Comparative Analysis of Saint and Common Forms of Bengali Text Classification Using Machine Learning

Md. Tanvir Ahmed Akash
Id: 221-15-5424
Department of Computer Science and Engineering
Daffodil International University, Dhaka, Bangladesh
akash15-5424@diu.edu.bd

Saiham Zaman Mridul
Id: 221-15-5849
Department of Computer Science and Engineering
Daffodil International University, Dhaka, Bangladesh
mridul15-5849@diu.edu.bd

Md. Foysal Sheikh
ID: 221-15-4865
Department of Computer Science and Engineering
Daffodil International University, Dhaka, Bangladesh
foysal15-4865@diu.edu.bd

Abstract:

Natural Language Processing (NLP) is an aspect of AI that enhances human language processing tasks for a machine or system. Text categorization, a prominent research subject, has aided natural language processing tasks, and this work focuses on language processing problems in Bangla, our native language. There has been relatively little study on Bengali text classification based on its two linguistic forms: saint (Sadhu bhasha) (elegant or delicate) and common (Cholito bhasha) (current or colloquial), and the dataset is small. As a result, this study focuses on categorizing Bengali texts based on these two types, using about 3005 sentences or data collected from Bengali literature, blogs, and articles. To classify Bengali texts, different supervised machine-learning algorithms have been applied, such as SVM, RF, and XGB. Before applying these algorithms, dataset preprocessing techniques were applied such as primary cleaning, regular expression removal, stopword removal, digit removal, null value removal, and tokenization. Using the Countvectorizer SVM and XGB has achieved a maximum output accuracy of 92.39% and RF, achieved 91.94% accuracy and by using TF-IDF Vectorizer SVM has achieved a maximum output accuracy of 93.06% and RF, XGB achieved 92.17% accuracy. This comparative study opens the way for future research in Bangla language processing, specifically text categorization.

Keywords: NLP, Machine learning, Bengali Text Classification, Shadu, Cholito, Saint, Common, CV, TF-IDF

Introduction:

Bangladesh's official and national language is Bengali, often known as Bangla. “Bangla" is written বাংলা in the Bengali script. This writing system (brahmic script) is the 6th most widely used one in the world [1,2,3]. Bangla belongs to the Indo-European language family and dates back to 3500 B.C [3]. More than 210 million people speak Bengali as a first or second language, with approximately 100 million in Bangladesh, 85 million in India, primarily in the states of West Bengal, Assam, and Tripura, and sizable immigrant communities in the United Kingdom, the United States, and the Middle East. Bengali has two standard speech patterns: Common (Cholito bhasa) (current or colloquial) and Saint (Sadhu bhasa) (elegant or gentle) . Common (Cholito bhasa) form is generally used for both writing and speaking and Saint (Sadhu bhasa), the literary style, which contains many words derived from Sanskrit and commonly used for writing [4].

But, recently we noticed that people are showing a tendency to mix Saint and Common forms of Bengali language, especially their social media posts and writing. Here the main problem is they are unable to do so and can not differentiate between two forms correctly. Some of the best Ai like ChatGpt, perplexity ai, gemini etc. are making big mistakes to write the Saint (Sadhu bhasa) form of Bengali language. Some impact occurs after introducing Bengali with English [5]. Those problems can not be ignored and The Dhaka University of Bangladesh published a paper on the language situation in Bangladesh [6].

To overcome such linguistic challenges in the Bengali language, we must focus on natural language processing (NLP) activities in Bangla, our own tongue. Natural language processing (NLP) is the intelligent and practical process by which computers evaluate, understand, and extract meaning from human discourse [7]. Natural Language Processing enables developers to organize and structure data for tasks like text categorization or classification, automatic text summarization, translation, named entity identification, connection extraction, sentiment analysis, voice recognition, and topic segmentation. NLP researchers are growing more interested in text classification, which is used with machine learning to improve language processing jobs. Text classification, often known as categorization, is an important field of study in natural language processing research. Natural Language Processing (NLP) is used to assess various texts and classify them into unique groups or classes.

Our main purpose of this research is to classify Saint (Sadhu bhasa) and Common (Cholito bhasa) from Bengali text using Machine Learning (ML) algorithms. Here our used machine learning classifiers will classify the forms of Bengali language. To do so Natural Language Processing (NLP) offers a variety of text classification approaches, including multi-level and binary classification. In this work, we will focus on binary-level classification, which is the categorization of text into two levels or classes.

The primary contributions of this proposed effort are listed below:

- 3005 self-made data (sentences) are collected from Bengali historical books, novels, and literary works.
- To ensure data quality and make the data machine efficient, a number of preprocessing techniques were applied to the data.
- We are using CountVectorizer and TF-IDF Vectorizer to transform text into a meaningful representation of numbers.
- We have selected three supervised machine learning algorithms based on our dataset.
- Selected machine learning algorithms were applied to preprocessed datasets and their performance was evaluated in terms of Accuracy, Precision, Recall, F1-score, TPR, TNR, and Error rate matrices.
- To enhance the performance of the applied machine learning models, their parameters were adjusted and better accuracy was achieved.
- Finally, the result has been analyzed with a comparative analysis of performance metrics, confusion matrices as well as prediction results of input data by each classifier.

The remaining portion of this paper is organized as follows: Section 2 describes a literature study of relevant studies, Section 3 illustrates the research technique, Section 4 depicts the experiment and result analysis, and Section 5 concludes and discusses future work.

Literature review:

The classification of Saint(Sadhu bhasa) and Common(Cholito bhasa) forms of Bengali has gained prominence in NLP research in Bangladesh, due to its relevance for applications like automated translation and sentiment analysis. The Saint form, rooted in classical texts, contrasts with the Common form of everyday speech. Researchers have employed various classifiers, such as SVMs, decision trees and neural networks, using datasets from literary works, news, and social media. Advances in deep learning and transfer learning, including word embeddings and BERT, have enhanced classification accuracy. This study integrates and extends previous methodologies to improve performance, addressing cultural and practical implications for more sophisticated Bengali NLP applications.

