Social Media Post Classification in Bangla Language: A Comparison Between Boosting and Traditional Machine Learning Algorithm

BY:

Faisal Ibn Salam
181-15-10867

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

Supervised By:

Mr. Md. Aynul Hasan Nahid

Lecturer

Department of CSE

Daffodil International University

Co Supervised By:

Md. Majidur Rahman

Lecturer

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

January 2023

APPROVAL

This Project/internship titled **"Social Media Post Classification in Bangla Language:A Comparison Between Boosting and Traditional Machine"**, submitted by **Faisal Ibn Salam, ID No.:181-15-10867** 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 28 January,2023.

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

Dr. Fizar Ahmed
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Taslima Ferdaus Shuva
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Md Sazzadur Rahman
Associate Professor
Institute of Information Technology
Jahangirnagar University

DECLARATION

We hereby declare that this thesis has been done by us under the supervision of **Mr. Md. Aynul Hasan Nahid, Lecturer, Department of CSE,** and co-supervision of **Mr. Majidur Rahman, Lecturer, and Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Mr. Md. Aynul Hasan Nahid

Lecturer
Department of CSE
Daffodil International University

Mr. Majidur Rahman

Lecturer
Department of CSE
Daffodil International University

Submitted by:

Faisal

Faisal Ibn Salam

ID: 181-15-10867
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty God for His divine blessing makes it possible to complete the final thesis successfully.

We are really grateful and wish our profound indebtedness to **Mr. Md. Aynul Hasan Nahid**, Lecturer, Department of Computer Science and Engineering Daffodil International University, Dhaka, Bangladesh. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning and NLP" to carry out this thesis. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting us at all stages have made it possible to complete this thesis.

We would like to express our heartiest gratitude to Prof. **Dr. Touhid Bhuiyan**, Professor and Head, Department of Computer Science and Engineering, for his kind help to finish my thesis and also to other faculty members and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate 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 passion of my parents.

ABSTRACT

To communicate with others, social media appears to have surpassed all previous methods. Millions of postings are made every day on social media. The Bangla language is widely utilized on Bangladeshi social media platforms. Classifying them based on textual information is often challenging. Social media posts defy easy categorization. Examination becomes more of a chore when the text is written in the Bangla language. Our goal is to classify these social network posts according to their emotional content, making them more accessible for searching, filtering, and organizing. As a means of gauging the persuasiveness of the posts, we conducted an analysis utilizing Sentiment Analysis. Moreover, we attempted to show by contrasting conventional machine learning with the boosting approach. GradientBoostingClassifier and XGBoost Random Forest Classifier were used for boosting, while Support Vector Classifier and Stochastic Gradient Descent were utilized for standard machine learning (SGD). To rate Bangla-language social media postings, we employed an algorithm with the most reliable results.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

In the past decade, the use of social media has skyrocketed in popularity. Around 40% of the world's population, or three billion people, are active users of social networking sites. The average daily time spent on social media by users is two hours [1]. As of January 2020, 66.44 million people in Bangladesh have access to the web. As of January 2020, there are 36 million people in Bangladesh who use social media. The number of persons utilizing social media grew by 3.0 million (+9.1%) between April 2019 and January 2020 [3]. The expectation in Bangladesh is that everybody who uses the internet will also have a social media profile. Sharing our experiences on social media is a priority. In the first stages of online communities, it served as a way of communication. But peculiarities have evolved considerably throughout the years. Consequently, the linguistic material on social media is a goldmine of information. Document representation, feature selection/representation, and algorithm estimation are the three stages of text categorization. Conceptual interpretations of document collections provided by text classification can have substantial effect in the real world. There is a vast variety of topics discussed in social media posts. People's views on social media aren't categorized or sorted in any meaningful way. There is a greater challenge in classifying when the postings are written in Bangla. Sentiment analysis, on the other hand, has been widely studied in English but less studied in Bangla. When it comes to classifying texts in Bangla, there hasn't been much research done. Our every move, every word, and every tone creates data that may be displayed and retrieved for analysis. We may be able to use this information to analyze and perhaps even predict human behavior with the help of natural language processing. Words and phrases can have their meanings and structures uncovered by use of algorithms used in natural language processing. Various statistical approaches are used in machine learning for NLP and text analytics to identify and categorize linguistic features such as words, names, and emotions. Natural language processing allows us to classify people's emotions and get insight into their minds. Extracted features from the textual information. We extracted the underlying sentiment and put the data into appropriate categories using a prediction approach applied to the dataset. We're trying to create a taxonomy of social media platforms' perspectives so that users may more quickly search, filter, and organize content. This was achieved with the use of a number of machine learning

(ML) toolkits and libraries, including Sci-kit Learn, Numpy, Matplotlib, and Pandas, and a number of natural language processing (NLP) tools, such as NLTK (Natural Language Toolkit) and TF-IDF. We processed and translated data collected from social media sites like Facebook, Twitter, and YouTube into labeled data, which we subsequently displayed. Several types of machine learning were crucial in helping us achieve success and wrap up the project. The main goal of this research is to develop a system that can categorize Bangla social media posts into several topics such as "emotion," "sports," "news," "politics," "food," and "events."

1.2 Motivation

People in general invest a considerable quantity of time perusing social media. Most native-speaking Bengalis will use the Bangla language when doing any sort of online activity. Many people share their thoughts and feelings through posts and comments on social media platforms like Facebook and Twitter on a wide range of topics, from current events and politics to food and the arts. However, not everyone enjoys reading a lot. There are a number of folks who would rather read a sports story than a political one. For some, reading political posts and comments is more entertaining than making their own. There are those who take pleasure in perusing entertainment boards. The ultimate choice rests with the person concerned. If the proper individuals agree with your assessment, your social media messages will become more effective. This is why it's essential to sort social media updates into relevant categories.

Occasionally, we find collections of projects written in English. Whatever the situation may be, the Bangla output is woefully inadequate. There are many potential uses for the Bangla language if it can consistently acquire exploration tokens. In addition, we think that the emphasis on program organization is a hallmark of modern life. Customers have faith that they will receive the larger things because of the system in place. A competent framework is one that can evaluate information presented by humans and make a choice. Skillful mining is required to select a monarch above others.

All of these things combined to pique our interest in working on this fascinating problem. Our research is focused only on Bangla-language social media posts and makes use of NLP and machine learning to categorize them.

1.3 Problem Definition

It's no secret that many automated tasks and robotics systems make use of techniques or ideas from the many English-language publications on Natural Language Processing (NLP). However, the capabilities of Bangla NLP are rather restricted. To realize more automated programs or to carry out much more competent Machine Learning procedures in Bangla, one must deal with Bangla text. This Bangla post categorization has given us the idea to do action. The notepad editors of today are clearly more sophisticated than their predecessors. Among the many fascinating options are automatic edits, grammatical checks, and suggestions for what to write next. NLP has been the saving grace behind these developments. However, these characteristics are most often linked to the English language. These sorts of flourishes don't appear frequently in Bangla literature. As a result of these, we have begun collaborating on blogs in both Bangla and Text. As this area pertains specifically to Bangla text, it is also important to analyze the problems and needs it faces before a suitable solution can be offered.

1.4 Research Questions

- Can the groups online provide a detailed definition of the article?
- How do regular people in Bangladesh use social media to share their thoughts and feelings with the world?
- How might tagging content help make social media more enjoyable?
- What degree of accuracy does the machine learning technique have in determining the type of postings being discussed?

1.5 Research Methodology

The Experiment data, pre-processing, tokenization, algorithm implementation, model training, and algorithm assessment will all be discussed here. At the conclusion of the chapter, we will evaluate the model's effectiveness.

1.6 Research Objectives

- To look into how a classifier algorithm may be used to sort and categorize Bangla threads.

- The goal of this project is to develop a system that can classify social media postings made in Bangla.
- Goal: To graphically represent the results of a statistical study of classifier algorithms applied to a dataset of Bangla forum posts.
- To use machine learning and natural language processing (NLP) to create a system for recognizing a wide variety of post types, including but not limited to: News, Food, Emotion, Events, Politics, Information, and Sports.

