SENTIMENT ANALYSIS OF CUSTOMER, BASED ON CUSTOMER REVIEWS IN BANGLA LANGUAGE USING MACHINE LEARNING

BY

SHANTONU SAHA

ID:211-25-946

This Report Presented in Partial Fulfillment of the Requirements for

The Degree of Master of Science in Computer Science and Engineering

Supervised By

Md. Sadekur Rahman

Assistant Professor

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

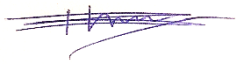
DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Project/internship titled **Sentiment Analysis of Customer, Based on Customer Reviews in Bangla Language Using Machine Learning,** submitted by **Shantonu Saha**, ID No: **211-25-946** 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 **22/01/22**.

BOARD OF EXAMINERS

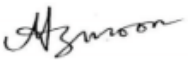


Chairman

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

Internal Examiner

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



Internal Examiner

Nazmun Nessa Moon
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

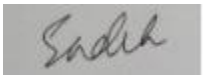
External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

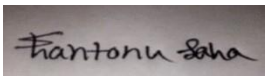
We hereby declare that, this thesis has been done by us under the supervision of **Md. Sadekur Rahman, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree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

Supervised by:



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

Submitted by:



Shantonu Saha
ID: 211-25-946
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

At first, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final thesis successfully.

We really very grateful and wish our profound our indebtedness to **Md. Sadekur Rahman**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep knowledge and keen interest of our supervisor in the field of "Natural Language Processing and Machine Learning" to carry out this thesis. His scholarly guidance, patience, constructive criticism, motivation, constant and energetic supervision, continual encouragement, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan, Professor & Head,** Department of CSE, for his motivation and appreciation. We are also very thankful to other faculty members 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 are very thankful to our parents and friends who were always motivate and criticize our work in a manner to improve our work. At least we thank all of them from the core of our heart.

ABSTRACT

Artificial Intelligence had made a revolution in our all aspect of life using machine learning techniques. It helps to understand hidden pattern of data and hence improve businesses. As a result, it is now widely used in analyzing customers preview to understand the behavior or demand of themselves which can beneficiary to any business owner. The aim of our research is analyzing customer reviews given in Bangla language. In order to conduct our research, we have first collected data from different online sources those have focused on food services. Later, these data were preprocessed adopt with six machine learning classifiers namely Multinomial Naïve Bayes, Random Forest, Decision Tree, Support Vector Machine, Extreme Gradient Boosting, K-Nearest Neighbors. Out of these KNN outperformed all other classifiers with an accuracy of 65%.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction:

In decision-making process it's really important "what other people thinking?". Nowadays, we are watching people debating about so many products on the different type of social media platforms and others web platforms. One of the most common topics is food when they are discussing. On different food delivery websites and on others web platforms, customers put reviews about restaurant, foods, and their delivery systems. Those reviews of those platforms help companies to make their future decisions. But it is too hard to handle this much information manage manually. Soo, they need an automated system for managing that information and making a decision easily. The thing that performs better here is sentiment analysis. If we take their customer feedback as an example, sentiment analysis (a form of text analytics) measures the attitude of the customer towards the aspects of a service or product which they describe in the text [1]. This typically involves taking a piece of text, whether it's a sentence, a comment, or an entire document, and returning a "score" that measures how positive or negative the text is [1]. So, in this sentiment analysis work, we have collected a lot of customer reviews from different platforms for training datasets. After applying all possible filtering methods, we take those as input and classify customers' emotions. This can help to enhance customers' experiences and improve customer service.

1.2 Motivation

Sentiment analysis approaches are important to identify and realize customers' emotions. Companies can enhance customers' experience and improve their customer service by generating

insights Using the sentiment analysis process. For sentiment analysis, customer reviews are important to identify customers' emotions. Customer reviews can help to improve customers' services and product quality. By analyzing customers reviews we can measure customer satisfaction. After a huge observation, I saw that on this site a lot of research work has been done on customer reviews which are in English. There are very limited works on Bangla customer reviews. I also saw that a limited amount of data is mainly responsible for that limited research work. This is my main motivation for choosing this topic and learning how to create Bangla datasets to do machine learning. And another thought of this paper was to observationally assess the elements influencing the decisions of customers while purchasing food from the website or food delivery app. The objective was to investigate customers' conduct in the arising business of online food delivery business in an arising economy, Bangladesh.

1.3 Rationale of The Study

The idea of online food delivery is extremely new in Bangladesh yet has acquired a ton of prevalence in its beginning phases. The prime objective of the research work is to identify the factors critical to success for online food delivery services in Bangladesh and enhance customer experiences.

1.4 Research Questions

The following research questions were adopted when doing research work:

Q1: may I collect Bangla customer's reviews?

Q2: how may I process those customer reviews?

Q3: which algorithm can justify positive or negative reviews from the Bangla texts dataset?

1.5 Expected output

A model that can classify and predict the level of customers’ reviews from a given Bangla text.

1.6 Project Management and Finance

I have financed in my own research work.

1.7 Report Layout

In our report we have total 6 chapters

- In Chapter 1 we mention our whole research work's outline and divided this chapter into multiple subchapters. For example, introduction, motivation, rational of the study, research question and expected output of our project.
- In Chapter 2 we have discussed about the previous work on Bengali text classification, the scope of the problem and challenges in this work.
- In Chapter 3 we will talk about our work procedure, methods and techniques to build a Bengali food review prediction model.
- In Chapter 4 we will discuss about the Experimental Results and Discussion of our build model.
- In Chapter 5 we will talk about the Impact of Society, Environment, Ethical Aspects and Sustainability plan of our work.
- In Chapter 6 we have discussed about the Summary, Conclusion and Further Study of the work.

CHAPTER 2

BACKGROUND

2.1 Preliminaries/Terminologies:

Pre-processing:

preprocessing in Machine Learning is a critical advance that helps upgrade the nature of the information to advance the extraction of significant experiences from the information. Information preprocessing in Machine Learning alludes to the procedure of planning (cleaning and coordinating) the crude information to make it reasonable for building and preparing Machine Learning models. In basic words, information preprocessing in Machine Learning is an information mining method that changes crude information into a justifiable and clear arrangement.

With regards to making a Machine Learning model, information preprocessing is the initial step denoting the inception of the interaction. Commonly, real-world information is deficient, conflicting, incorrect (contains mistakes or anomalies), and frequently needs explicit property estimations/patterns. This is the place where information preprocessing enters the situation – it assists with cleaning, designing, and putting together the crude information, subsequently preparing it to-go for Machine Learning models.

