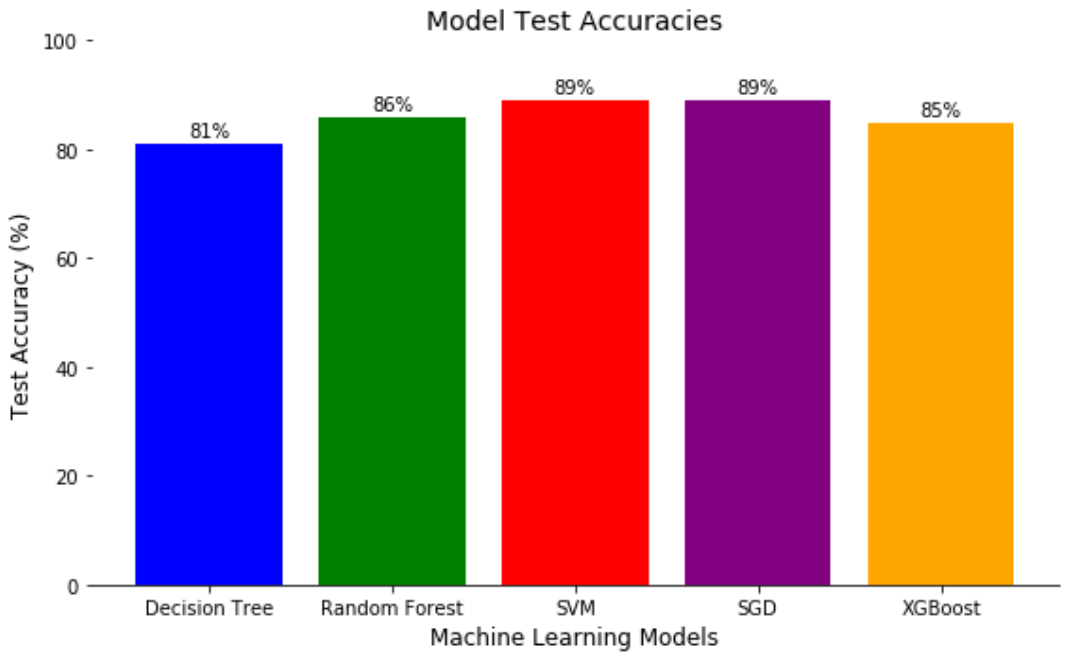


Figure 4.4.1: Comparative Analysis of Model Performance Metrics

The bar chart allows comparing all models with each other using all performance measures known and indicating which model performs best for each of them. For example, if among the classifiers mentioned above, SVM and SGD has the highest accuracy and F1-score this will mean that this classifier is preferable for our application. This visualization enables one to make the comparison in a very simple way and this makes the stakeholders understand the results well.

4.4.2 User Interface:

Figure 4.4.2.1: Predict negative sentimental text

Figure 4.4.2.2: Predict positive sentimental text

4.5 Discussion

Below the results obtained are discussed using the perspective of model performance and highlighting its features. Observations might include:

SVM Performance: The algorithm shows that SVM has a better presentation of accuracy and F1-Score scores over the other evaluation metrics which are used to state that this model has the ability to manage the complexity of Bangla text sentiment analysis.

Overfitting Analysis: Checking the learning curves allows to determine whether any of the models had to be stopped due to overfitting and as a result of which the model performs well on the training material but poorly on the validation material.

Confusion Matrix Insights: The confusion matrix used in the study can be useful also to identify cases where specific types of reviews are not predicted well by the model for example very negative reviews may be predicted as positive and vice versa.

Combined with the visuals and direct comparison of the metrics, the analysis given in this chapter gives a detailed idea about the potential and the issues of each model in Bangla sentiment analysis. It will be based on it to drive the identification of areas that require enhancements and choices of the right model for implementation.