1.7 Research Layout

The report will have appeared as regards:

The study is summarized in the first chapter. Research in the introduction is the most vital aspect of the first segment. In any case, this chapter provides sufficient justification for engaging in such investigation. Specifically, this chapter's Problem Definition section is crucial. In this chapter's final part, we lay out the specific aims and hypotheses that will be tested.

Chapter 2 is dedicated to a critical evaluation of previous research on this topic. Many projects, some of which are based on the Bangla language and others of which use machine learning and natural language processing.

Chapter 3 explains the research process in detail, including the statistical methods that were applied. This chapter also provides examples of the methods and procedures used by the Machine Learning classifier. Everything from raw data to processed data to data analysis to tokenization to classifier approaches and everything in between is discussed.

The results are analyzed in Chapter 4. Meaning, it's about what comes out of the study. Images depicting the final product of the planned work may be seen in this section.

Research participants are the focus of Chapter 5. This section is capable of delivering the whole performance report per the requirements. The chapter finishes by illuminating the shortcomings of our work and how they can effect future professionals in the field.

1.8 Expected Outcome

To prescribe given Bangla posts as for the welded model of a prepared dataset, this exploratory investigation is anticipated to yield a consideration or accumulate the producing policy. The expected result is a categorization of Bangla postings into seven groups, allowing for the identification of which posts are assigned to which regions.

CHAPTER 2

BACKGROUND

2.1 Introduction

An extremely common use of natural language processing is in the field of sentiment analysis. The question of how to categorize texts in various languages has been examined extensively. Past work in this area was discussed. This chapter provides evidence of the vital procedures that have been successfully completed by a number of specialists in the aforementioned field during the past several years.

2.2 Related Works

One of the most common uses of natural language processing is in sentiment analysis. Language document categorization is a hot issue. Previous research of relevance were mentioned here.

In [4], S. Limon et al. proposed making a very basic website to label or classify news. Nine distinct data sources were used to categorize six distinct types of news stories into a total of six TSV (tab-separated values) files. These six TSV recordings served as raw information. Many different ML techniques, including Naive Bayes, SVM, Decision Tree, KNN, and Random Forest, were used. Accuracy was provided by Nave Bayes, which they used 76.94% of the time.

Using several ML methods, M. S. et al. [5] developed a Bangla news classifier. They used Nave Bayes, Decision Tree, KNN, SVM, and Random Forest classifiers, among others. To calculate precision, they used a confusion matrix. After a thorough analysis, it was determined that Nave Bayes is the most accurate of the ML algorithms. For Nave Bayes, that efficiency equated to about 85%.

Internet news classification was the topic of research undertaken by U. Suleymanov et al. [6]. The stories published in Azerbaijani media served as the source material. They played around with the dataset to try out different ways of converting the textual information into TF-IDF numbers. They used an unsupervised clustering method called K-means. Support Vector Machines and an ANN were also used. The Artificial Neural Network outperformed the rest of the methods by a significant margin (89 percent).
It was the goal of L. Nahar et al. [7] to categorize Bangali political and sports news on social media platforms. Data was gathered from a wide range of online sources, including news articles, social media posts, and user comments. Preprocessing included activities such as tokenization, stemming, text extraction, and dictionary word matching. Later, many classifiers such as the SVM, Neural Networks, and the Naive Bayes classifier have been implemented. The Naive Bayes technique yielded the best results.

The goal of T Islam et al[8] .'s Facebook political bias filter was to reduce the spread of extremist views. Two hundred data points were collected (fifty comments and one hundred fifty posts), and a lexicon of 239 unique phrases was developed for comparison. With the help of the Nave Bayes method, they were able to boost their accuracy to an astounding 83%.

Researching the Bangla news headline, M. M. Islam et al. [9] created a system for classifying the emotional tone of news headlines in Bangla. Researchers incorporated extraction and Tally Vectorizer for word count within data preparation. They created a two-tiered news system. SVM accuracy was found to be 75% with 1600 datasets, whereas Naive Bayes accuracy was found to be 73% with the same number of datasets.

Articles written in Bangla may be classified using supervised algorithms, which A. K. Mandal et al. [10] investigated. The feature extraction process used TF-IDF and normalization, and the four classifier algorithms used were KNN, SVM, Nave Bayes, and a Decision tree. A k-fold approach was used to divide 1000 data points into smaller groups. In addition, they factored in the estimated time required for training each approach, finding that SVM required the least amount of time.

By reading 1020 comments and labeling them as either favorable or negative, Shafin et al. [11] studied collaborative filtering in online product reviews. The maximum accuracy (88.81%) was attained using SVM when TF-IDF vectorizers were used as model features.

Similar to this, M. H. Rahman et al. [12] compiled and categorized 6281 Bangla book reviews sourced from various web sources. To classify the study, they used several models, with CNN-LSTM achieving the highest accuracy (97.22%).

2.3 Comparison of related work

RELATED WORK	ACCURACY RATE
Strategies for supervised learning in the classification of Bangla language online documents	89.14%
Tokenization is the first step in natural language processing.	87%
A Research Project on the Application of Machine Learning Techniques to the Categorization of Online News	89%
Separating Textual Information from Social Media Posts About Politics and Sports in Bengali,	90%
Comparison of Multiple Supervised Learning Algorithms for Sentiment Classification of Bengali News Headlines	75%
Extraction of political opinion from social media posts	83%
Institution of a Tabulation Scheme for Bangla Newspapers	85%
Using ML to Categorize Bangla News,	76.94%
Preprocessing of source code data using natural language processing: an empirical study	85%

Table 2.1 Comparison Table of related work

As we've seen, there is a dearth of research on sentiment analysis in Bangla. If we look at the relevant work, we can see that our model outperforms the competition thanks to its larger dataset and its performance across more categories.

2.4 Research Summary

The aforementioned analysis was conducted on several research papers by various research organizations, and it is still being refined to show that there is an increasing body of new study on Bangla text. Multiple promising results support this claim. Despite the lack of sufficient resources, there is hope that the area will become more prosperous over time.

2.5 Challenges

The primary challenges of this research are related to the dataset. Unfortunately, there aren't many tried-and-true methods that can be used to successfully conduct the dataset refining. A lack of resources is also a problem for this endeavor.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Our work may be broken down into seven distinct phases, each of which is illustrated in Figure 3.1. Specifically, here is what you need to do:

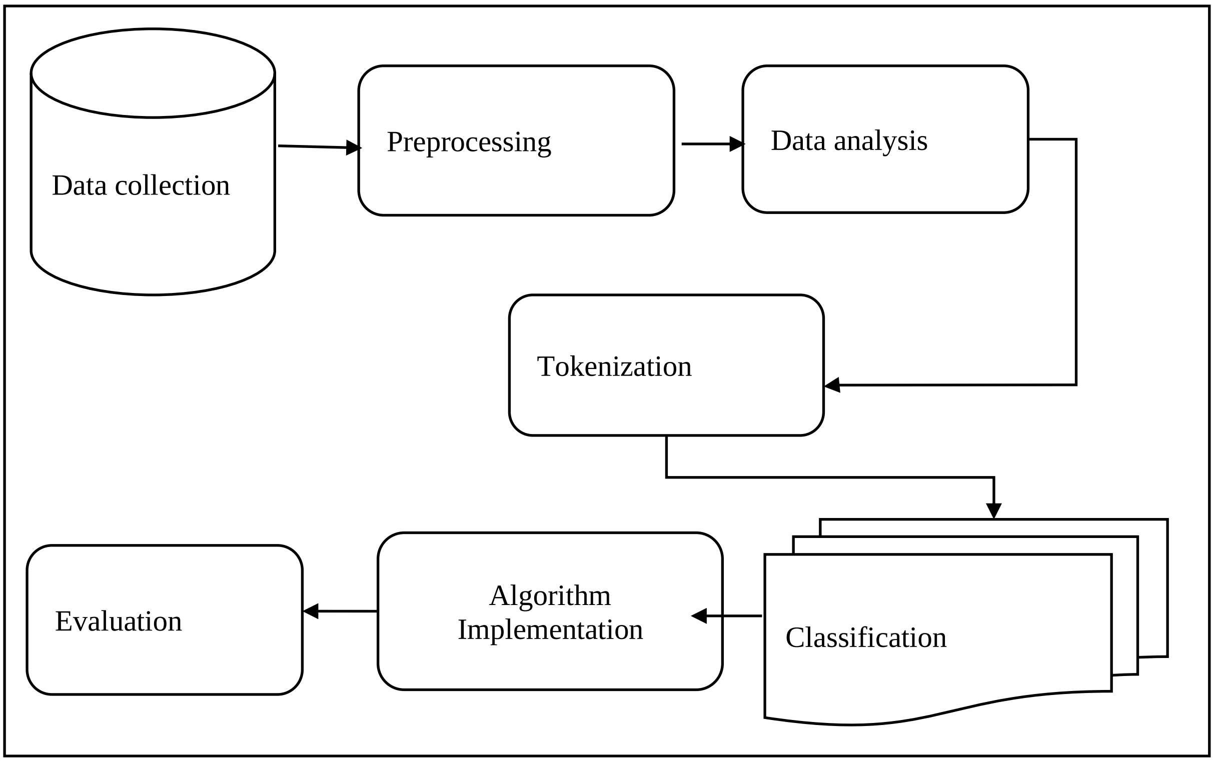


Figure 3.1: Methodology diagram

3.2 Data Collection

Every every day, there are literally billions of fresh articles that are made available online. There is an emotional component to every point of view. All of our social media efforts were concentrated on Bangla-language material because Bangla was the only language we worked with. Then, we assembled Bangla-language social media posts that met our criteria for being both credible and threatening. Over 4,012 comments have been gathered up for examination at this time.

Source	Amount of data
Facebook	1000
Youtube	1200
Instagram	800
Twitter	500
Different Bangla blog	512

Table 3.1 Data Sources

Table 3.1 represent the source and data collection amount. We collected bangla sentence from facebook about 1000, youtube 1200, Instagram 800 twitter 500 different Bangla blog 512.

3.3 Pre-Processing

X has confirmed this. The analysis of primary data reportedly takes a lot of time, as stated in L et al. [13]. As a consequence, preparation is necessary; during this time, the data must be transformed into a format that is structured. We started cleaning up the data by deleting any punctuation and stop words that were written in Bangla. Following that, the emoji in question was removed from the primary post content. Because emoji are such an important part of the feeling that is transmitted by a piece of writing, we have decided to include one at the end of each article so that you can make a more in-depth analysis of the piece's disposition. Figure 3.2 represents the data preprocessing steps.

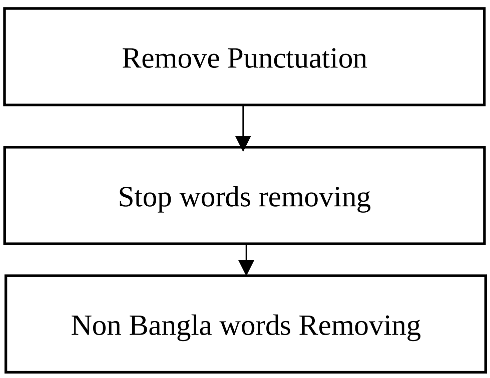


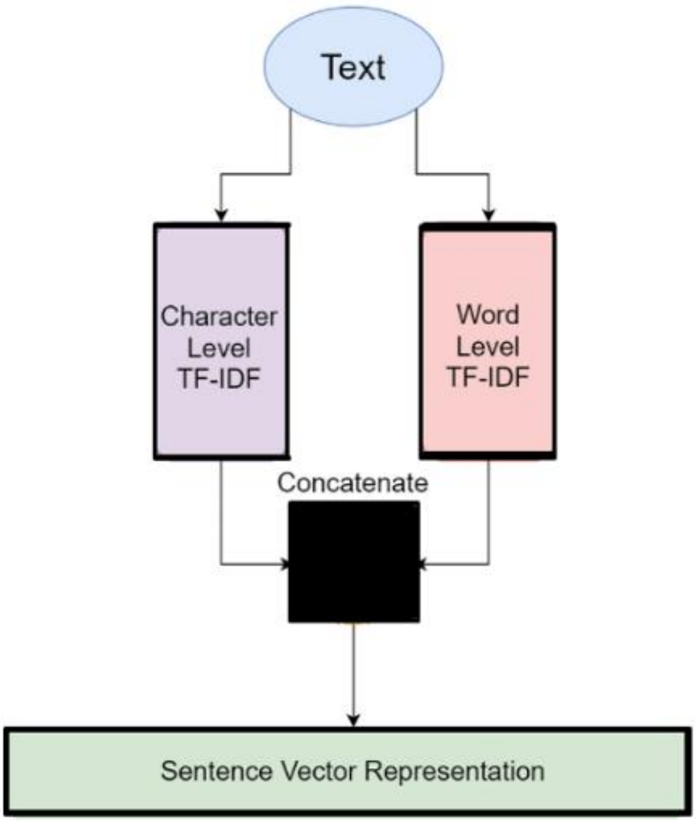
Figure 3.2: Data Pre-processing steps

3.4 Data Analysis

Following the removal of duplicates, we sorted the information according on how we felt about it based on the post-analysis results (food, event, emotion, sports, politics, informative and news). because computer programs can never comprehend string representations in their entirety. As a direct result of this, we were required to put all of our research findings into numerical form. In order to accomplish this, we made use of a method known as TF-IDF (Term Frequency-Inverse Document Frequency). When searching for Term I in File j:

$$W_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right) \quad (1)$$

The TF-IDF score is denoted by (1) W i, j. The frequency with which I occurs in set j is denoted by tf i, j. N is the number of files that need to be processed. The df i variable represents the total number of documents that include i. Figure 3.2 represents the types of tfidf algorithm. from this figure we can see that there are two sector of tfidf. In our work we



used word level tifidf.

Figure 3.3: Types of tfidf

3.5 Tokenization

Tokenization is the act of replacing personally identifiable information (PII) with anonymous information (or "tokens") that may be kept and used in a database or internal system without exposing any PII to the public. PII refers to information that can be used to identify an individual. Tokens are independent values, yet they maintain some features of the original data, such as the length and format of the data, to guarantee that business may proceed unabated. After that, the original, secret data is stored in a safe location that is disconnected from the business's network. According to J. J. Webster and colleagues [14], tokenization is an essential part of natural language processing (NLP).

Raw Data	Type	Tokenized data
প্রতিটি পদ ভালোভাবে শক্ত করে দখল করুন তা না হলে টিকতে পারবেন না	Political	'প্রতিটি', 'পদ', 'ভালোভাবে', 'শক্ত', 'করে', 'দখল', 'করুন', 'তা', 'না', 'হলে', 'টিকতে', 'পারবেন', 'না'
বৃষ্টির সময় বাসায় বসে পিজ্জা খাওয়ার আনন্দই অন্যরকম	Food	'বৃষ্টির', 'সময়', 'বাসায়', 'বসে', 'পিজ্জা', 'খাওয়ার', 'আনন্দই', 'অন্যরকম'
করোনা জয় করে এ পর্যন্ত সুস্থ হয়েছেন ১০ হাজার ৭৬ পুলিশ সদস্য	News	'করোনা', 'জয়', 'করে', 'এ', 'পর্যন্ত', 'সুস্থ', 'হয়েছেন', '১০', 'হাজার', '৭৬', 'পুলিশ', 'সদস্য'
মেস্তির অসাধারণ গোল অনেকদিন পরে গোল দেখে রনের কথা মনে পড়ে গেলো	Sports	'মেস্তির', 'অসাধারণ', 'গোল', '', 'অনেকদিন', 'পরে', 'গোল', 'দেখে', 'রনের', 'কথা', 'মনে', 'পড়ে', 'গেলো'
আজ বিজয় দিবস ইচ্ছা ছিল না কিন্তু আর পারলাম না অনেকটা অপারগ হয়েই পোস্ট দিলাম	Events	'আজ', 'বিজয়', 'দিবস', '', 'ইচ্ছা', 'ছিল', 'না', 'কিন্তু', 'আর', 'পারলাম', 'না', '', 'অনেকটা', 'অপারগ', 'হয়েই', 'পোস্ট',' দিলাম"

Table 3.2 Tokenization Table

Tokenization allowed us to extract individual words for further study of the phrase. Table 3.1 is an effort at visualizing the tokenization process. Each string must be tokenized before the TF-IDF method can be applied to it. Table 3.2 displays the input that was derived from genuine social media data. After that, the commas and the stop word were taken out. Then, we tokenized our initial data using porter stemmer, a popular tokenization tool.