Tokenization:

This method involved with separating a piece of text into little units called tokens. A token might be a word, part of a word, or simply characters like accentuation. It is one of the most basic NLP tasks and a troublesome one, in light of the fact that each language has its own linguistic

developments, which are regularly hard to record as rules. Tokenization characterizes what our NLP models can communicate.

Digit removal:

An overall text record might contain just as many digits. Yet, as significant words don't contain digits, eliminate these digits by utilizing their Unicode portrayal.

Punctuation removal:

Eliminating special characters (&, @, #, $ etc.), symbol ({;}, {:}, {“ ”}, {^}, {<}, {>} etc.), Emojis (😭, 😮, 😋, 😍 etc.), from text data. Reduced excess use of spaces, tabs, shift, from texts data.

Stop words removal:

This procedure is the most usually utilized preprocessing venture across various NLP applications. The thought is basically eliminating the words that happen ordinarily across every one of the archives in the corpus. Ordinarily, articles and pronouns are by and large named stop words. eliminate these words from the text credentials.

Stemming:

By using the stemming process reducing a word to its base form. This is an important part of pipelining process in natural language processing.

Feature extraction:

By using feature extraction process decrease the number of properties needed for processing without losing significant or appropriate data. Using this method feature extraction procedure

become easier from those words. The assortment of words that are left in the report after that multitude of steps are considered as a conventional representation of the record.

Training and testing:

In this stage a dataset, that will use to prepare for calculation, remember that-piece of the information will be utilized to check how well the training goes. This implies that your information will be parted into two sections: one for training and the other for testing.

Performance Measure:

There are lots of evaluation metrices when comes the text classification part. In this project used the performance measure including precision, recall and F-measure.

Precision:

By using this measures the quantity of positive class predictions that actually belongs to positive class.

Recall:

This procedure measures the quantity of positive class predictions made out of all positive examples in the dataset.

F-Measure:

This step delivers a single score that balances both the concerns of precision step and recall step in one number.

Accuracy and training time:

After applying all possible procedures need to checked the accuracy and training time of all datasets. How many times it takes train classifiers is another important fact. Low training time is always preferable here.

2.2 Related Works

Automatic text classification has always been an important application and research topic since the inception of digital documents. Text categorization is an active research area of text mining where the documents are classified with supervised, unsupervised or semi-supervised knowledge. Among various machine learning approaches in document categorization, most popular is supervised learning where underlying input-output relation is learned by small number of training data and then output values for unseen input points are predicted. Various number of supervised learning techniques are Neural Network, K-Nearest Neighbour, Decision Tree, Naïve Bays, Support Vector Machine, and N-grams, has been used for text document categorization. Some of the important works done on Bangla Language focused on text classification is analyzed in the next few paragraphs along with their limitations and strengths.

Nusrat Jahan Ria et al. applied various classifiers in Bengali text to classify them. They used various machine learning supervised algorithms to classify their Bengali dataset [1]. Alberto Holts et al. proposed five representations which were based on text documents. Frequency representation, Binary representation, tf.rf representation, tf.idf representation, tf representation were the proposed five representations. In the dataset they had a predefined set of categories and each pair of documents had a boolean value [2]. Mita K. Dalal et al. worked on automatic text classification. They classified unstructured text which wasn't in a specific format. They trained a

set of text documents and after preprocessing they applied various machine learning models to classify their dataset. Mainly the preprocessing was for removing stemming, removing HTML tags, stop-words etc. [3]. Sheikh Abujar et al. worked on text summarization. They described the importance of data collection to create a dataset and also the importance of preprocessing techniques. And also, they talked about the barrier of collecting Bengali dataset due to the structure of Bengali text. After applying preprocessing to their dataset their data was tokenized and sequenced the input & N-gram. When the pad sequence was generated then the model was created. They trained the model using RNN (Recurrent neural Network) and predicted the output of their model [4].

Ashis Kumar Mandal et al. also worked on the same topic. They categorized the Bengali web document with machine learning supervised methods. They used machine learning supervised methods like Naive Bayes, K- Nearest Neighbor, Decision Tree and Support Vector Machines in their dataset [5]. Fabrizio Sebastiani et al. worked on automated text categorization. They described the Boolean Information Retrieval Systems. Where in this system each of the documents has one or more than one keywords. These keywords describe its contents. A finite set of these key phrases and keywords were included in the keywords. In their work they used the DNF formula which helps to get good effective results. Though this formula also has some limitations in machine learning. Information Retrieval Techniques also applied in this research work. Document indexing also has been used in this work [6]. Another work done by Fang Miao et al. they classified Chinese news text using machine learning algorithms. They used machine learning supervised algorithms like K-nearest Neighbor (KNN), Naive Bayes (NB), and Support Vector Machine (SVM) as their classification algorithm while doing the research. They pretreatment their train data set as well as test data set before applying classifiers. To do so at first, they removed the stop words and various

punctuation marks for rough dimension reduction and also did word segmentation. After that they did text representation to make the Chinese language into binary so that the machine could easily recognize the word. After data pretreatment they applied their desired machine learning algorithms to classify Chinese text [7]. Bengali text summarization using the word2vector approach was introduced by Sheikh Abujar et al. In their work they summarized the Bengali text using word2vector approach where word2vector takes input as text corpus and returns the outcome as vector. To produce word embedding Word2vec is used. They built their data set by collecting data from different web portals, news portals and social media pages. They used the Skip Gram model to identify those words which were based on other words in a similar sentence. To remove the hidden layer they used a CBOW model. To make all the words in the same position they applied this model to their data set. They used T-distributed Stochastic Neighbor Embedding to visualize the words. Vocabulary was the limitation of their work [8].

Abu kaiser Mohammad Masum et al. did text summarization with sequence to sequence RNNs. Their method created an automatic text summarizer. To do their work they collected the amazon fine food reviews dataset [9]. By the help of neural network technique Sharun Akter Khushbu et al. worked on Bengali News Headline Multi Classification. In their work they used 8000 data which were the headlines of different newspapers. They applied neural network classification which gave satisfying accuracy on the dataset. Classifiers like Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) were also applied in their dataset [10]. To classify hate speech Bjorn Gamback et al. used neural network technique. In their dataset they had four types of categories. After applying Convolution Neural network technique, they also got satisfactory accuracy over their dataset [11]. Using deep learning models Jingjing Cai et al. also worked on text classification. They used the Convolutional Neural Network (CNN)

model and RNN model to classify the text of their data set [12]. M. Ikonomakis et al. classify text using machine learning. In their work they define a single category to each document which they defined as a hard categorization [13].