Bitto et al.[8] examined machine learning techniques for categorizing Bengali social media text as "good" or "bad". They test Logistic Regression (LR), Decision Tree Classifiers (DTC), Random Forests (RF), Multinomial Naive Bayes (MNB), and K-Nearest Neighbors (KNN) on a dataset of 1499 Bengali text documents, all achieved great accuracy, with MNB coming out on top with 89.31%. Hasan et al.[9] investigated emotion recognition in Bengali speech using recurrent neural networks (RNNs). They categorized emotions based on Bengali music and movie speech as well as the researchers obtained an accuracy rate of 47.66% to 51.33% for six emotions: joy, sorrow, anger, surprise, fear, and disgust. Rahman et al.[10] proposed a dynamic approach using a Word2Vec model for sentiment classification in Bengali literature. Using a dataset comprising 11,000 Bengali articles, newspapers, and continuous bags of words (CBOW), they were able to reach a 75% accuracy rate for sentiment classification of happy, furious, and excited categories. Khushbu et al.[11] used Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR), Neural Networks (NN), and Naive Bayes (NB) to classify tenses in Bengali news headlines. They attained a 90% accuracy rate with a dataset of 8000 Bengali news headlines in 11 different tenses. Khushbu et al.[12] studied the use of Long Short-Term Memory (LSTM) networks to categorize Bengali news articles into six categories: entertainment, national, sports, city, state, and foreign. They attained 84% accuracy with a dataset of 13,445 Bengali news items.Ria et al.[13] studied the application of four machine learning techniques, namely Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest

Neighbors (KNN), to categorize Bangla web pages. They achieved 77% accuracy with a dataset of over 1200 mixed sentences gathered from several Bangla websites . Using a multilingual BERT model that has already been trained, Islam et al.[14] investigate sentiment analysis in Bengali news articles. Sentiment analysis is the process of determining if a text contains positive or negative sentiment. Using a dataset of 17,852 Bengali news entries, they were able to categorize Bengali news articles into three sentiment categories (positive, negative, and neutral) with an accuracy of 71% . [15] Haque, Rezaul, et al. A large dataset of 84,072 Bengali social media comments was collected and classified into four categories: political, religious, sexual, and acceptable. The collection is much larger than previous freely available Bengali hate speech datasets. Machine learning models were tested on the dataset and achieved an accuracy of 85.8%. The dataset and outcomes indicate the viability of detecting hate speech in the Bengali language on social media at scale using supervised learning techniques. Wadud et al.[16] used Long Short-Term Memory (LSTM) networks to categorize Bengali social media text by race, religion, ethnicity, sexual orientation, gender, and physical impairment. They achieved 93.38% accuracy with a dataset of 20,000 Bengali social media posts. Bijoy et al.[17] used machine learning methods such as K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Multinomial Naive Bayes (MNB), and Support Vector Machines (SVM) to detect objectionable language in Bengali discussions and blogs .They achieved 74.03% accuracy using a dataset of 740 Bengali text documents that ranked first in search results . Mandal et al.[18] explored many machine learning methods, including NB, SVM, DT, KNN, Random Forest (RF), and Logistic Regression (LR), for categorizing Bangla news items from diverse sources into business, sports, health, technology, and education categories. They achieved 89.14% accuracy with a collection of 1000 Bangla news documents. Hasan et al.[19] examined the sentiment categorization of Bangla documents using several machine learning methods such as Random Forest (RF), Decision Tree (DT), and LSTM. They attained the greatest accuracy of 89.42% for positive sentiment and 88.20% for negative sentiment on a dataset of 1824 Bangla text documents. Bitto et al.[20] study sentiment analysis in analyses of Bangladeshi food delivery businesses using machine learning (ML) and deep learning (DL) approaches . They examined a variety of techniques, including Logistic Regression (LR), Random Forest (RF), Multinomial Naive Bayes (MNB), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Recurrent Neural Networks (RNN) . Their greatest findings came from Machine

Learning techniques, with Random Forest scoring 89.64% for positive sentiment and 91.07% for negative sentiment. Roy et al.[21] used several machine learning techniques to identify Bengali news items, including Support Vector Machines (SVM), Random Forest (RF), Multinomial Naive Bayes (MNB), Decision Tree (DT), K-Nearest Neighbors (KNN), and Logistic Regression (LR) . They attained the greatest accuracy of 0.9261 and 0.9521 using a Support Vector Machine (SVM) ,Xo-Bert respectively on a dataset of 14,451 Bengali news pieces. Alam et al.[22] evaluated the effectiveness of using BERT and XLM-RoBERTa big language models, as well as a beam search decoder, for text summarization of web pages, social media posts, online news portals, emails, online shops, user reviews, and customer care question and answer sessions. They found 93.8% on a dataset of Bengali and English documents .Sarkar et al.[20] investigated a framework for identifying Bangla news articles and comments with Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks. They obtained an accuracy of 83.77% using a dataset of 13,803 Bengali news items and comments. Sarkar et al.[23] examined a method for deep learning that uses Convolutional Neural Networks (CNNs) and a softmax classifier to classify Bengali social media comments into political, religious, social, and mixed categories . Using a pre-trained XLM-RoBERTa big language model with a dataset of 84,072 Bengali social media comments, they attained an accuracy of 95.12% . Naughton et al.[24] used Support Vector Machines (SVMs) to classify text from diverse genres into event categories such as "die" and "attack". They test two SVM models on a set of English text data from news articles and blogs. Their SVM models performed well on the event categorization challenge. Sen et al[25] looked at the classification of queries from a real-world contracts dataset into predetermined categories using rule-based machine learning . They obtained an accuracy of 48% on a dataset of questions on real-world contracts and 69% on a dataset of TREC questions.Using a parallel corpus, Gong et al.[26] developed a classifier-based tense model for SMS categorization . They used a dataset of 230,456 Chinese SMS sentences and obtained an accuracy of 84.39% .A skip-gram model, a variation of n-grams, is used by Yajni et al.[27] to investigate sentiment analysis in Nepali literature. Sentiment analysis is the process of determining if a text contains positive or negative sentiment. Using a dataset of 4,200 Nepali sentences, they were able to classify positive and negative statements with an accuracy of 89% .