3.6 Classification

According to T, one of the most important considerations when selecting a choice is your level of confidence. Ghorpade et al. [15], classifying into several categories. The Bangla content was divided into seven distinct groups. Therefore, all of the information in our database has been separated into the following zero through six digit categories: cuisine, media, events, politics, education, and feelings. In each Bangla forum, there were about 500 total postings.

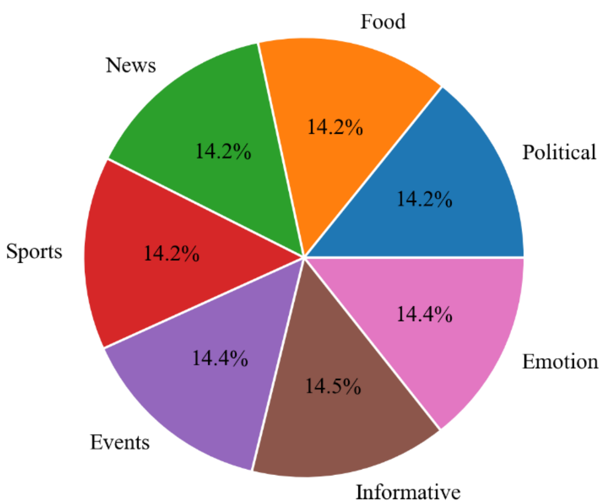


Figure 3.4: Classification of dataset.

Figure 3.2 depicts the dataset segmentation. Further, it breaks down the data by class and displays the percentages for each. In most cases, 14.2 percent of data is lost.

3.7 Algorithm Implementation

Whenever there is a variable in a model whose value cannot be determined by inspecting the data, we refer to it as a hyperparameter. Procedures utilize them frequently to aid in parameter estimation for models. They are typically outlined by the doctor themselves. Heuristics can be used to set them frequently.

Algorithms	Details
Gradient Boosting Classifier	random_state=0
XGBRF classifier	random_state=0
SVM	Kernel = linear
SGD classifier	max_iter=1000, tol=1e-3,n_jobs=2,penalty="l2"

Table 3.3 PARAMETER USAGE

Table 3.3 details the settings and other inputs utilized in our implementations of the chosen algorithms.

3.7.1 Gradient Boosting Classifier

A functional gradient approach known as "Gradient Boosting" repeatedly chooses a function that points in the direction of a weak hypothesis or a negative gradient in order to minimize a loss function. This is done in order to maximize the likelihood of a desired outcome. A potent predictive model can be created using the gradient boosting classifier by combining many less effective learning models.

The following are the three fundamental components that make up gradient boosting:

Functioning at a Loss

The objective of the loss function is to determine, given the information at hand, how accurate the model's predictions are. This could be different depending on the specific nature of the problem at hand.

Weak Learner

The data are classified, but a weak learner does a poor job of it and makes many errors in the process. In most cases, these take the form of decision trees.

Additive Model

The addition of the trees is carried out in this manner in a manner that is successive, iterative, and gradual. With each iteration, you should be moving closer and closer to the final version of your model.

3.7.2 XGBRF classifier

An ensemble learning approach known as eXtreme Gradient Boosting Random Forest, or XGBRF for short, was developed by combining the benefits that are offered by the XGBoost and Random Forest algorithms. It is a variant of the XGBoost method that, in the process of building each tree, employs random subsets of the features, functioning in a manner analogous to that of the Random Forests algorithm. By taking this method, overfitting can be reduced, and the model's performance in terms of generalization can be improved. In addition, XGBRF incorporates a degree of randomness into the process of finding splits, which has the potential to substantially improve the performance of the model. In general, XGBRF is a robust machine learning technique that is commonly utilized for a variety of classification and regression related projects.

How this algorithm works?

The XGBoost and Random Forest learning algorithms are combined in the XGBRF algorithm to provide a more powerful learning tool. It makes use of gradient boosting, just like XGBoost does, in order to fit an ensemble of trees to the data. The technique known as "gradient boosting" involves adding trees to an existing ensemble in an iterative fashion, with each new tree making an effort to improve upon the performance of the trees that came before it. However, in contrast to conventional gradient boosting, XGBRF constructs each tree based on a random selection of the features rather than a predetermined order, much in the manner that Random Forests operate. By taking this method, overfitting may be reduced, and the model's performance in terms of generalization can be improved.

The data are initially used to tailor an initial tree to the algorithm's specifications. In each succeeding iteration, a new tree is fitted to the negative gradient of the loss function, in comparison to the collection of trees that came before it. After that, the new tree is brought into the mix, and the process is continued in this manner until a predetermined termination point is reached.

XGBRF additionally incorporates a randomness into the process of split discovery, which has the potential to significantly improve the performance of the model. When determining which attributes to take into account at each node of the tree in order to obtain the optimal split, the algorithm chooses those features at random. This strategy assists in de-correlating the trees that are a part of the ensemble, which can lead to enhanced overall performance.

In general, XGBRF is a robust machine learning technique that is commonly utilized for a variety of classification and regression related projects. It does this by combining the advantages of two different machine learning algorithms—the XGBoost and the Random Forest—to produce a model that is more resilient and generalizable than either approach on its own.

3.7.3 SVM

Support Vector Machine, or SVM for short, is a well-known supervised learning technique that may be applied to classification and regression problems. Finding a decision boundary, also known as a hyperplane, that categorizes the data into distinct groups is the primary objective of the support vector machine (SVM) technique. When selecting the hyperplane, care is taken to ensure that it optimizes the margin, which is defined as the distance between the decision boundary and the data points belonging to each class that are located closest to it. Support vectors are the terms used to refer to these nearest data points.

When the data contains a feature space with a high dimension or when the number of observations is less than the number of features, support vector machines (SVMs) are especially helpful. This is due to the fact that SVM builds a decision boundary by locating a linear combination of the input characteristics that maximizes the margin, which might make it less prone to overfitting than other methods.

Through the utilization of a kernel technique, SVMs are also capable of handling non-linearly separable data well. A function known as a kernel is one that takes the data that is being inputted and maps it into a higher-dimensional feature space in which the data may be separated linearly. Because of this, SVMs are able to simulate intricate decision boundaries. Kernels such as linear, polynomial, and radial basis function (RBF) kernels are examples of those that are used often.

The concept of regularization is also an essential part of SVM. Overfitting may be avoided with the use of a technique called regularization, which works by rewarding models with lower complexity. This may be accomplished by including a term in the optimization problem whose value is determined by the values of the model's parameters.

SVMs have found widespread usage in a variety of applications, including the categorization of images and texts, bioinformatics, and speech recognition, among others. SVM's benefits include the fact that it is effective and efficient even when dealing with high-dimensional and non-linearly separable data, that it provides a regularization parameter to prevent overfitting, and that it enables the use of a variety of kernel functions. However, it can be sensitive to the choice of kernel function and regularization parameter, and it can be computationally expensive for big datasets. Additionally, it can be sensitive to the choice of kernel function.