2.3 Comparative Analysis and Summery

For our work we reviewed some previous work related with us. We are working with Bangla NLP, that's why we related some paper which is related with Bangla test data. Basically, we are going to check which machine learning algorithms perform very well with Bangla text data. In this section we are going to compare one previous with other work. In Table 2.1 shows the comparison between previous Bangla NLP work.

Table 2.1: Comparison Between Bangla NLP Previous Work

No	Author	Year	Algorithms	Accuracy
1.	Nusrat Jahan Ria et al.	2020	KNN NB XGB DT RF SVC	63.78% 76.76% 75.68% 65.95% 75.14% 69.19%
2.	Alberto Holts et al.	2010	SVM NB	- -
3.	Mita K. Dalal et al.	2011	NB DT Neural Network SVM	- - - -
4.	Sheikh Abujar et al.	2020	General LSTM Using LSTM	93% 97%
5.	Ashis Kumar Mandal et al.	2014	NB KNN SVM	-
6.	Fabrizio Sebastiani et al.	2002	NB DT LR KNN	81.5% 88.4% - -

			SVM	92%
7.	Fang Miao et al.	2018	KNN NB SVM	92% 92.1% 95.7%
8.	Sheikh Abujar et al.	2019	CBOW Skip-Gram	- -
9.	Abu kaiser Mohammad Masum et al.	2019	NMT RNN	- -
10.	Sharun Akter Khushbu et al.	2020	Neural Network SVM NB LR RF	90% 43% 40% 39% 38%
11.	Bjorn Gamback et al.	2017	CNN	86.68%
12.	Jingjing Cai et al.	2018	TextCNN TextRNN	85% 82%
13.	M. Ikonomakis et al.	2005	SVM DT RF NB	- - - -

We reviewed many previous works which is related with text classification problem. Because that will be very helpful for us to compare our proposed model. From the above table we saw that most of the work is related with Bangla NLP and Bangla Text classification.

2.4 Scope of the Problem

The online food delivery system is very new concept in our country. Soo many sentiments analysis work has been done before in different languages but I found no one working on Bangla customers reviews data and online food delivery companies in our country facing huge trouble to manage those data manually. So here is a great research scope available in Bangladesh.

2.5: Challenges:

Collecting Bangla reviews manually from online food delivery system is not so easy because there are no Bangla customers reviews dataset available in our country. Dataset preprocessing applying all possible filtration process is another major challenge. finally overcome all of those challenges successfully.

CHAPTER 3

RESEARCH METHODOLOGY

This segment of the documentation is based on information gathering and information pre-processing procedures. Here I have given a proper explanation of our dataset creation and data filtration processes. This section also shows how we preprocessed all of our data using a different types of machine learning algorithms.

3.1 Research Subject and Instrumentation:

The Bengali language has two different types of forms one is saint form and another one is usual form which one is people regularly used. The sentiment analysis work on Bangla language in very trendy topic. Machine learning classifier were used to identify negative and positive sentiment. I have used machine learning supervised algorithms like KNN, DT, NB, RF, SVC and XGB to classify the dataset.

k-Nearest Neighbor (KNN):

It's called one of the simplest machine learning algorithms based on supervised learning tactics. this algorithm stores every one of the accessible information and characterizes another information point dependent on the closeness. This algorithm mostly uses to solve classification issues [15].

Decision Tree:

This algorithm is another supervised learning algorithm. It solves classifications problems too.it works as decision support tool in machine learning process [16].

Naive Bayes:

This is actually a collection of classification algorithms based on bayes theorem. This a family of algorithms where every single of them shares a common principle [17].

Support Vector Machine:

This algorithm is used for both classification and regression analysis and it's a supervised machine learning algorithm. Most of the time this is used for classification. The aim of using SVM is to generate the best decision boundary which can separate n-dimensional space into classes so in the future we can easily keep the latest data point in the precise category [18]

XGB:

This algorithm is a tree-based ensemble machine learning algorithm that is famous for its scalability. It drives fast learning methods through parallel and distributed computing and offers efficient memory usage [19]

Confusion matrix:

Basically, it is a performance measurement method for the classification model in machine learning. Using this method can provide a superior thought of what your classification model is getting right and what kind of mistakes it is making.

There are four terms in the confusion matrix which are given below:

1. True positive
2. False positive

3. True negative

4. False negative

Table 3.1: Sample Confusion Matrix

n	Predicted No	Predicted Yes
Actual No	TN	FP
Actual Yes	FN	TP

Precision:

Precision refers to how close two or more measurements to each other are. Precision is independent. Precision is a positive predictive value. It is a fraction among the instances.

$$\boldsymbol{precision} = \frac{\boldsymbol{True\ positive(TP)}}{\boldsymbol{True\ Positive(TP) + False\ Positive(FP)}}$$

Recall:

Recall is the sensitivity. It is also a fraction of retrieved relevant instances. It can be viewed as a probability.

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positive(TP)}}{\boldsymbol{True\ Positive(TP) + False\ Negative(FN)}}$$

Accuracy:

Accuracy is the known or standard value. It refers to the nearness of measured value.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{True\ Positive + True\ Negative}}{\boldsymbol{(True\ Positive + False\ Positive + True\ Negative + False\ Negative)}}$$

3.2 Data Collection Procedure:

The first and most basic rule for text classification big quantity of data is required. A large amount of data for this research work were collected from different types of online food delivery platforms. I have collected all of our data from those platforms manually. For this research work, I have collected data from foodpanda and HungryNaki online food delivery platforms.

Table 3.2: Bangla Dataset of food review

Sentence	Class type
ভালো ছিল। নরমালি অফারের খাবারগুলোর গুণগতমান খারাপ করে দেয়, এখানে ঠিকই ছিল। এটাই সবচেয়ে ভালো লেগেছে।	5
স্বাদ অসাধারণ,তবে মাংসের পরিমাণ খুব কম ছিল প্রাইস হিসেবে	4
মোরগ পোলাও এর পোলাও টা তেমন ভালো ছিলো না। আগে কখনো এমন পাই নাই। স্বাদহীন	3
চেয়েছি একটা আর তারা দিয়েছে আরেকটা। এমনকি তারা কল দিয়েও কনফার্ম করে নাই।	2
খিচুড়ি অর্ডার করেছিলাম, পোলা তো দুরের কথা, মনে হলো সাদা ভাতের মধ্যে হলুদ দিয়ে খিচুড়ি বানায়া দিসে। তাতে আবার পাথর দুইবার দাঁত ভাঙ্গার জোগাড়। এখন পর্যন্ত আমার খাওয়া বাজে খিচুড়ি। অথচ দাম কিন্তু সেই লালবাগের কিন্তু মানে আকাশ পাতাল পার্থক্য। ইচ্ছা করছিলো দোকানে গিয়ে দুই গালে দুইটা সেটিয়ে আসি।	1

3.3 Statistical Analysis

After collecting data from various source, we are able to collect 400 data and there were 5 classes. Distribution of data in each class are shown in Figure 3.1.