Table -01: Literature Summary and Comparison with Proposed work

Authors	Data Source	Language	Class	Dataset Size	Accuracy
Bitto, Abu Kowshir, et al.[8]	Social media, Facebook, YouTube & Bengali Blogs.	Bengali	Good or Bad Discourse	1499 text	89.31%,
Hasan et al. [9]	Bengali speech dataset from songs and movies	Bengali	Joy, sadness, anger, surprise, fear, & disgust	21,000 voice clips & sentences.	51.33%
Rahman, Mafizur, et al.[10]	Bengali articles, newspapers	Bengali	Happy, angry, and excited.	11000 sentences.	75%
Khushbu, Sharun Akter, et al.[11]	Bengalis Newspapers	Bengali	11 classes	8000 headline sentences.	90%
Khushbu,Sharun Akter,, et al.[12]	Newspapers	Bengali	Entertainment, national,	13,445 news	84%.

			sports, city, and state news.		
Ria, Nushrat Jahan, et al. [13]	Bengali written sources	Bangla	Saint Common	1200 sentences	77%
Islam, Khondoker Ittehadul et al.[14]	Online news	Bengali	Positive, Negative, Neutral.	17,852	71%
Haque, Rezaul, et al.[15]	Social Media Comment	Bengali language	Political, Religious, Sexual, Acceptable Combined	84,072 Comments	85.8%
Wadud, Md Anwar Hussen, et al.[16]	Bengali online newspapers	Bengali Text	Race, Behavior, Physical, Sexual, Orientation, Class, Gender, Ethnicity,	Dataset containing 20,000 posts,	93.38%

			Disability, Religion, Revile, Addiction, Others		
Bijoy, Md Hasan Imam, et al.[17]	Social media, people conversation, and Bengali blogs	Bengali	NA	740 text document	96.39%
Mandal, Ashis Kumar, and Rikta Sen.[18]	Bangla News web sources	Bangla	Business, sports, health, technology, education	1000 text	89.14%
Hasan, Mehedi, et al.[19]	Self-made	Bengali	Assertive, interrogative, imperative, optative, or exclamatory	1824 Text	89.42%
Bitto, Abu	Facebook pages	Bengali	Positive, Negative.	1400 sentiments.	91.07%

Kowshir, et al.[20]					
Roy, Amartya, Kamal Sarkar et al.[21]	Leading Bengali newspapers	Bengali	NA	14,451	0.9512
Alam, Tanviru et al.[22]	Web pages, social media, online news portal, emails, online shops, user reviews, and questions and answers from customer Services.	Bengali and English	strongly positive, positive, neutral, negative, and strongly negative, Other	NA	93.8%
Sarkar et al.[23]	Bangla News comments	Bengali and English	NA	13803 records	83.77%
Naughton et al. [24]	News	English	Die, Injure, Attack, Meet, Tras, Charge	230,183 sentences	F1 92%
Sen, Prithvira	TREC comprising	English	NA	NA	69% on TREC

Methodology:

Classification of Saint (Sadhu bhasa) and Common (Cholito bhasa) from Bengali texts has become an important topic in recent times. To do so, choosing the appropriate methodology is also a pretty hard task. Because this step determines the effectiveness of categorization and prediction analysis, and a proper assortment for the method analysis will result in a system with specifications. In this section, we present our methodology workflow to classify Saint (Sadhu bhasa) and Common (Cholito bhasa) from Bengali text.

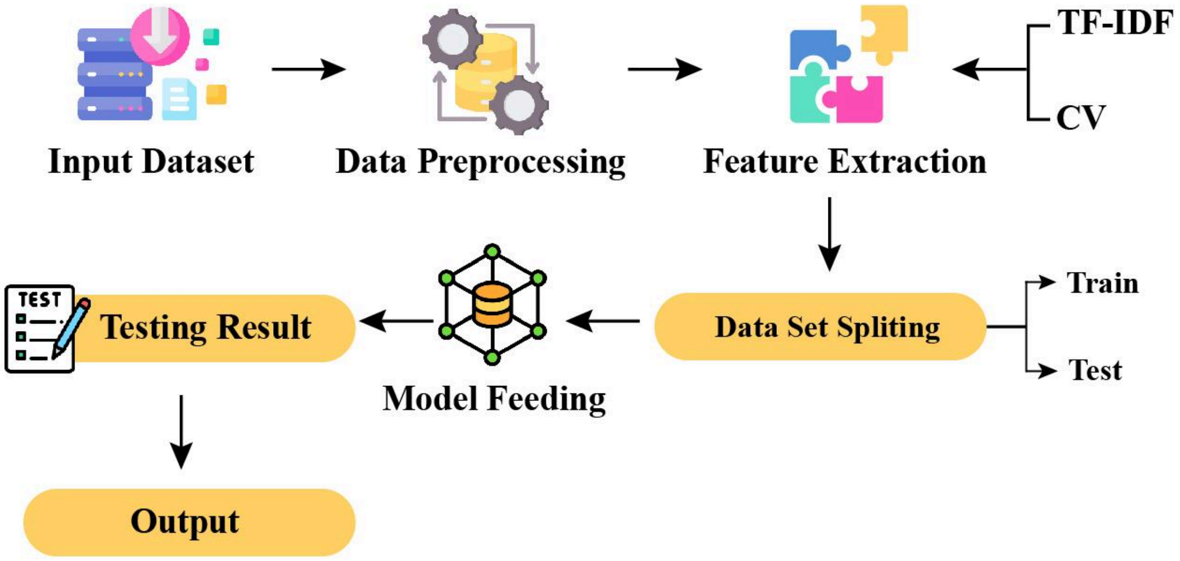


Fig.-01: Research Methodology

3.1 Dataset Collection and Properties:

Datasets play a vital role in research, and machine learning gains traction as data is gathered more regularly to feed the computer. We have collected 3005 Bengali texts in two classes—saint and common—covering a variety of subdomains, including agriculture, medicine, crime, education, social media, feelings, food, finance, etc. Some of the saint sentences are made by ourselves according to those subdomains, while others are taken from numerous historical books, novels, and literary works, such as চোখের বালি, চরিত্রহীন, লালসালু, etc. With the help of this Saint and Common Bengali Language dataset, machine learning (ML) can be used to learn the machine from the data, leading to revolutionary results in language processing.

Table-02: Sample Bengali Text data in Saint and Common Form

Sentences	Class
শিক্ষা প্রতিষ্ঠান সমাজ গঠনে গুরুত্বপূর্ণ ভূমিকা পালন করে	Common
মাটির উর্বরতা বজায় রাখার জন্য ফসলের আবর্তন কৃষিতে একটি সাধারণ অভ্যাস	Common
যাহা চাহিয়াছিলাম তাহা পাই নাই	Saint
তাহার জন্য আমার কথায় ব্যাঘাত কিছু ঘটিয়াছে	Saint

Our dataset contains 1510 Saint (0) form sentences and 1467 Common (1) form sentences of Bengali text.