3.7.4 SGD Classifier

In the field of machine learning, the term "Stochastic Gradient Descent," abbreviated as "SGD," refers to a common optimization approach that is used for training models. This algorithm is particularly useful in situations in which the dataset is vast or the number of features is high. The idea behind SGD is to update the model's parameters based on the gradient of the loss function with respect to the parameters, while only using a small subset of the data (also referred to as a mini-batch) at each iteration of the process. This is done according to the gradient of the loss function with respect to the parameters. This methodology is referred to as stochastic, and it gets its name from the fact that the update at each stage is based on a random sampling of the data rather than the whole dataset.

SGD is frequently utilized in the training of linear models like logistic regression and linear regression, in addition to more complicated models like neural networks. An initial set of parameters is used as a foundation for the method, which then modifies those parameters in an iterative manner depending on the gradient of the loss function. A smaller portion of the data is chosen at each iteration in order to calculate the gradient, and this smaller subset of the data is the only one that is used. The parameters are then updated based on this gradient, using a learning rate that regulates the step size of the update. This process is repeated until all of the parameters have been updated.

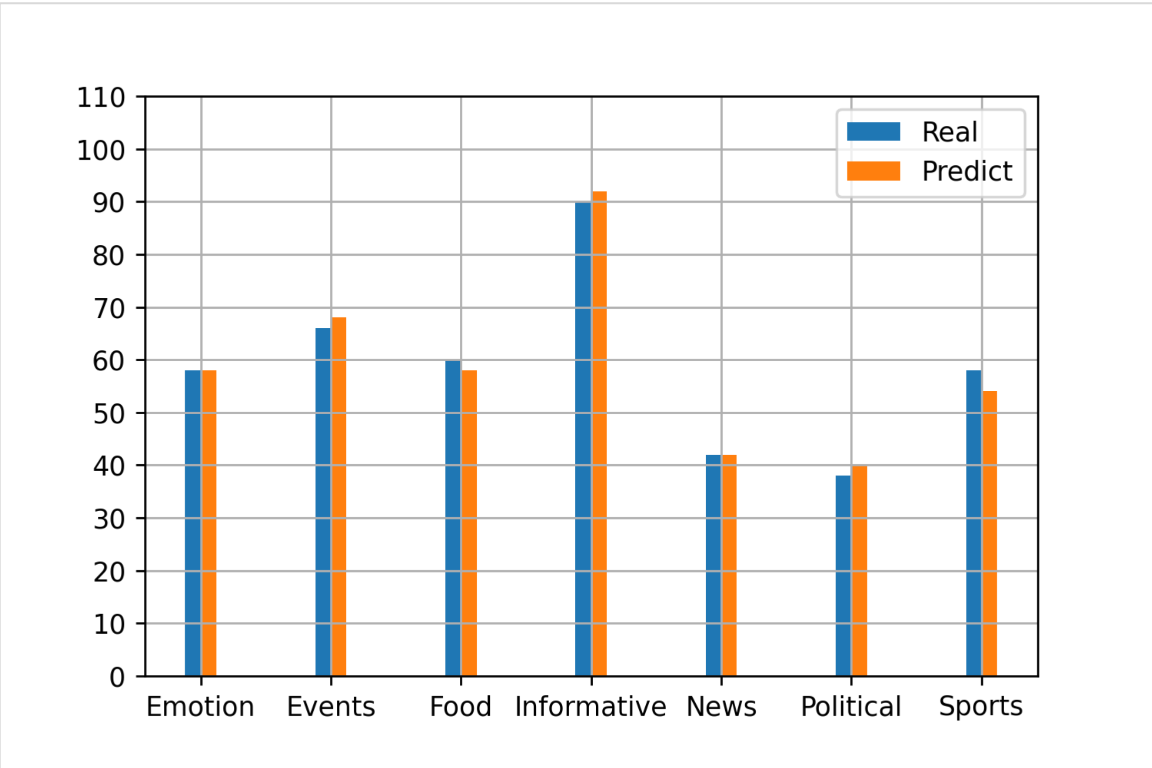
One of the primary benefits of SGD is that it can be utilized to train models on very big datasets that may not be able to fit in memory. This is one of the most significant advantages of SGD. Additionally, it enables online learning, during which the model may be changed in accordance with the arrival of fresh data. In addition to this, it is computationally efficient since during each iteration, it is only necessary to compute the gradient with regard to a limited portion of the data. This makes it possible to complete the calculation quickly.

However, employing SGD does come with a few limitations that you should be aware of. One of them is that the algorithm may converge to a solution that is less than optimum because of the unpredictability in the selection of the mini-batch. It may also be sensitive to the learning rate that the user chooses, so they need to be careful not to overshoot the best answer or oscillate around it. The algorithm may also be sensitive to the presence of noise or outliers in the data it processes.

Techniques such as learning rate schedules, momentum, and adaptive learning rate approaches such as Adagrad, Adadelta, and Adam have been presented as potential solutions to the problems described above.

3.8 Evaluation

Figure 3.3 shows our project evaluation. New information was culled from social media and classified as real data for each category. Figure 3.3 displays the system's effectiveness in the emotion, event, food, informational, news, political, and sports domains. A total of 413 sentences are organized into 5 classes. Emotion is correctly identified in 58/58 images. According to the timeline of events, we can deduce that four slip-ups occurred. For example, there are only 90 records, despite the system's prediction of 93. As a result, there were three mistakes made. There are no mistakes in the news. Due to political reasons, an error has



occurred at the level 2. And there were five miscues in sports.

Figure 3.5: Comparison of real and predicted classes

CHAPTER 4

RESULT ANALYSIS

4.1 Introduction

The results portion of the research paper should endeavor to tell the findings without attempting to analyze or evaluate them and also serve as a guide for the discussion section. The analysis is revealed when the results are reported. The writer explains what was done with the data found in the analysis section. Knowing what the analysis included is crucial for writing the analysis part, but this does not imply that data is required. The analysis must be completed before writing the results section. The experimental results and statistical analysis of those results were the primary topics of this section.

4.2 Experimental Result

I evaluated the efficacy of our study using five separate approaches applied to data that had already been preprocessed. To easily understand and evaluate the efficacy of various algorithms, I compiled the accuracy value they offered into an accuracy Table 4.1. I used 30%, 40%, 50%, 60%, and 70% complete data sets as training samples for our accuracy analysis. The result of comparing these five approaches is very remarkable. My results showed that XGB, Decision Tree, and Random Forest all performed consistently and accurately. On the other hand, the results were very comparable across SVM and Logistic Regression. Any study performed with logistic regression is the same as using statistical methods, as stated by R. E. Wright [16]. According to Table 4.1's red rectangular border-box, the logistic regression approach had the best accuracy of the five algorithms, at 81.23 percent, while using 70 percent training data and 30 percent testing data. Each yellow-shaded column shows the algorithms' best achievable accuracy at a given level of data usage. In terms of efficiency, the SVM method is our runner-up. In Machine Learning and Pattern Recognition, V. Vapnic [17] shown that the Support Vector Machine (SVM) achieved superior results due to the decrease of measured data. In our trial, SVM achieved a 79% by 40% data usage rate. P. O. Gislason et al. [18] claim that Random Forests can handle multi-source classification and high-dimensional data through the management of a large number of trees. Accuracy of 71.43 percent was achieved by Random Forest on 30 percent of test data. To achieve 72.92% accuracy, the naive bayes algorithm used just 30% of the data for testing, while the remaining 70% was used for training. When given a target accuracy of 70%, the XGB Classifier algorithm was able to accomplish it at a 30% data use rate.

Test Data usage rate	Algorithms			
	GB	*XGBRF*	*SVM*	*SGD*
20%	71.86%	45.08%	92.03 %	91.66 %
25%	71.49%	43.37 %	90.83 %	89.83 %
30%	72.01%	44.35 %	88.46 %	87.71 %
35%	70.82%	43.99 %	86.69 %	86.33 %

Table 4.1 ACCURACY TABLE

4.3 Classification report

In machine learning, a classification report is used as a measure of success. It demonstrates the reliability of your trained classification model by displaying its accuracy, recall, F1 Score, and support.