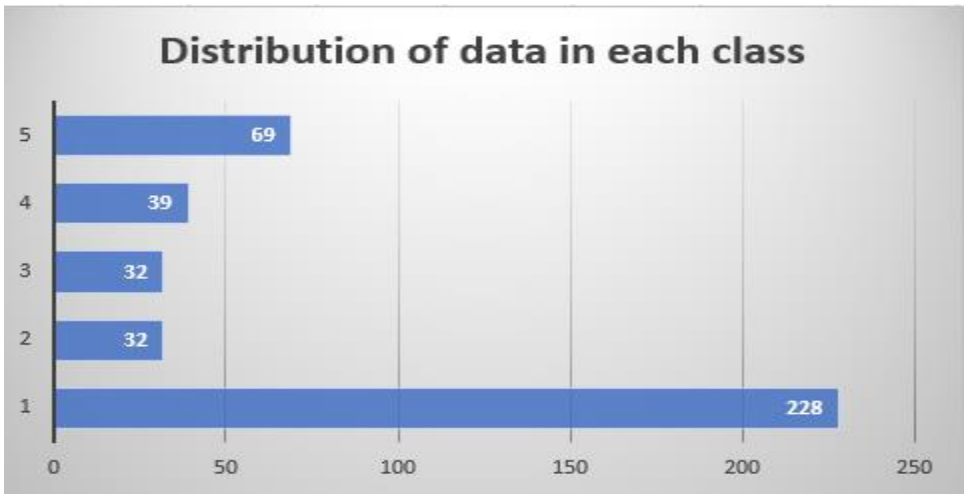


Figure 3.1: distribution of data in each class

3.4 Proposed Methodology

We are going to discuss about our research methodology in this following section. In our work, we use six supervised machine learning classifiers MNB, DT, RF, KNN, SVM and XGB to classify food reviews from Bengali sentences. To apply this classification algorithm, we make our own dataset. Though it is hard to find the appropriate resource for Bangla Language but had tried our level best to make our work accurate. For this we divided our work into some steps. Figure 3.2 represents the steps of our methodology. We discussed data collection procedure in 3.2 section. Rest of the methodology steps are described below.

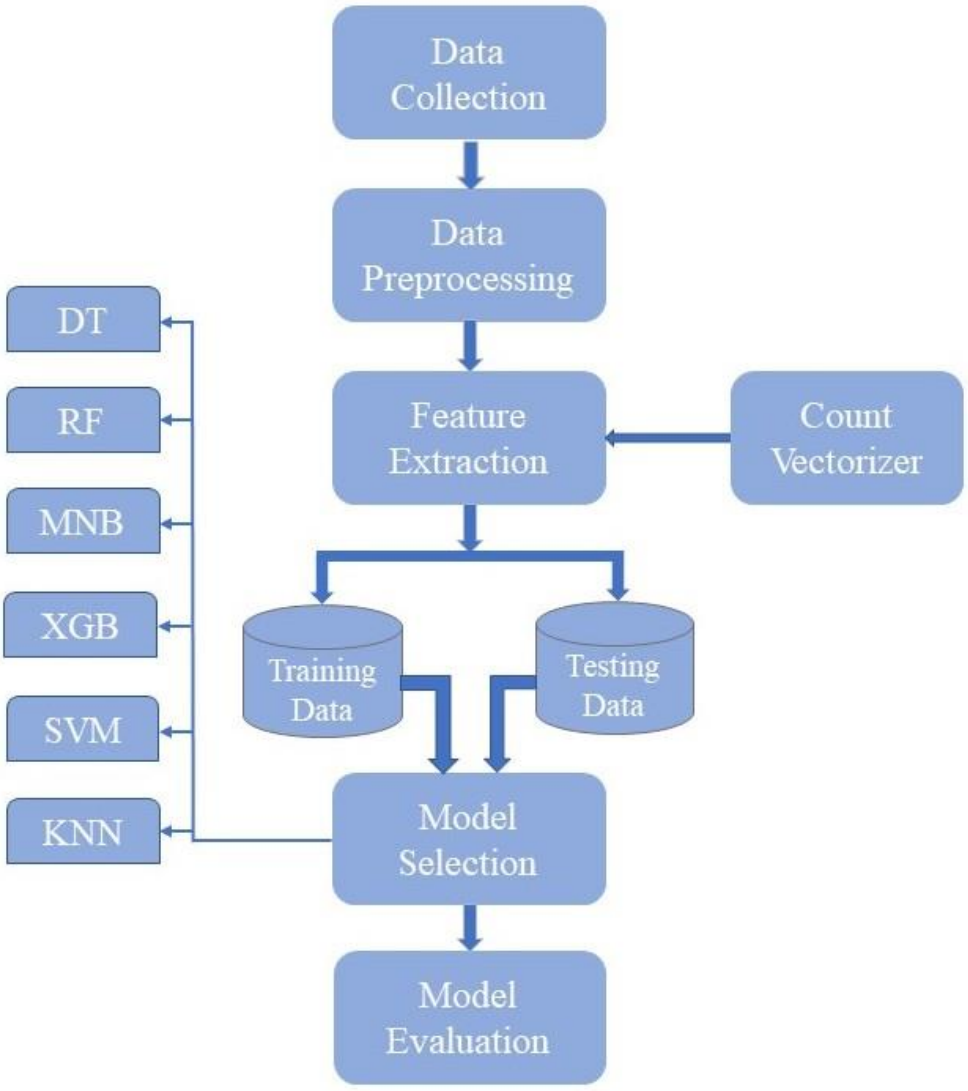


Figure 3.2: Proposed Methodology

3.4.1 Data Preprocessing

We cannot use raw text data to feed our classifier model. Because sometimes raw text data have some characters or symbols which is not essential and suitable for our classifier model. This unwanted characters and symbols sometimes reduce our classifier model accuracy. So, before feeding our model we need to apply some preprocessing techniques on the raw text data. In our raw text data we found some special characters and symbols *, #, !, @ etc. We remove those special characters and symbols from our text. Our text data contains some numerical values such