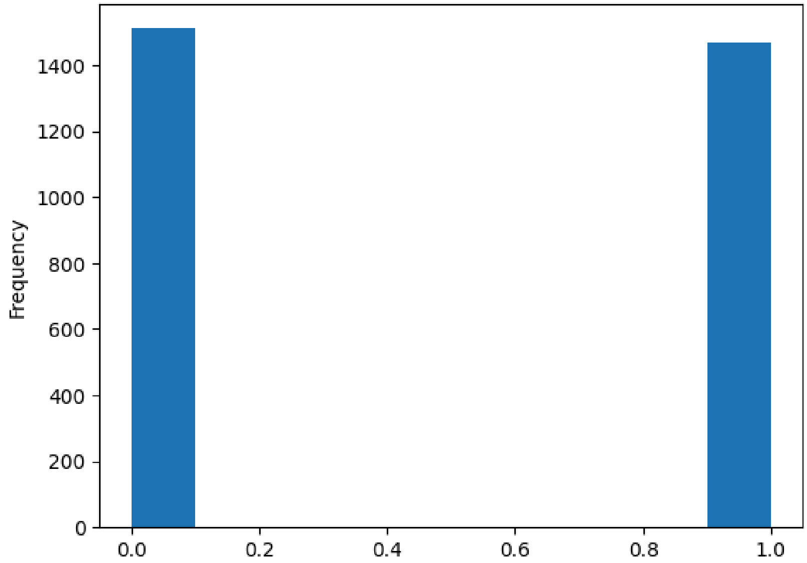


Fig.02: Statistical Analysis of Data

3.2 Dataset Preprocessing:

Data preprocessing is like a warm-up before a race. Preprocessing data actually optimizes the complexity of the data. For accurate predictions it is important to understand the data properly by machine. Only machine-accessible data can provide accurate predictions in any context. To get accurate accuracy we labeled our data in two forms as 0 for Saint(sadhu) form and 1 for Common(cholito) form of Bengali text. Table-04 illustrates the data label mapping or data class mapping

Table-03: Data Label Mapping

Sentences	Class
পৃথিবীতে অন্যায় দেখে আমার রাগে ভরে যায়	0
একটি ইতিবাচক মানসিকতা গড়ে তুলুন	0
তিনি নিজে গিয়াই ক্ষমা করিয়া আসিবেন	1
আপনি না গেলে ইঁহাদের চড়িভাতিতে যে কাণ্ডটা হইবে	1

Then we evaluated our data to check if the sentences had the three attributes of coherence, clarity, and imagery, as well as if the words in primary clean were accurately spelled. If any problems arose, we removed them. We may scrape the same sentence numerous times because we have a vast volume of data. To avoid model overfitting, we check for and eliminate duplicate values from our dataset. Textual data may include emoticons, unicode, numerals, punctuation, special characters, and mixed letters or words from multiple languages. Thus, deleting these unneeded characters improves the regular expression's ability to be cleaned up while also generating smooth and readable text forms.The table shows the sample data as well as the cleaned data after regular expressions are removed.

Table-04: Cleaned data after Removing Regular Expression

No.	Sample Input Text	Cleaned Data
01	“দেশের রাজনীতি দিনকে দিন পচে যাচ্ছে।!! **ধোরৈনকাএম কোমিকাএং**. How unfortunate. সুস্থ থাকা দায়.	দেশের রাজনীতি দিনকে দিন পচে যাচ্ছে সুস্থ থাকা দায়
02	আলো প্রতিফলিত হয়ে আশপাশের তাপমাত্রাও বাড়িয়ে তুলছে।!!!😓😓🥵🥵🥵	আলো প্রতিফলিত হয়ে আশপাশের তাপমাত্রাও বাড়িয়ে তুলছে

We washed the dataset by deleting null values, unnecessary words, and stop words. Stop words are common words that contribute little meaning, such as articles (a, the), prepositions (in, on), and conjunctions (and, but). In Bangla, these include terms like "আমি", "এবং", "সে", and "আরও", “ও”. By removing stop words it leaves behind more meaningful words that Improves accuracy and efficiency of natural language processing.

For instance, the stop words in the sentence "আমি এবং সে আরও একটি জয় পেয়েছি" are “আমি”, “এবং”, “সে”, “আরও”, “একটি”. After preprocessing stop words, the phrase may become “জয় পেয়েছি”.

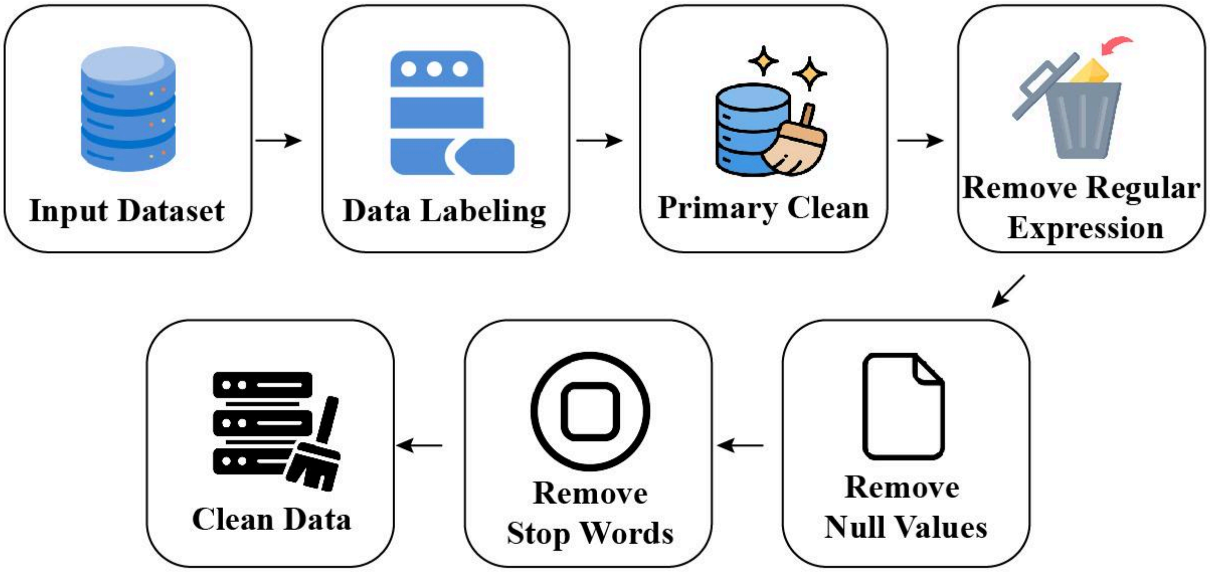


Fig.03: Data Preprocessing Techniques

After the data preprocessing is complete, a clean text suitable for classification is generated, and the total amount of updated and cleaned data is 2977. We selected 85% of the data for training our model and 15% for testing. To achieve higher levels of precision, we employed a large amount of data during training.