Target	Precision	Recall	F1-Score	Support
Emotion	0.90	1.00	0.95	9
Events	1.00	1.00	1.00	15
Food	1.00	1.00	1.00	7
Informative	0.95	1.00	0.97	39
News	1.00	1.00	1.00	11
Political	1.00	1.00	1.00	7
Sports	1.00	0.77	0.87	13
Accuracy			0.97	101
Macro AVG	0.98	0.97	0.97	101
Weighted AVG	0.97	0.97	0.97	101

Table 4.2: Classification report on support vector classifier

The accuracy of a classification algorithm may be evaluated with help from a Classification report. Find out what percentage of your forecasts came true and what percentage were off. In particular, I employ True Positives, False Positives, True Negatives, and False Negatives to make predictions about the metrics. Table 4.3 represents the classification report on support vector classifier. Emotion, events, foods, informative, news, political and sports are the main target of our work. In this part I tried to find out classification report on testing data. For support vector machine lowest precision rate is 0.90 achieved by emotion. The highest precision rate is 1, achieved by events, food, news, politicalm sports. For recall the lowest value is 0.77 achieved by sports. The highest f1 score is 1 achieved by events food news, political. The lowest f1 score is 0.87 achieved by sports.

Figure 4.1 represents the comparison between highest and lowest score of Support vector machine algorithm. We achieved this graph from classification report. From this graph we can see that the highest accuracy is 98 and lowest is 97 percent. Highest f1 score is 100 and lowest is 87 percent. Highest recall is 100 and lowest is 77. Highest precision is 100 and lowest is 90.

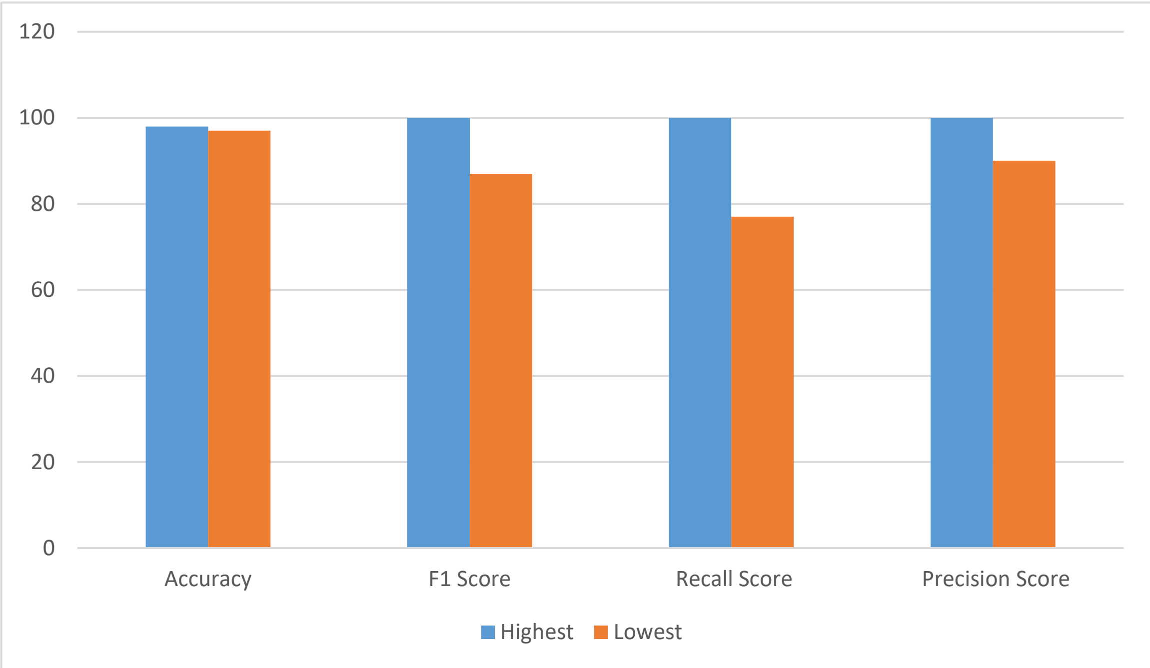


Figure 4.1: Comparison of highest and lowest score of Support Vector Classifier

Target	Precision	Recall	F1-Score	Support
Emotion	0.69	1.00	0.82	9
Events	1.00	1.00	1.00	15
Food	1.00	1.00	1.00	7
Informative	0.97	0.90	0.93	39
News	1.00	1.00	1.00	11
Political	1.00	1.00	1.00	7
Sports	1.00	0.92	0.96	13
Accuracy			0.95	101

Macro AVG	0.95	0.97	0.96	101
Weighted AVG	0.96	0.95	0.95	101

Table 4.3: Classification report on GradientBoosting classifier

A Classification report can be used to assess a classification algorithm's performance. Evaluate how accurate your predictions were and where you went wrong. In order to produce accurate forecasts on the metrics, we use a variety of techniques, including True Positives, False Positives, True Negatives, and False Negatives. You can see the support vector classifier classification report in Table 4.4. Most of what we do focuses on aspects of human experience, including emotions, experiences, foods, information, news, politics, and sports. Here we sought a report on the testing data's categorization. Emotion allows for the lowest accuracy rate of 0.69 in GradientBoosting. There are six categories where accuracy is at a perfect 1: events, cuisine, news, politics, and sports. The recall rate was almost 90.0 percent for informative. Events, food news, and politics have the highest f1 score possible, at 1. Feelings are the only ones to get a f1 score below 0.82.

The Gradient Boosting Classifier algorithm's top and lowest scores are contrasted in Figure 4.2. This graph was created using a categorization report. This graph shows that 97 percent accuracy is the highest and 95 percent accuracy is the lowest. The highest f1 score possible is 100, while the lowest is 82%. The recall ranges from 92 to 100. 69 is the lowest precision and 100 is the highest.

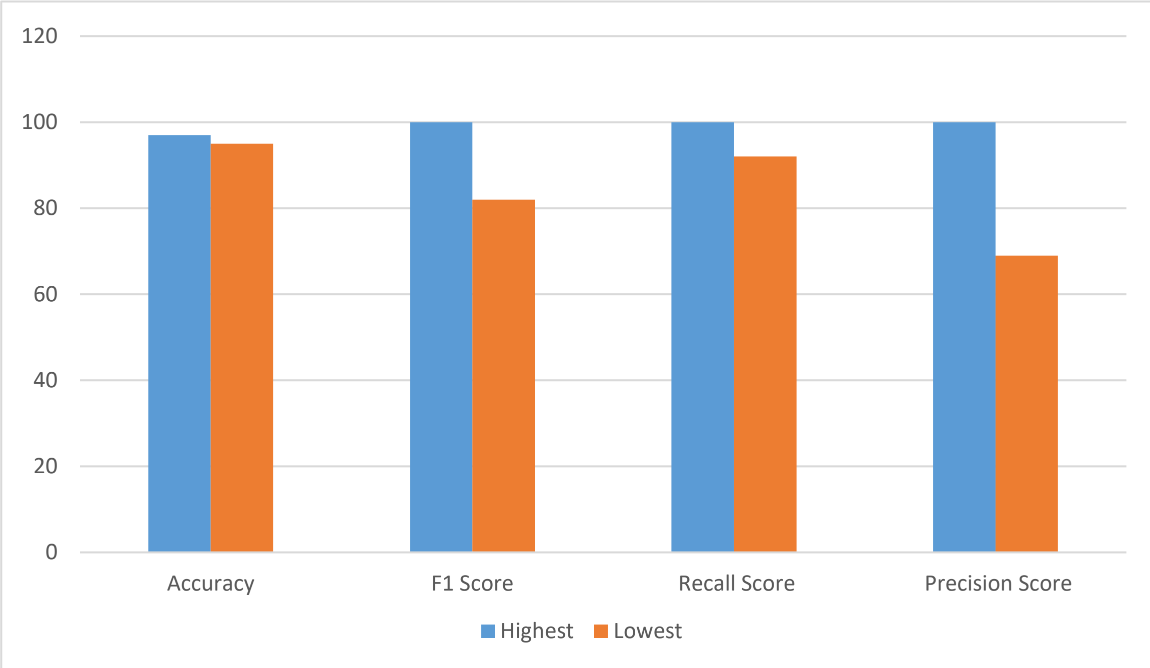


Figure 4.2: Comparison of highest and lowest score of Gradient Boosting Classifier