CHAPTER 5

Impact on Society, Environment, and Sustainability

5.1 Impact on Society

The Bangla customer review's sentiment analysis using machine learning is a beneficial factor for society. In a way customer sentiment analysis works to enhance the quality of services provided, resolve complaints lodged and initiate the production of services or products which can address social needs and wants. In a broader context, the use of sentiment analysis tools makes it easier for businesses in Bangladesh or any other developing country to level the playing ground of getting good and reliable analysis that would help them to operate fully online businesses thanks to the advanced insights that such analytic tools offer. Additionally sentiment analysis in Bangla help in archiving the language and when there are new applications such as data science Bangla speaking people can easily access them. It creates opportunities for other Bangladeshi segment to contribute towards online forums, thereby getting their thoughts, comments and feedback quantified fairly. This can encourage social inclusion and participation of the people as they are given an opportunity to exercise their freedom of speech and be understood in their own languages.

5.2 Impact on Environment

Big data models including machine learning models present the challenge of being computer intensive which may lead to environmental outputs. In recent years, there has been the realization that model training consumes a lot of energy and as such machine learning models, as well as the use of large-scale training sets in model training are environmentally unfriendly. Comparing to baseline and neural models, methods like SVM and SGD are computationally intensive in this project. In the following some suggestions on possible measures that can be taken ecologically in the given project can be mentioned the main ones are the following: during the creation of training models, one can use cloud platforms with data centers, which use the green energy. moreover, it is possible to train fewer complex models which consume less energy but make no significant difference in

the final results. Furthermore, it can also help to reduce computational load associated with language model by using transfer learning or pre-trained embedding in some cases could further reduce computational expenses and is good for the environment as well as the financial field.

5.3 Ethical Aspects

Ethics is a bigger factor in the creation of tools for sentiment analysis especially when dealing with customer information and being fair on the outputs. Several ethical considerations arise in this project:

Privacy: Prominent hygiene should be observed in relation to the user reviews to eradicate individual identification when the method is in practice.

Bias in Data and Model: He further explains that machine learning models mirror bias if they occur in the training dataset and will make unfair or wrong predictions. For instance, if the semantic data is biased then the sentiments which may be attributed to special dialects or contexts could be wrong. By ensuring equal data distribution and checking how often the model is wrong gives this risk some buffer.

Transparency: In the models there must be the clear way of thinking of approaching the sentiment classification and every decision made must be expounded. This is additive to fortifying credibility among users of these SAH's for authoritative purposes for the sake of decision-making.

Introducing sentiment analysis systems in Bangla has an extra touch of ethicality to it as it gives the people within Bangladesh an opportunity of using digital platforms without necessarily discarding their root language.

5.4 Sustainability Plan

To this end it is crucial to introduce ideas that would help to maintain the ongoing effectiveness and development of this project over time and its scalability. A sustainable plan for this sentiment analysis system would involve the following components:

Regular Model Updates: Language also dynamically changes and new trends in customer sentiment may pop up after some time. To keep its relevance it is needed

that the sentiment analysis models need to be refined periodically with newer data containing the newer words, expressions and context of Bangla.

Energy-Efficient Practices: As aforementioned energy efficiency is a consideration in an endeavor to ensure that the project impacts the environment in a minimum way possible. Such techniques as reducing recurrent transmissions of a model using less resource-prolific algorithms for updates operationalizing energy-saving hardware can be used to realize this.

Community Involvement: With business and end-users interacting with the system continual refinement can be carried out to the models to align with the Bangla-speaking communities rather than giving precedence to the existing biases. A community-driven approach also makes it possible for the system to retain relevance with what the society actually uses in their day-to-day language.

Open-Source Contributions: Some of the aspects of the project can be made available for public use in order to make contribution from other researchers and developers in improving the efficiency of the tool hence encouraging innovation and cutting the number of costs incurred by the developers.