3.3 Feature Extractions:

Once the data has been cleaned up, we tokenize our text. The method of tokenizing a statement involves reducing it to a single word. Firstly, the count vectorizer is used to tokenize our data and it has broken down the individual text data into single words based on whitespaces. For example the text as “বিজয়ী বীর অভ্যর্থনা পেল “ has been transformed into an array individual words: [“বিজয়ী”,”বীর”,”অভ্যর্থনা”,”পেল”] for applying tokenization techniques on data.Secondly, countVectorizer and TF-IDF vectorizer has built a vocabulary of all unique words presented throughout entire text corpus and as an output it returns a vocabulary with 67 unique words by analyzing our dataset. Finally,it converts each unique word into a numerical vector form and provides a lexicon of the same density of 35.38.

3.4 Dataset Splitting:

Once our data preprocessing and feature extraction we got out cleaned and simplified the dataset. After completing data preprocessing we got 2977 preprocessed data. For training we used 85% data and used 15% for testing. To achieve higher accuracy, we used a large amount of data during training.The training data is 2530 and the test data is almost 447.

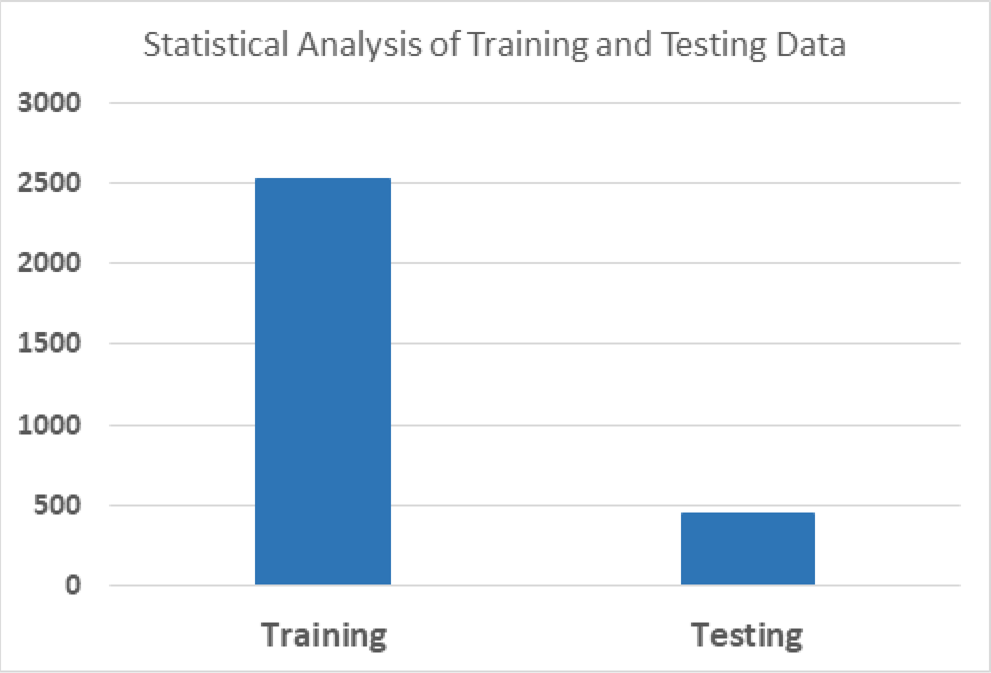


Fig.04: Statistical Analysis of Training and Testing Data

3.5 Model Feeding:

Many supervised and unsupervised models are available in machine learning to improve language processing tasks, particularly the classification of texts. With two vectorizers we applied three supervised machine learning models or classifiers on our Bengali dataset to accurately categorize Bengali text or sentences into two categories: Saint (sadhu) and Common (cholito). In this paper we used Support Vector Machine (SVM), Random Forest (RF) and Xgboost (XGB). A brief summary of each classifier, along with their performance based on our dataset analysis is provided below:

A. Support Vector Machine(SVM): Support vector machines are linear machine learning algorithms that address difficult text classification problems by utilizing supervised learning models.Finding the ideal hyperplane in an N-dimensional space to divide the data points into distinct classes in the feature space is the primary goal of the SVM method.This classifier is fit in self-made Bengali text dataset to classify Bengali text into two forms: Saint and Common and it has acquired the highest accuracy as 92.32%.Mathematical expression for linear SVM is mentioned as follows:

$$w^T x + b = 0 \text{ --------- } (1)$$

B. Random Forest(RF):A machine learning system called Random Forest uses trees to classify texts.When training, this ensemble model generates a large number of decision trees; for classification problems, the random forest's output is the class that the majority of the trees choose.Random forest is feeded on dataset and obtained 91.87% accuracy in terms of text data classification.Random forest with bootstrap aggregating looks like this:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \text{ --------- } (2)$$

C. Xgboost Classifier(XGB):A powerful machine-learning technique called XGBoostthat can assist in better understanding of data and decision-making. It is also an application of gradient-boosting decision trees. Xgboost classifier has obtained 91.19% accuracy by analyzing our dataset.The mathematical representation of this classifier is given below:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \text{ --------- } (3)$$

4. Experiment and Result Analysis:

We have used three machine learning models twitch. One for Counter Vectorizer and another for TF-IDF Vectorizer. All classifiers performed very well on our dataset, some performances are better than others. Here, using Counter vectorizer SVM and XGB prediction accuracy is 92.39% and RF got 91.94%. But using TF-IDF Vectorizer on those classifiers we achieved highest accuracy 93.06% with SVM and RF and XGB both prediction accuracy is 92.17%.