Target	Precision	Recall	F1-Score	Support
Emotion	0.25	0.22	0.24	9
Events	1.00	0.93	0.97	15
Food	1.00	0.57	0.73	7
Informative	0.63	0.87	0.73	39
News	0.86	0.55	0.67	11
Political	0.80	0.57	0.67	7
Sports	1.00	0.69	0.82	13

Accuracy			0.72	101
Macro AVG	0.79	0.63	0.69	101
Weighted AVG	0.76	0.72	0.72	101

Table 4.4: Classification report on XGBRF classifier

A Classification report can be used to assess a classification algorithm's performance. Evaluate how accurate your predictions were and where you went wrong. In order to produce accurate forecasts on the metrics, we use a variety of techniques, including True Positives, False Positives, True Negatives, and False Negatives. The results of the support vector classifier classification are shown in Table 4.5. Most of what we do focuses on aspects of human experience, including emotions, experiences, foods, information, news, politics, and sports. Here we sought a report on the testing data's categorization. In the context of a support vector machine, the lowest achievable rate of accuracy is 0.25. When it comes to pinpoint accuracy, only events, cuisine, news, politics, and sports can compete at a rate of 1. Emotion might provide the lowest rating of 0.22 for recall. Events with a f1 score of 0.97 are the best possible.

The greatest and lowest scores of the XGBRF algorithm are compared in Figure 4.3. Using a classification report, we created this graph. We can observe from this graph that 72 percent accuracy is the highest and 63 percent is the lowest. The maximum f1 score is 97, while the lowest is 24%. Lowest recall is 77 and highest recall is 93. the lowest precision is100, while the highest is 25.

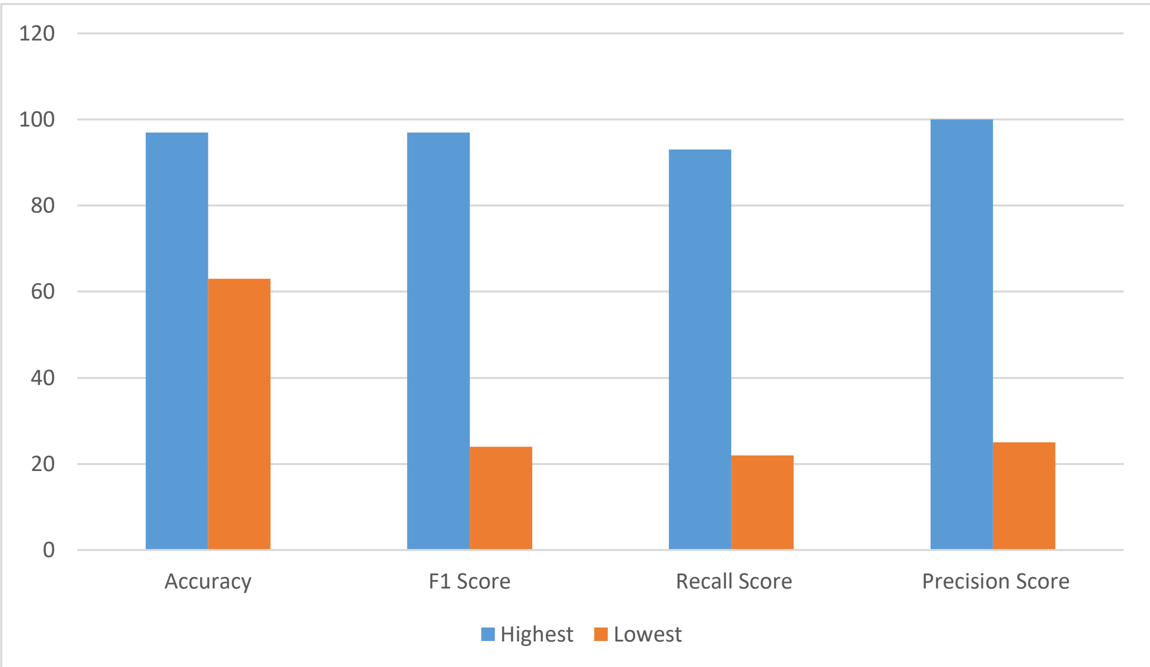


Figure 4.3: Comparison of highest and lowest score of XGBRF Classifier

Target	Precision	Recall	F1-Score	Support
Emotion	1.00	1.00	1.00	9
Events	0.88	1.00	0.94	15
Food	0.78	1.00	0.88	7
Informative	0.97	0.90	0.93	39
News	1.00	1.00	1.00	11
Political	1.00	1.00	1.00	7
Sports	1.00	0.92	0.96	13
Accuracy			0.95	101
Macro AVG	0.95	0.97	0.96	101
Weighted AVG	0.96	0.95	0.95	101

Table 4.5: Classification report on SGD classifier

A Classification report can be used to assess a classification algorithm's performance. Evaluate how accurate your predictions were and where you went wrong. In order to produce accurate forecasts on the metrics, we use a variety of techniques, including True Positives, False Positives, True Negatives, and False Negatives. Data on support vector classifier classification is shown in Table 4.6. Most of what we do focuses on aspects of human experience, including emotions, experiences, foods, information, news, politics, and sports. Here we sought a report on the testing data's categorization. Food has the lowest accuracy rate of any category in the SGD classifier, at 0.78. The news, politics, and sports ca theories all have an accuracy rate of 1. For remember the lowest value is 0.90 attained by informative. An emotional or political response receives a perfect f1 score of 1. Food has the lowest f1 score, at 0.88.

The greatest and lowest scores of the SGD algorithm are compared in Figure 4.4. Using a classification report, we created this graph. We can observe from this graph that 97 percent accuracy is the highest and 95 percent is the lowest. The maximum f1 score is 100, while the lowest is 88%. Lowest recall is 90 and highest recall is 100. the lowest precision is 8, while the highest is 100.

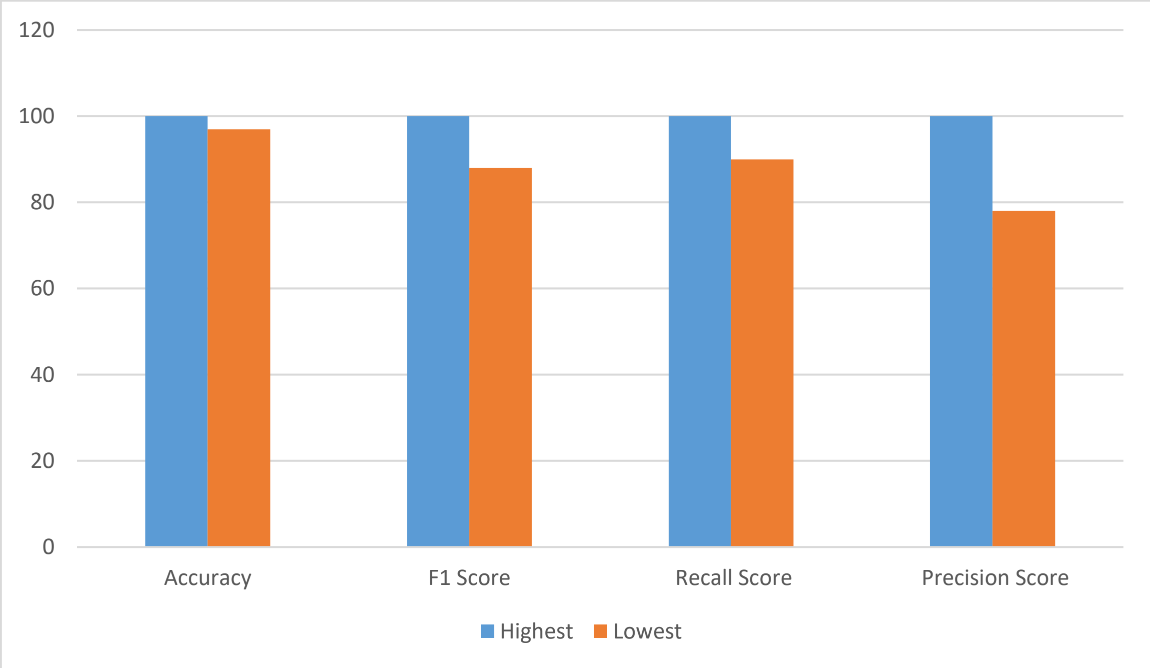


Figure 4.4: Comparison of highest and lowest score of SGD Classifier

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Facebook now has close to a quarter of the world's population as active users. This website is utilized by close to 80 percentages of all internet users in the Bangladesh region. Because they are sustained by the interactions that take place between users, social networks gain strength as their memberships expand. Because of the internet, people who hold unpopular opinions may now realize that they are not the only ones who feel this way. And when these individuals connect with one another through social media, they have the ability to do things, such as develop memes, publications, and entire online worlds that strengthen their worldview, and then burst into the mainstream.