as English digits and Bangla digits. We also found some punctuation marks such as full stop, comma, question marks, quotation marks etc. We use python regular expression library to remove this unwanted data. Table 3.3 shows the characters details which we have remove from our data in preprocessing phase. In our raw text we found some Bangla short form "ইঞ্জিঃ", "ড.", "রেজিঃ" etc. We elaborate the short form on Bangla text as "ইঞ্জিঃ" => "ইঞ্জিনিয়ার", "ড." => "ডক্টর", "রেজিঃ" => "রেজিস্ট্রেশন" etc. For this we make a python dictionary with short form and elaborate form, after that we split our text data and compare the data with dictionary. If short form found we replace those data with elaborate form. Table 3.4 provides the details of Bangla probable short form and their elaborate form. In NLP work there are some words in every language which are commonly used and this word are unimportant for machine learning model, this set of data are called stop words. So, in Bengali language we have some stop words such as "ও", "এবং", "অতএব", "অথচ", "অথবা" etc. [13] In data analysis or when we apply classifier model it creates problems. We need to use Bangla stop words corpus to remove stop words from our dataset. Beside this in Bangla stop word corpus we found some words that are important for our work we filter out those word. We remove stop words from our dataset using our own modified Bangla stop words corpus. Table 3.5 shows the raw text data and preprocessed text data and Figure 3.2 shows the text data preprocessing steps.

Table 3.4: Characters Details Considered Removing in Preprocessing

Characters Category	Characters
Special Symbols	@, #, $, %, ^, &, *, (,), /, \, {, }, ……
English Digits	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Bangla Digits	০, ১, ২, ৩, ৪, ৫, ৬, ৭, ৮, ৯
Punctuation Marks	?, !, “ ”, ., :, ;, ……..
English Alphabets	A to Z; a to z
Bangla Full Stop (দাঁড়ি)	।

Table 3.5: Bangla Short Form and Elaborate Form

Short Form	Elaborate Form
বি.দ্র.	বিশেষ দ্রষ্টব্য
ড.	ডক্টর
ডা.	ডাক্তার
ইঞ্জিঃ	ইঞ্জিনিয়ার
রেজিঃ	রেজিস্ট্রেশন
মি.	মিস্টার
মু.	মুহাম্মদ
মো.	মোহাম্মদ

Overall data preprocessing steps are shown in figure 3.3.

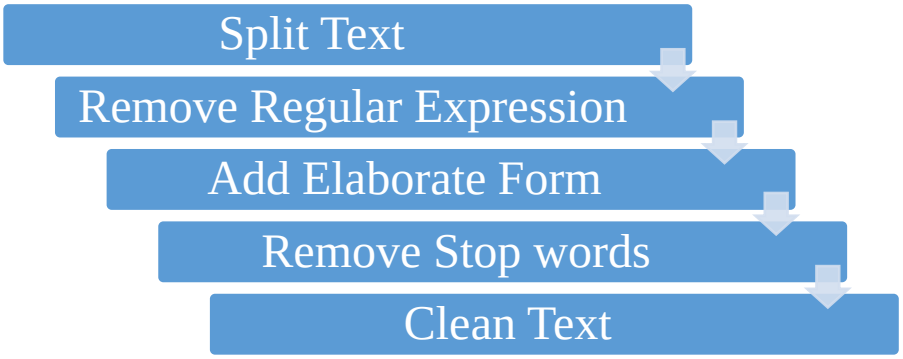


Figure 3.2: Data Preprocessing Steps

3.4.2 Feature extraction

After preprocessed our data we have clean text, but we can't feed our machine learning model with this text. We need extract the features from clean text. Feature extraction means we need to convert the text data into numerical value. But we need to extract the features in proper way, because extracting proper features from the text have an impact on the machine learning model performance. For this we need to apply some techniques which convert our text data into vector, that means in numerical value. This is called one hot encoding. Here we use count vectorizer and tf-idf (term frequency-inverse document frequency) vectorizer. Count vectorizer is most useful feature extraction method in NLP. Basically, Count vectorizer make a vector from the text data based on the word frequency (count) of each word which is occurs in the sentences. This is very helpful method for sentiment analysis or any NLP related work. Count vectorizer creates a vector or matrix where unique words are represented as matrix columns and each row text data from dataset represented as matrix row. After that it count the word frequency and put that value on the matrix. TF-IDF is advance and common method for features extraction from processed Text data. Sometimes features extraction using TF-IDF vectorizer method increase the proposed model accuracy.

3.4.3 Model Selection

In machine learning techniques there are two types of leaning exist. Supervised Learning and Unsupervised Learning. In our work we have input and output data to train a model so we need to use Supervised Learning algorithms. In our work we use some supervised classifier algorithms, we apply six different algorithms DT, RF, XGB, MNB, SVM and KNN on our dataset. After vectorized our data we divided our vector data into two parts training and testing data. We split our data into 80% and 20%. For training purpose, we keep 80% and rest of 20% data for testing

purpose. We apply six different classifiers on our training data and evaluate the model based on testing data. Multinomial Naive Bayes classifier perform very well on our dataset.

3.5 Implementation Requirements

Hardware and Hardware and Software:

- Intel Core i5 8th gen integrated with 8GB ram
- 1 TB Hard Disk
- Google Colab with 12GB GPU and 350GB ram
- High Speed Internet Connection

Advance Libraries and Tools:

- Windows 10
- Python 3.8
- Pandas
- NumPy
- NLTK
- Matplotlib
- Scikit-Learn

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

In this section we are going to describe about our model performance which we apply on our dataset. In our work we use six classification algorithms which we already discussed in the Methodology section. All classification algorithms perform very well but some of them are classify the text data accurately. When we try to apply Machine Learning algorithms in raw data, we found accuracy below the 10%. After that we preprocessed the raw data. In the Methodology section we mentioned the technique we use to preprocess our data. Then we apply six machine learning algorithms in our clean data. Our MNB, RF, DT, SVM, KNN, XGB models came with the accuracy of 58%, 62%, 50%, 60%, 63%, 60% respectively. These classifiers perform very well with text data. Table 4.1 shows the all-models accuracy. Only based on the accuracy score we can't consider our model as a perfect model for our dataset.

Table 4.1: Model Accuracy

Model Name	Accuracy
Multinomial Naïve Bayes	58.75%
Random Forest	63.74%
Decision Tree	50%
Support Vector Machine	60%
K-Nearest Neighbors	65%
Extreme Gradient Boosting	62.5%

Among the all-algorithms **K-Nearest Neighbors** gives us highest accuracy which is **65%**.

4.2 Experimental Results and Analysis

We have used supervised algorithms like Decision tree, Naive Bayes, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Extreme Gradient Boosting (XGB) to classify the food reviews of customers from Bengali text. After implementing the model, we took a sample input to check which of the algorithms can predict the correct form of the input, either its saint or its common. The outcome of this experiment is given below.