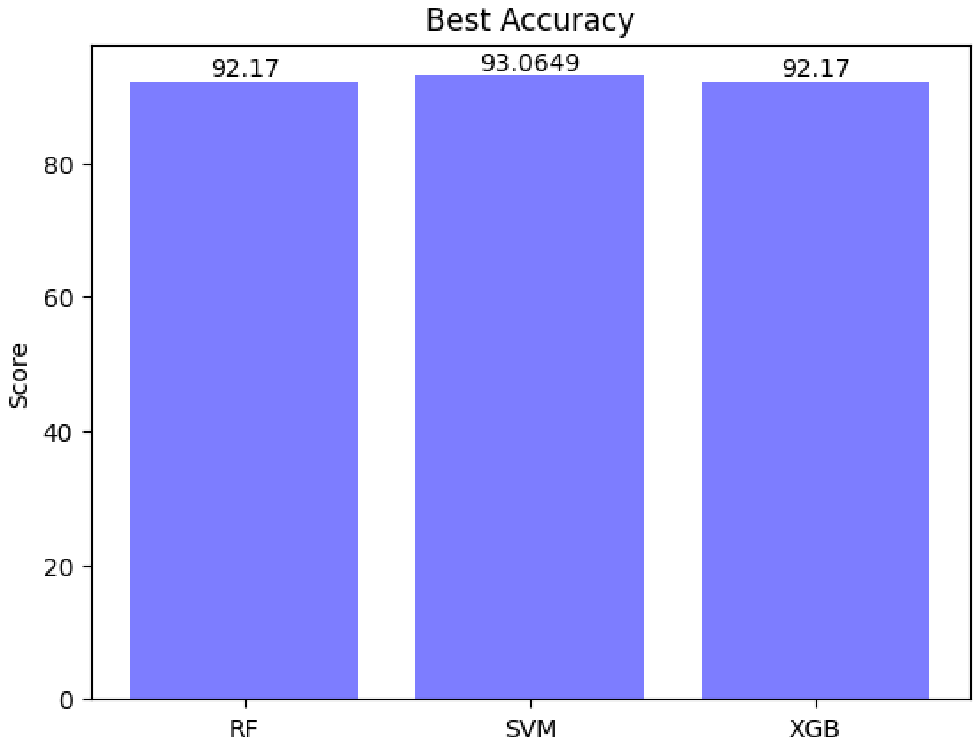


Fig.05: Accuracy Summary of Bengali Text Classifiers using TF-IDF Vectorizer

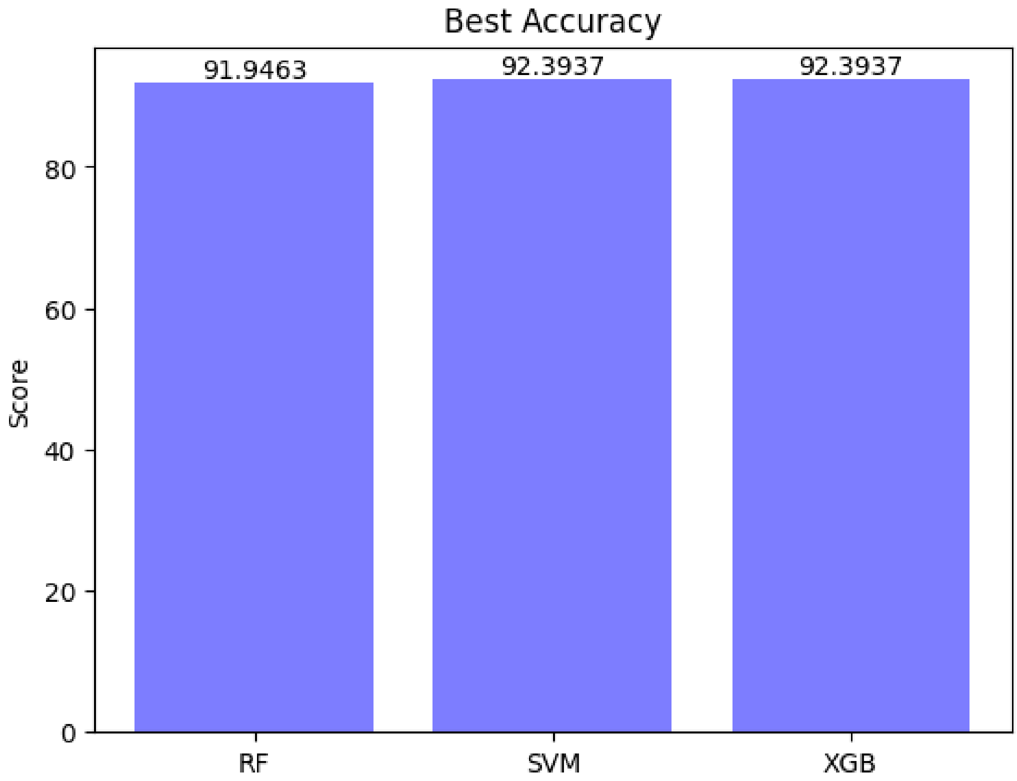


Fig.06: Accuracy Summary of Bengali Text Classifiers using Counter Vectorizer

We classify our data into two parts. One is Saint((Sadhu) and another is Common(Cholito)and the corresponding confusion matrix is as follows:

Table-05: Confusion Matrix

	Saint(Predict)	Common(Predict)
Saint(Actual)	Saint(True)	Common(False)
Common(Actual)	Saint(False)	Common(True)

We have measured performance of applied methods with parameters as: Accuracy, Recall, F1-Score, Precision and Support, TPR, TNR, and Error rate of each machine learning model.

Table-06: Comparison of Performance Matrix of Used three Models using Counter vectorizer

Algorithm		Precision	Recall	F1-Score	Support	Accuracy	TPR	TNR	Error Rate
SVM	0	0.94	0.90	0.92	208	92.39%	93.5 %	91.49%	7.60%
	1	0.91	0.95	0.93	239				
RF	0	0.95	0.88	0.91	208	91.94%	94.79%	89.80%	8.05%
	1	0.90	0.96	0.93	209				
XGB	0	0.93	0.91	0.92	208	92.39%	92.64%	92.18%	7.60%
	1	0.92	0.94	0.93	239				

Table-07: Comparison of Performance Matrix of Used three Models using TF-IDF Vectorizer

Algorithm		Precision	Recall	F1-Score	Support	Accuracy	TPR	TNR	Error Rate
SVM	0	0.94	0.91	0.92	208	93.06%	93.90%	90.8%	7.82%
	1	0.92	0.95	0.94	239				
RF	0	0.94	0.89	0.91	208	92.17%	94.02%	92.27%	6.93%
	1	0.91	0.95	0.93	239				
XGB	0	0.92	0.91	0.92	208	92.17%	91.78%	92.5%	7.82%
	1	0.93	0.93	0.93	239				

Here, Table-06 and Table-07 shows the comparison between Count vectorizer and TF-IDF Vectorizer with three text classifier algorithms Support Vector Machine(SVM), Random Forest(RF) and Xgboost Classifier(XGB). Although with the highest accuracy among other related works we also tried an example input from an online newspaper to see whether algorithms could correctly anticipate the input's common or typical form, and algorithms could predict the output exactly.