The exposure of social, ethical, environmental, and political problems would be significantly reduced if social media did not exist. The balance of power has moved from the hands of a select few to those of the general populace as a result of increased visibility of the issues. Teenagers feel the pressure to conform to their peers' expectations, excel socially, and outdo their peers. Even before the widespread adoption of social media, this procedure was notoriously difficult. When you add platforms like as Facebook, Twitter, Snapchat, and Instagram to the mix, you immediately have a situation in which adolescents are forced to feel the pressure to mature prematurely in an online environment.

As a result of the conversation that has just taken place, I can see that utilizing social media may have led to both positive and negative outcomes. If the material of social media is organized in an appropriate manner, then it will be highly beneficial for society.

5.2 Impact on Environment

Changes in the natural or built environment that are the direct result of an activity and that can have a negative effect on the air, land, water, fish, and animals as well as the people who live in the ecosystem are referred to as environmental impacts. My project does not affect any of them.

5.3 Ethical Aspects

When doing research on human subjects, it is important to limit risks and harms while maximizing benefits; to respect human dignity, privacy, and autonomy; to take extra measures with vulnerable groups; and to make an effort to properly distribute the benefits and burdens of study. Everything that was covered earlier was accounted for in the work that I did.

5.4 Sustainability Plan

I claim that a project is sustainable when problems are resolved, new requirements are satisfied, maintenance in the future is eased, and the project can adapt to the environment in which it is located. The long-term viability of our initiative is wholly dependent on the dataset. If I am successful in gathering all of the Bangla data from social media, then it is safe to say that this project will have a very long lifespan.

CHAPTER 6

SUMMARY, CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

There is little doubt that the field of Natural Language Processing is active, with much of the research focusing on the English language. This field of research has recently expanded as a result of the seismic shift in how we utilize computers brought on by the cumulative effects of a wide variety of works. The research that comes out of these kind of projects has some quite astounding practical applications. However, the lack of similar linguistic research on the Bangla language is also troubling. But I expect that many academics from other countries have already begun investigating this area. As part of my study, I use a number of different methods to my Bangla posts in order to classify them.

6.2 Conclusion

My research involved developing a system to categorize social media posts by their written content. The categorization efficiency of five different machine learning methods was compared while being applied to Bangla status updates. For boosting, I used the GradientBoostingClassifier and the XGBoost Random Forest Classifier; for traditional machine learning, I used the Support Vector Classifier and the Stochastic Gradient Descent (SGD). My model successfully categorized the posts as Food, Emotion, Event, Sports, and News using the proposed strategy. The limited number of available tags and the ambiguity of some social media posts are our major limitations. The other issue is that my method can't determine what kind of post it is if the content is shown as an image.

6.3 Recommendations

Here are some well-regarded suggestions for that.

- The results of this study may be enhanced if a better data set were to be compiled.
- This might be because more information is included in the dataset, which could improve the outcome.

6.4 Future Work

The next steps in developing this study will involve:

❑ Creating secondary categories to further narrow down search results. To do this, more data will be needed.

❑ To this purpose, we want to develop a smart system that can automatically classify articles or blogs, as well as classify textual pictures, for use across a wide range of blog sites and social networks.

❑ Adding extra folders will make this project easier to manage.

REFERENCES

[1] J. Brown. (2018). What the science suggests. Available: https://www.bbc.com/future/article/20180104-is-social-media-bad-for-you-the-evidence-and-the-unknowns

[2] D. Chaffey. (17 april 2020). Global social media research summary 2020 Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

[3] S. KEMP. (17 FEBRUARY 2020). DIGITAL 2020: BANGLADESH. Available: https://datareportal.com/reports/digital-2020-bangaldesh

[4] S. Limon, M. Ahmad, and F. N. Mishu, "Bangla News Classification Using Machine Learning," 2018.

[5] M. S. Salayhin, "Development of a Bangla news classification system," 2019.

[6] U. Suleymanov, S. Rustamov, M. Zulfugarov, O. Orujov, N. Musayev, and A. Alizade, "Empirical Study of Online News Classification Using Machine Learning Approaches," in 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1-6: IEEE.

[7] L. Nahar, Z. Sultana, N. Jahan, and U. Jannat, "Filtering Bengali Political and Sports News of Social Media from Textual Information," in 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1-6: IEEE.

[8] T. Islam, A. R. Bappy, T. Rahman, and M. S. Uddin, "Filtering political sentiment in social media from textual information," in 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 663-666: IEEE.

[9] M. M. Islam, A. K. M. Masum, M. G. Rabbani, R. Zannat, and M. Rahman, "Performance Measurement of Multiple Supervised Learning Algorithms for Bengali News Headline Sentiment Classification," in 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), 2019, pp. 235-239: IEEE.

[10] A. K. Mandal and R. J. a. p. a. Sen, "Supervised learning methods for bangla web document categorization," 2014.

[11] M. A. Shafin, Md. M. Hasan, Md. R. Alam, M. A. Mithu, A. U. Nur, and Md. O. Faruk, "Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language," in International Conference on Computer and Information Technology (ICCIT), DHAKA, Bangladesh, 2020

[12] M. H. Rahman, M. S. Islam, M. M. U. Jowel, M. M. Hasan and M.S. Latif, "Classification of Book Review Sentiment in Bangla Language Using NLP, Machine Learning and LSTM," in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021

[13] X. L. Xiaobing Sun "Empirical studies on the NLP techniques for source code data preprocessing."

[14] J. J. Webster and C. Kit, "Tokenization as the initial phase in NLP," in COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics, 1992.

APPENDIX

I had several challenges while doing the research, the first of which was determining the methodological strategy for our work. It was not typical job, and there had not previously been much work done in this sector. As a consequence, we couldn't obtain much aid from anyone. Another impediment was data processing, which proved to be a significant issue for me. Because there was no accessible source for a Bangla text pre-processing system, I created a corpus for data collecting. In addition, I began manually collecting data. Furthermore, categorizing various postings is a difficult task to do. I were able to do this after a lengthy period of hard labor.

Social Media Post Classification in Bangla Language: A Comparison Between Boosting and Traditional Machine Learning Algorithm

ORIGINALITY REPORT

24%	20%	12%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	Md. Rejaul Alam, Afsana Akter, Minhajul Abedin Shafin, Md. Mehedi Hasan, Antara Mahmud. "Social Media Content Categorization Using Supervised Based Machine Learning Methods and Natural Language Processing in Bangla Language", 2020 11th International Conference on Electrical and Computer Engineering (ICECE), 2020 Publication	6%
3	Submitted to Daffodil International University Student Paper	4%
4	Mohd. Istiaq Hossain Junaid, Faisal Hossain, Udyan Saha Upal, Anjana Tameem, Abul Kashim, Ahmed Fahmin. "Bangla Food Review Sentimental Analysis using Machine Learning", 2022 IEEE 12th Annual Computing	1%