Sample input: ভালো ছিল। নরমালি অফারের খাবারগুলোর গুণগতমান খারাপ করে দেয়, এখানে ঠিকই ছিল। এটাই সবচেয়ে ভালো লেগেছে।

Table 4.1: One sample prediction result of each classifier

No	Classifier	Prediction	Result
01	MNB	5	right
02	RF	5	right
03	DT	5	right
04	SVM	1	wrong
05	KNN	5	right
06	XGB	5	right

As we can see from the table for the sample input Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, Extreme Gradient Boosting (XGB) algorithms show correct prediction and algorithm Support Vector Machine (SVM) show wrong prediction.

Then we started implementing the algorithms in our dataset to measure the accuracy, and precision, recall, f1 score and support were also calculated. We calculated the accuracy from the confusion matrix. Confusion matrix is a convenient way of performance measure of any predictive model.

Accuracy shows how often the classifier is correct. Again, Precision determines when the model predicts yes, how often it is correct. The overall result of the performances are shown in the figure 4.1.

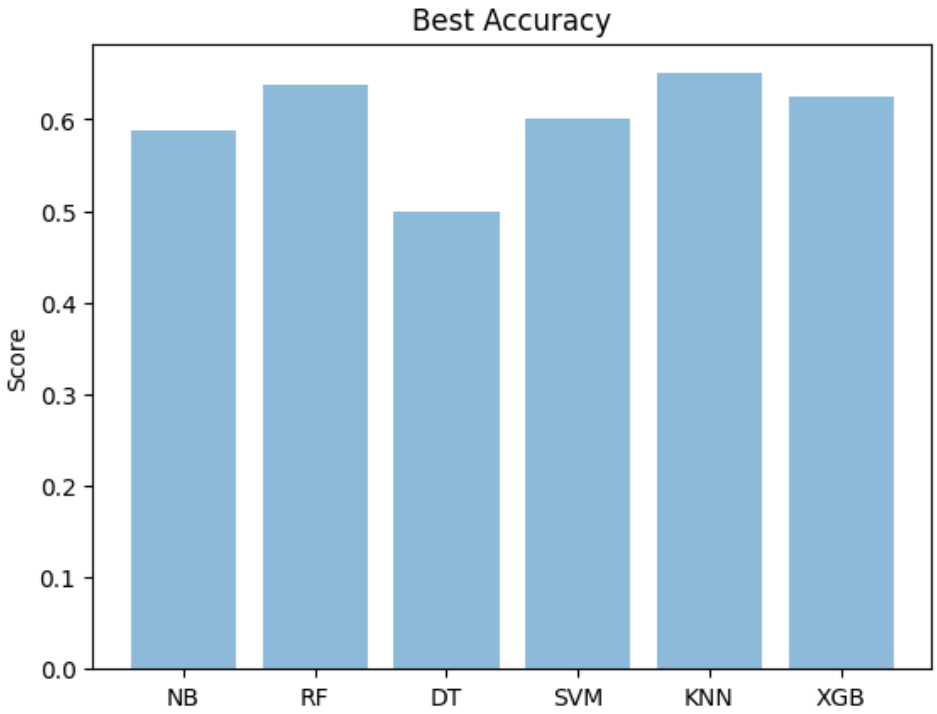


Figure 4.1: Performance of Classifiers

4.3 Discussion

In the end we are try to contribute in the Bangla research domain. Worldwide there are lots of work in the different language. We are trying to added our Bangla language in that work list. We dreamt to detect the Bangla food reviews from Bangla texts. So, we are very happy that we make our dream comes true. In this work we are able to make a machine learning model which is able to detect the class level of an customers' review. We use six different algorithms in our model. All model's accuracy is shown in the Figure 4.1.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Customer review classification creates a massive impact on Bangla language that is the reason we have created this model so by this anyone can easily have the idea of what his or her customers are thinking about his or food or service quality.

5.2 Impact on Environment

As we are expecting that it will be a convenience project so that it will create a great impact on society then it must be having a great impact on environment too. Because an environment is made with a society & society made with people. If people are having adequate knowledge about their own opinion, then it also dominance in their environment. So, for, our project will help to predict class level of a customer review which in course will help the business owners to improve his food or service quality.

5.3 Ethical Aspects

In order to conduct the research, we tried to adopt all sorts of ethical issues related with research. We have given due credits to all the authors whose papers we have cited and who helped us to gather our knowledge and also to implement the codes finally.

5.4 Sustainability Plan

Our plan is to help upcoming generation so they can find better world. Food is one the most important part of life. At the time our daily life is getting mechanical day by day. People have

hardly any time to cook at home now. Therefore, almost everyone tries order foods to ease the extra pressure of their life. But, the quality of the food or the service provided by restaurant can spoil all the fun and can create extra pressure on a person. He or she will definitely feel mental pressure in that situation. So, it will definitely help improve our world.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the study

Our work is related with Bengali NLP. Working with NLP is a challenging work for researcher. In research work dataset is main important for execute the work. In this work we are making a machine learning model for detecting the level of Bangla customer review. While building this model we faced some problem and so more things we face during this work. All the steps and work summery is given below step by step.

Step 1: Planning about this work

Step 2: Problem formulation

Step 3: Data collection from various books and newspapers

Step 4: Data Labeling

Step 5: Data cleaning

Step 6: Data Vectorization

Step 7: Train and Test Data Separation

Step 8: Model Selection

Step 10: Model evaluation and performance testing

After execute all of the steps we are finally able to make our model that can Detect class level of Bangla customers' review.

6.2 Conclusion

There are so many works about Bangla test classification, fake news detection. Some classification algorithms in supervised learning perform very well for text data. In our work we proposed to build a model on six different algorithms such as DT, RF, MNB, XGB, KNN and SVM using our Bangla Dosh dataset. In our task making dataset is very hard and tough work for us. Because there is no previous work available related with our data. We need to make our data manually. Finally, in our work we use 400 data to make a model. All classification algorithms perform very well, but KNN came out with highest performance accuracy with our dataset.

6.3 Recommendation

We have some recommendations for our work. In this section we will increase our dataset for improve our model accuracy. In our work we use some supervised machine learning classification algorithms. And in the text data transformation section we use only one vectorization technique. There are so many techniques and algorithms for large number of datasets. So, that model and techniques will predict more accurately Bangla Guruchandali Dosh sentence. Some recommendations of our work are given below.

- Big dataset for Bangla Customer Reviews
- Understand the transformation techniques and improvised the techniques for Bengali dataset.
- Try to make a better classification model

- Try to get more performance accuracy

6.4 Implication for Further Research

We have some limitations and drawback in our work. For example, we only use machine learning algorithms in our model. Beside this we only use Count Vectorized text transformation techniques. Also, our data is not sufficient. So, we extend our data count. Because, we have plan to apply deep learning algorithms like RNN, LSTM, BiLSTM etc. in our dataset. Without increasing the data volume, we cannot get the better accuracy from deep learning model. We also use Word Embedding techniques to vectorized the text data in to numeric value.