CHAPTER 6

Overview of the Study, Conclusion, and Future Work

6.1 Overview of the Study

This work has been specifically tailored to conduct sentiment analysis on Bangla customer's reviews with great employ of BNLP and ML. To initiate the customer sentiment analysis process, different algorithms, including Decision Tree Classifier, Random Forest Classifier, Support Vector Machine (SVM), Stochastic Gradient Descent (SGD) Classifier, and XGBoost Classifier were tested in order to determine the most appropriate model to use in order to obtain a high level of accuracy. The Bangla sentiment data set was gathered resolved and normalized then these steps were performed Feature engineering and oversampling. Using quantitative measures such as accuracy, F1-score, precision and recall for each model, this study brings out an understanding of the effectiveness and constraint of machine learning for sentiment analysis in Bangla. The work also underscores the importance of applying sentiment tools for Bangla increasing the availability of language resources for minorities and improving the decision making of companies focusing on the Bangladeshi market.

6.2 Conclusions

Through the present work it can be justified that machine learning methodologies can be appropriately applied to sentiment analysis of Bangla customer reviews. For all the algorithms tested their performances differed and XGBoost and Random Forest were preferred due to their stability in dealing with the complexity. This research also shows the possibility of employing sentiment analysis as a competitive advantage for organizations to learn about customer's responses in real-time to show adjustments to service or goods. Furthermore, the conclusions stress that language independent works such as the presented paper need to progress the identification of features in underrepresented languages like Bangla because it extends the possibility to engage Bangla speaking populations in digital analysis practices. Hence the project directly benefits the digital ecosystem of Bangladesh by offering a chance of better customer interactions and offering the right kinds of understandings that could help in making superior decisions.

6.3 Limitations

Several limitations were encountered during this study which affected both the methodology and the outcomes:

Dataset Limitations: As for the dataset used for training and testing the models there were some constraints, for example a sufficient number of balanced samples in all the given sentiment classes. Besides, oversampling was used to deal with class imbalance while constructing a more diverse, larger dataset would improve model portability

Computational Constraints: Building up SVM and SGD training machine learning models was computationally intensive, thus we could not delve deeper into enhancing the model complexity as well as using ensemble methods. Subsequent enhanced versions of the same hardware could present better performance to the model and shorter training sessions.

Language Nuances: Just like any other Bangla, the text writing has some of the features that can pose difficulties to most NLP approaches. Some of the facets that had challenges with the models include, Use of sarcasm Contextual meanings Mixed Language Communications, Use of both Bangla and English. This led to certain amount of reduction in the accuracy of the sentiment judgements that were output in our programmed.

Generalization to Other Domains: As this model target on Bangla customer reviews it is less generalizable to the other text data where we cannot find such as news articles or social media content. However to widen the utilization of the model further training on reference domain data is necessary.

6.4 Future Work

To build upon the results and address the limitations encountered several future directions are proposed:

Expanding the Dataset: Obtaining a bigger and a more variants sample will allow to increase the precision of given models and enhance generalization. The authors could actually extend the corpus with more variations of Bangla dialects and with

the most frequently used phrases to enhance the interpreter's ability to recognize colloquial language.

Integration of Deep Learning Models: Although feature extraction can be improved, perhaps future work could extend the model, incorporating tuples into deep learning methods like recurrent neural networks (RNN), or long short-term memory (LSTM) or more recent structures in natural language processing. Conceptually, these models should cope with context and long-term dependencies in a language, which can impact the sentiment classification.

Hybrid Model Development: It is hypothesized that when merging the results of using machine learning and utilizing deep learning to perform sentiment analysis a more extensive system would be created to cover all the subtleties of sentiment. Probabilistic fusion could combine multiple classifier's characteristics and improve efficiency in various structures of the language.

Real-Time Sentiment Analysis Implementation: A valuable extension would be to use the model as an online sentiment analysis tool. This would entail the establishment of a large customizable API for usage by businesses and users to track Bangla customer sentiment in real time.

Incorporation of Cultural Nuances and Sarcasm Detection: The future work can be the addition of the NLP techniques for sarcasm detection and the analysis of sentiment in a given context. This would make the tool more valuable for the businesses that want to gain more detailed understanding of Bangla-speaking customer's behaviors and attitudes.