Table-08:Predictions of Saint Form for both Vectorizers

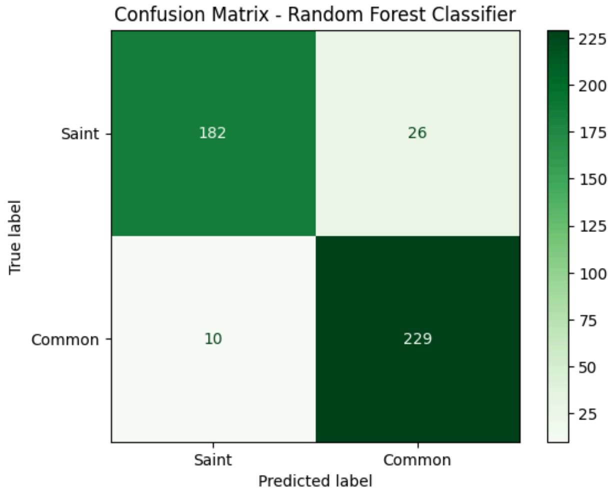
Original Text	অনেকক্ষণ আর কেহ কোনো কথা কহিল না।
Original Prediction	Saint
Input Text	অনেকক্ষণ আর কেহ কোনো কথা কহিল না।
Prediction of Algorithms	
XGB	Saint
RF	Saint
SVM	Saint

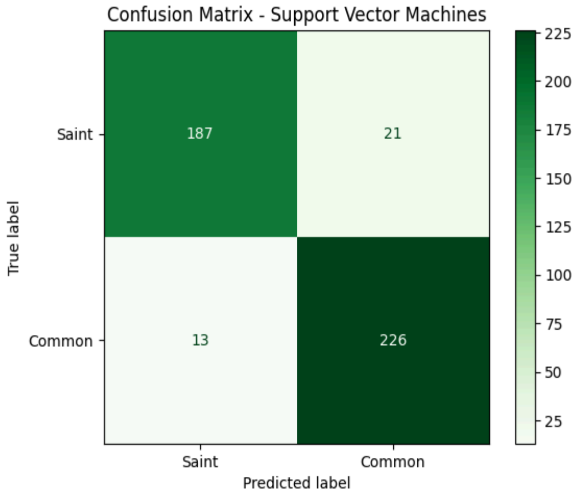
Table-09: Predictions of Common Form for both Vectorizers

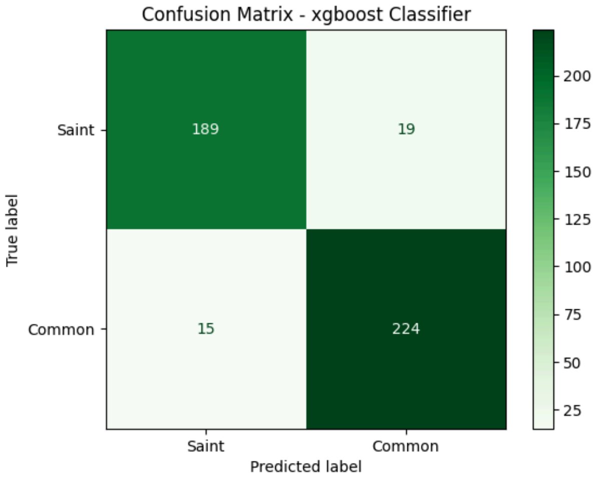
Original Text	আলো ও বাতাসের জন্য আগের ভবনগুলো ছিল জানালা নির্ভর।
Original Prediction	Common
Input Text	আলো ও বাতাসের জন্য আগের ভবনগুলো ছিল জানালা নির্ভর।

Prediction of Algorithms	
XGB	Common
RF	Common
SVM	Common

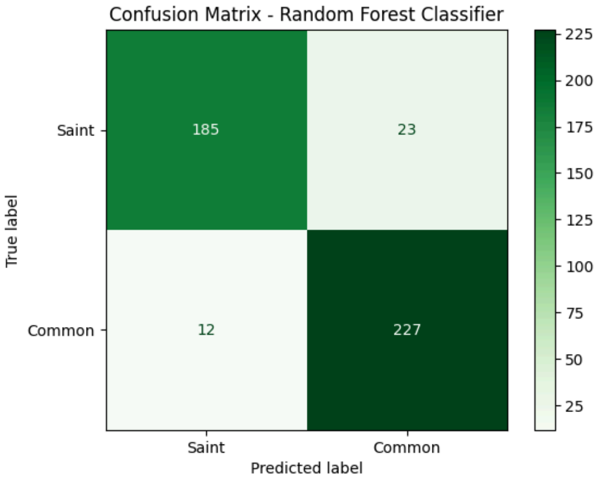
Confusion Matrices of while using Counter Vectorizer on three classifiers or models given below:

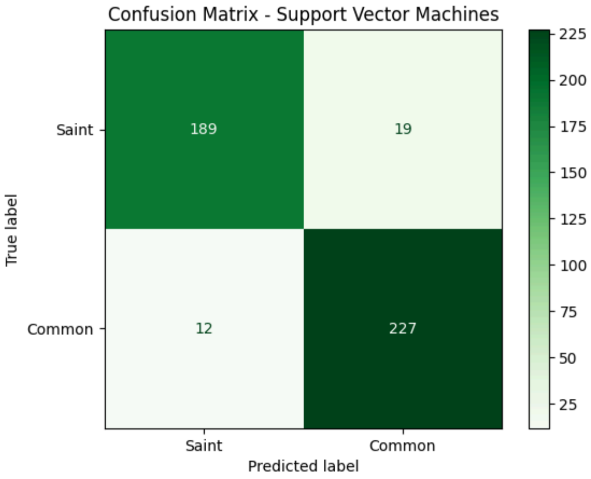


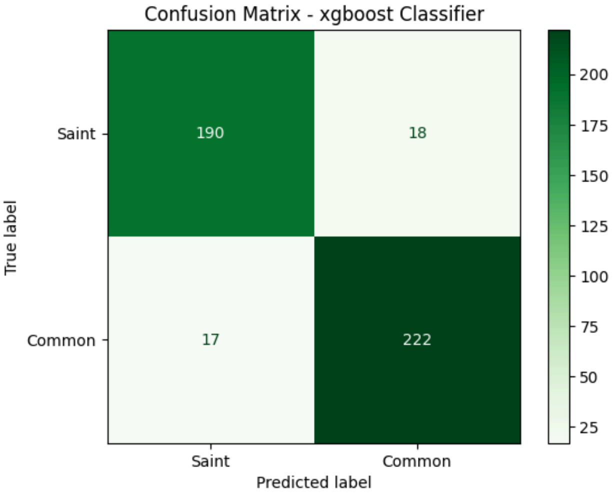




Confusion Matrices of while using TF-IDF Vectorizer on three classifiers or models given below:







Conclusion and Future Work:

The classification of Bengali texts with a focus on Saint and Common forms is a developing and exciting area of natural language processing in Bangla. This study shows a dataset with approximately 3005 sentences(data), which is an optimal size for testing two vectorizers with three classifier methods each, to effectively distinguish between Saint and Common forms in Bengali texts. The research demonstrated diversity in algorithm application. Using the CountVectorizer, SVM and XGB achieved a maximum accuracy of 92.39%, while RF achieved 91.94% accuracy. With the TF-IDF Vectorizer, SVM reached a peak accuracy of 93.06%, and both RF and XGB achieved 92.17% accuracy. This comparative study aimed to identify Bengali sentences accurately, distinguishing between Saint and Common forms. Future work will explore deep learning methods to develop a model that enhances Bengali text classification, applying Bangla natural language processing techniques to a larger dataset.

References:

[1] https://lingo-star.com/bengali-language/?v=4326ce96e26c

[2] "Bengali language." *New World Encyclopedia,* . 27 Sep 2023, 09:13 UTC. 19 Apr 2024, 05:41 <https://www.newworldencyclopedia.org/p/index.php?title=Bengali_language&oldid=1123421>

[3] https://www.thedailystar.net/shout/cover-story/news/evolution-bangla-1705177

[4] Britannica, The Editors of Encyclopaedia. "Bengali language". *Encyclopedia Britannica*, 15 May. 2024, https://www.britannica.com/topic/Bengali-language. Accessed 17 May 2024.

[5] Afrin, Sazia, and Mohammad Rezaul Islam. "Impact of Bangla Language on English Language Learning of the Undergraduate Students of Dhaka University." *Journal of Teacher Education* 8.1 (2023): 99-110.

[6] Faquire, A. B. M. R. K. "Language situation in Bangladesh." *The Dhaka University Studies* 67.2 (2010): 63-77.

[7] Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82.3 (2023): 3713-3744

[8] Bitto, Abu Kowshir, et al. "Approach of Different Classification Algorithms to Compare in N-gram Feature Between Bangla Good and Bad Text Discourses." *Machine Intelligence Techniques for Data Analysis and Signal Processing: Proceedings of the 4th International Conference MISP 2022, Volume 1*. Singapore: Springer Nature Singapore, 2023.

[9] Hasan, HM Mahmudul, and Md Adnanul Islam. "Emotion recognition from bengali speech using rnn modulation-based categorization." *2020 third international conference on smart systems and inventive technology (ICSSIT)*. IEEE, 2020.

[10] Rahman, Mafizur, et al. "A dynamic strategy for classifying sentiment from Bengali text by utilizing Word2vector model." *Journal of Information Technology Research (JITR)* 15.1 (2022): 1-17.

[11] Khushbu, Sharun Akter, et al. "Neural network based bengali news headline multi classification system: Selection of features describes comparative performance." *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2020.

[12] Khushbu, Sharun Akter, et al. "Neural network based bengali news headline multi classification system: Selection of features describes comparative performance." *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2020.

[13] Ria, Nushrat Jahan, et al. "Toward an enhanced bengali text classification using saint and common form." *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2020.

[14] Islam, Khondoker Ittehadul, Md Saiful Islam, and Md Ruhul Amin. "Sentiment analysis in Bengali via transfer learning using multi-lingual BERT." *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2020.

[15] Haque, Rezaul, et al. "Multi-class sentiment classification on Bengali social media comments using machine learning." *International journal of cognitive computing in engineering* 4 (2023): 21-35.

[16] Wadud, Md Anwar Hussen, et al. "How can we manage offensive text in social media-a text classification approach using LSTM-BOOST." *International Journal of Information Management Data Insights* 2.2 (2022): 100095.

[17] Bijoy, Md Hasan Imam, et al. "An automated approach for Bangla sentence classification using supervised algorithms." *2021 12th international conference on computing communication and networking technologies (ICCCNT)*. IEEE, 2021.

[18] Mandal, Ashis Kumar, and Rikta Sen. "Supervised learning methods for bangla web document categorization." *arXiv preprint arXiv:1410.2045* (2014).

[19] Hasan, Mehedi, et al. "Multiple Bangla Sentence Classification using Machine Learning and Deep Learning Algorithms." *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2022.

[20] Bitto, Abu Kowshir, et al. "Sentiment analysis from Bangladeshi food delivery startup based on user reviews using machine learning and deep learning." *Bulletin of Electrical Engineering and Informatics* 12.4 (2023): 2282-2291.

[21] Roy, Amartya, Kamal Sarkar, and Chintan Kumar Mandal. "Bengali Text Classification: A New multi-class Dataset and Performance Evaluation of Machine Learning and Deep Learning Models." (2023).

[22] Alam, Tanvirul, Akib Khan, and Firoj Alam. "Bangla text classification using transformers." *arXiv preprint arXiv:2011.04446* (2020).

[23] Sarkar, Ovi, et al. "An experimental framework of bangla text classification for analyzing sentiment applying CNN & BiLSTM." *2021 2nd International Conference for Emerging Technology (INCET)*. IEEE, 2021.

[24] Naughton, Martina, Nicola Stokes, and Joe Carthy. "Sentence-level event classification in unstructured texts." *Information retrieval* 13 (2010): 132-156.

[25] Sen, Prithviraj, et al. "Learning explainable linguistic expressions with neural inductive logic programming for sentence classification." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

[26] Gong, Zhengxian, et al. "Classifier-based tense model for SMT." *Proceedings of COLING 2012: Posters*. 2012.

[27] Yajni, Archit, and Ms Sabu Lama Tamang. "CHUNKER BASED SENTIMENT ANALYSIS AND TENSE CLASSIFICATION FOR NEPALI TEXT."