References

[1] R. Dumbleton, "Sentiment Analysis: Definition, Uses, Examples + Pros /Cons", Thematic, 2022. [Online]. Available: https://getthematic.com/insights/sentiment-analysis. [Accessed: 20- Jan- 2022].

[2] Nusrat Jahan Ria, Sharun Akter Khushbu, Mohammad Abu Yousuf, Abu Kaisar Mohammad Masum, Sheikh Abujar, "Syed Akhter Hossain, Toward an Enhanced Bengali Text Classification Using Saint and Common Form", 11th ICCCNT, 2020

[3] Alberto Holts, Claudio Riquelme, Rodrigo Alfaro, "Automated Text Binary Classification using Machine Learning approach", XXIX International Conference of the Chilean Computer Science Society, 2010.

[4] Mita K. Dalal, Mukesh A. Zaveri, "Automatic Text Classification: A Technical Review", International Journal of Computer Applications, 2011.

[5] Sheikh Abujar, Abu Kaisar Mohammad Masum, Md. Sanzidul Islam, Fahad Faisal and Syed Akhter Hossain "A Bengali Text Generation Approach in Context of Abstractive Text Summarization Using RNN", 7th ICICSE, 2020.

[6] Ashis Kumar Mandal, Rikta Sen, "SUPERVISED LEARNING METHODS FOR BANGLA WEB DOCUMENT CATEGORIZATION", International Journal of Artificial Intelligence & Applications (IJAIA), 2014.

[7] Fabrizio Sebastiani, Consiglio Nazionale delle Ricerche, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, 2002.

[8] Fang Miao, Pu Zhang, Libiao Jin, Hongda Wu, "Chinese News Text Classification Based on Machine learning algorithm", 10th International Conference on Intelligent Human-Machine Systems and Cybernetics, 2018

[9] Sheikh Abujar, Abu Kaisar Mohammad Masum, Ohidujjaman, Syed Akhter Hossain, "An Approach for Bengali Text Summarization Using Word2Vector", 10th ICCCNT (International Conference on Computing, Communication and Networking Technologies), 2019.

[10] Abu kaiser Mohammad Masum, Sheikh Abujar, Md Ashraful Islam Talukder, AKM Shahriar Azad Rabby, Syed Akhter Hossain, "Abstractive method of text summarization with sequence-to-sequence RNNs", 10th ICCCNT, 2019.

[11] Sharun Akter Khushbu, Mohammad Abu Yousuf, Abu Kaisar Mohammad Masum, Sheikh Abujar, Syed Akhter Hossain, "Neural Network Based Bengali News Headline Multi Classification System: Selection of Features describes Comparative Performance", 11th ICCCNT, 2020

[12] Bjorn Gamback, Utpal Kumar Sikdar, "Using Convolutional Neural Networks to Classify HateSpeech", 2017.

[13] Jingjing Cai, Jianping Li, Wei Li, Ji Wang, “Deep Learning Model Used in Text Classification”, ICCWAMTIP, 2018.

[14] M. Ikonomakis, S. Kotsiantis, V. Tampakas, “Text Classification Using Machine Learning Techniques”, 8, Volume-4, PP.966-974, 2005.

[15] "K-Nearest Neighbours - GeeksforGeeks", *GeeksforGeeks*, 2022. [Online]. Available: https://www.geeksforgeeks.org/k-nearest-neighbours/. [Accessed: 20- Jan- 2022].

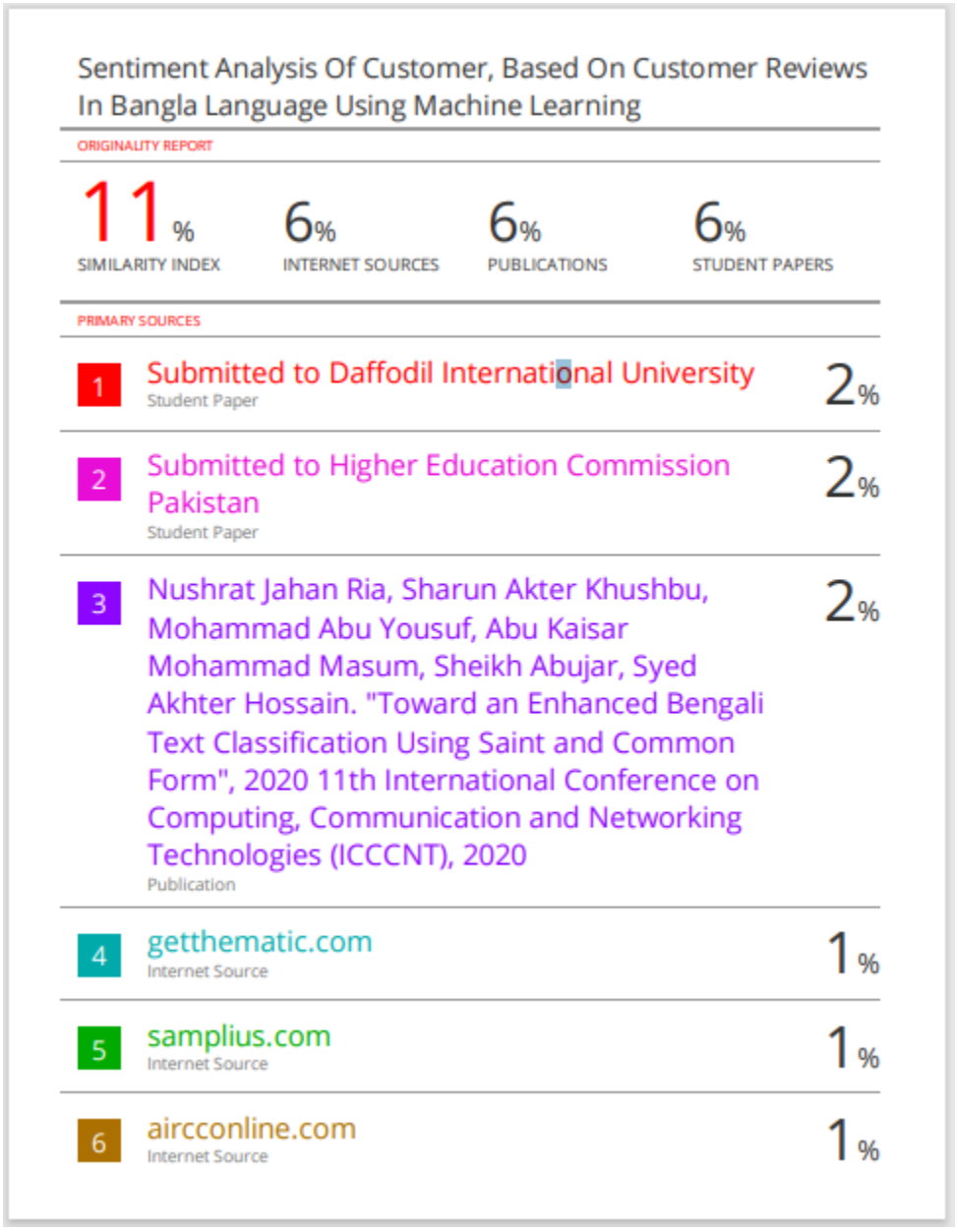
[16] N. Chauhan, "Decision Tree Algorithm, Explained - KDnuggets", *KDnuggets*, 2022. [Online]. Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html. [Accessed: 20- Jan- 2022].