Cross-Language Sentiment Analysis: A lot of Bangla speakers switch between Bangla and English when giving reviews. It is highly valuable to extend the model to work with code mixing data such as Bangla English because it approximates the real language usage by the population. Efforts to refine approaches for dealing with mixed languages would expand the applications of the model in the Bangladeshi online environment.

References

[1] Rahman, M., et al. "Sentiment Analysis of Bangla Movie Reviews using SVM with TF-IDF Vectorization." Journal of Language Processing, 2020.

[2] Hossain, F., et al. "LSTM for Sentiment Analysis in Bangla Customer Reviews." Asian Journal of Computer Science, 2021.

[3] Alam, M., et al. "Naive Bayes Classifier in Bangla Text Sentiment Analysis." International Journal of Data Mining, 2019.

[4] Khan, I., et al. "CNN-based Bangla Text Classification." Computer Vision and Language Processing Journal, 2020.

[5] Ahmed, J., et al. "Integrating Word2Vec and SVM for Sentiment Analysis." Proceedings of the NLP Conference, 2022.

[6] Chowdhury, R., et al. "Random Forest and Ensemble Methods in Bangla Sentiment Analysis." Machine Learning Journal, 2018.

[7] Akter, S., and Nasrin, M. "Logistic Regression and Decision Trees for Product Review Sentiment." Bangla NLP Journal, 2020.

[8] Sarker, A., et al. "Enhanced Sentiment Prediction using Bi-LSTM in Bangla." Journal of Artificial Intelligence Research, 2021.

[9] Mollah, T., et al. "Using XGBoost for Bangla Customer Feedback Sentiment Analysis." Data Science Journal, 2019.

[10] Hasan, M., and Alam, R. "Comparative Analysis of KNN and SVM in Sentiment Analysis." International Journal of Machine Learning, 2017.

[11] Rahim, A., and Parvin, T. "The Role of GloVe Embeddings in Bangla Sentiment Analysis." Proceedings of the International Language Processing Conference, 2020.

[12] Fahim, R., and Ahsan, U. "Multilevel Hybrid Deep Learning Models for Sentiment Analysis." Asian NLP Studies, 2021.

[13] Islam, N., et al. "BERT-Based Transfer Learning for Bangla Sentiment Classification." Natural Language Processing Journal, 2022.

[14] Akash, M., et al. "LSTM Techniques for Bangla Language Sentiment." Bangla Computational Linguistics Journal, 2018.

[15] Sultana, F., and Karim, R. "Comparative Study on Neural Networks and Random Forest for Bangla Sentiment." Machine Learning in South Asia, 2019.

[16] Kamal, S., et al. "Effectiveness of Attention Mechanisms in RNN-based Bangla Sentiment Analysis." Journal of Language Modeling, 2021.