[17] "Naive Bayes Classifiers - GeeksforGeeks", *GeeksforGeeks*, 2022. [Online]. Available: https://www.geeksforgeeks.org/naive-bayes-classifiers. [Accessed: 20- Jan- 2022].

[18] "XGBoost Algorithm - XGBoost In Machine Learning", *Analytics Vidhya*, 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/. [Accessed: 20- Jan- 2022].

[19] "Support Vector Machine (SVM) Algorithm - Javatpoint", *www.javatpoint.com*, 2022. [Online]. Available: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm. [Accessed: 20- Jan- 2022].

Plagiarism Report



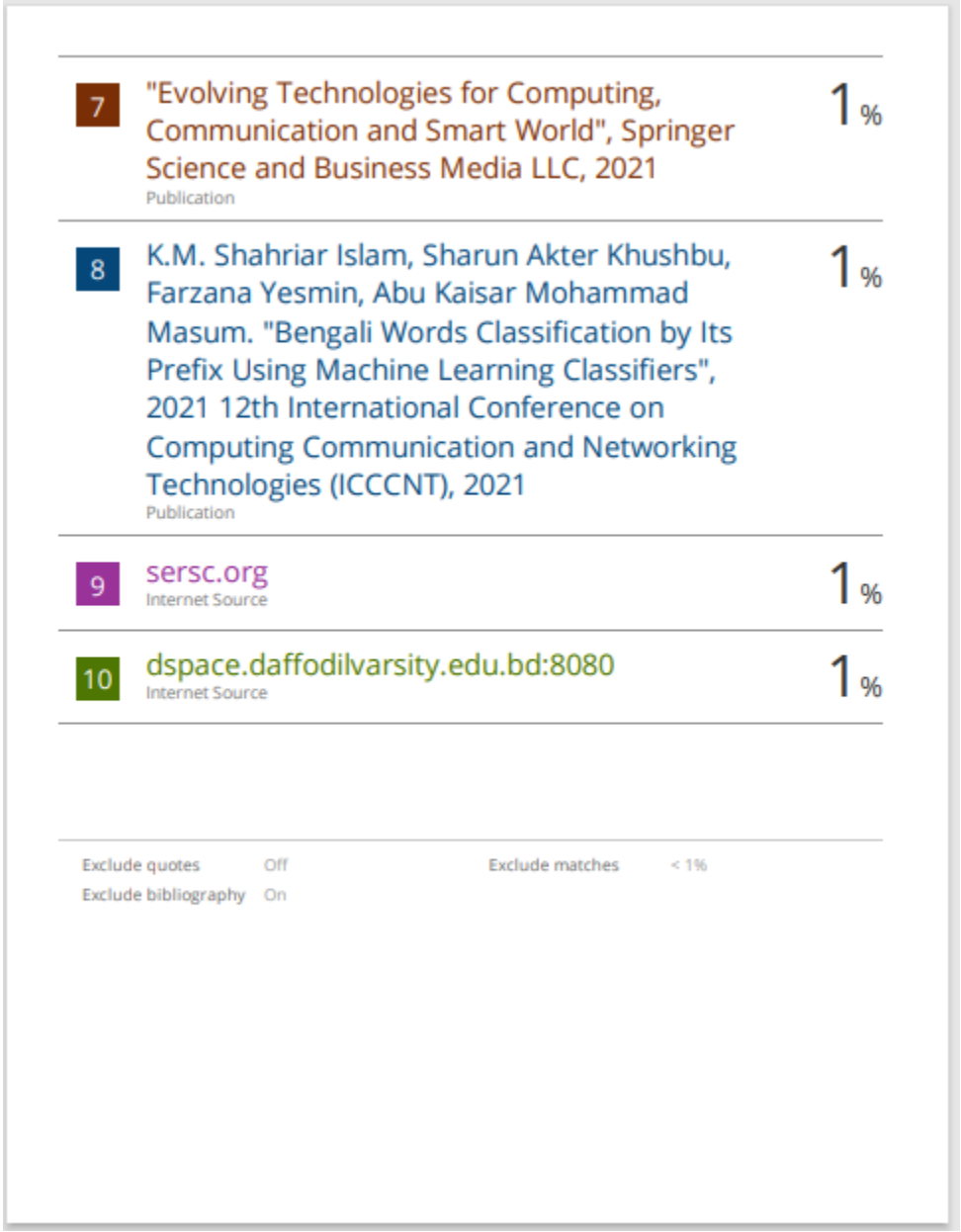
Sentiment Analysis Of Customer, Based On Customer Reviews In Bangla Language Using Machine Learning

ORIGINALITY REPORT

11%	6%	6%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	Submitted to Higher Education Commission Pakistan Student Paper	2%
3	Nushrat Jahan Ria, Sharun Akter Khushbu, Mohammad Abu Yousuf, Abu Kaisar Mohammad Masum, Sheikh Abujar, Syed Akhter Hossain. "Toward an Enhanced Bengali Text Classification Using Saint and Common Form", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020 Publication	2%
4	getthematic.com Internet Source	1%
5	samplius.com Internet Source	1%
6	aircconline.com Internet Source	1%



7	"Evolving Technologies for Computing, Communication and Smart World", Springer Science and Business Media LLC, 2021 Publication	1 %
8	K.M. Shahriar Islam, Sharun Akter Khushbu, Farzana Yesmin, Abu Kaisar Mohammad Masum. "Bengali Words Classification by Its Prefix Using Machine Learning Classifiers", 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021 Publication	1 %
9	sersc.org Internet Source	1 %
10	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1 %

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		