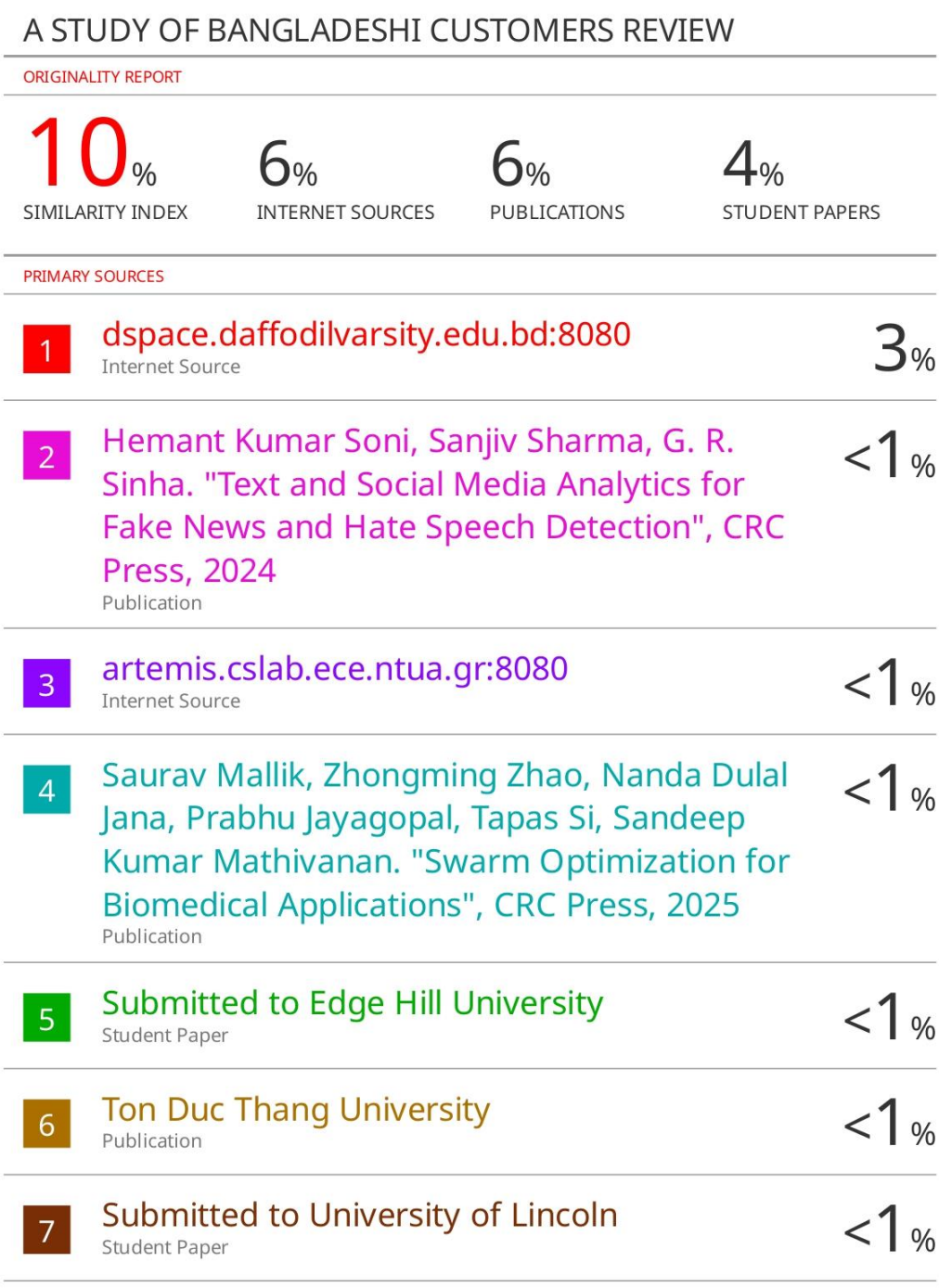
[17] Rashid, N., and Parvin, K. "Ensemble of Naive Bayes and SVM for Improved Bangla Sentiment Analysis." Computer Science and Information Systems, 2020.

[18] Nahar, A., et al. "Incorporating Sentiment Lexicons in Bangla ML Models." Journal of Sentiment and Text Mining, 2018.

[19] Mahmud, L., et al. "Developing an Annotated Dataset for Bangla Sentiment Analysis." Bangla NLP and AI Review, 2020.

[20] Mitu, P., and Alam, M. "Combining LSTM and CNN for Bangla Text Sentiment Analysis." Journal of Data Engineering, 2022

Plagiarism Report



A STUDY OF BANGLADESHI CUSTOMERS REVIEW

ORIGINALITY REPORT

10%	6%	6%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	Hemant Kumar Soni, Sanjiv Sharma, G. R. Sinha. "Text and Social Media Analytics for Fake News and Hate Speech Detection", CRC Press, 2024 Publication	<1%
3	artemis.cslab.ece.ntua.gr:8080 Internet Source	<1%
4	Saurav Mallik, Zhongming Zhao, Nanda Dulal Jana, Prabhu Jayagopal, Tapas Si, Sandeep Kumar Mathivanan. "Swarm Optimization for Biomedical Applications", CRC Press, 2025 Publication	<1%
5	Submitted to Edge Hill University Student Paper	<1%
6	Ton Duc Thang University Publication	<1%
7	Submitted to University of Lincoln Student Paper	<1%