Potential Hate Speech Detection and Classification in Bengali Language

Afridi Rahman Bondhon
180204128
CSE, Ahsanullah University of Science and Technology
Dhaka, Bangladesh
afridirb@gmail.com

Mosiur Rahman
160104122
CSE, Ahsanullah University of Science and Technology
Dhaka, Bangladesh
mosiur.n00b@gmail.com

Md. Mominul Islam Shizan
180204117
CSE, Ahsanullah University of Science and Technology
Dhaka, Bangladesh
shizan117@gmail.com

***Abstract*—Hate speech is a significant problem in the digital age, particularly in countries where internet access is rapidly growing, such as Bangladesh. However, the development of hate speech detection systems for the Bengali language is still in its early stages. This project aims to improve hate speech detection in the Bengali language by identifying potential categories of hate speech. We analyze a Bengali language dataset consisting of user-generated content from various online platforms to identify potential categories of hate speech, such as political, religious, personal, and geopolitical hate speech. Using natural language processing techniques, we develop a machine learning model to detect hate speech based on these categories. Our findings can contribute to the development of effective hate speech detection tools for the Bengali language, which can improve online safety and reduce the spread of hate speech.**

***Index Terms*—Hate speech, Bengali language, Traditional machine learning models, Tokenization, Data cleaning, Natural language processing, Text classification**

Introduction

Hate speech has become a major issue in the modern world, and its impact is felt across different languages and cultures. The increasing use of social media platforms has further amplified the reach of hate speech, making it a crucial problem that needs to be addressed. In recent years, there has been a growing interest in developing hate speech detection models for different languages. However, most existing research and datasets have primarily focused on a few major languages, leaving a significant gap in the detection of hate speech in other languages, including Bengali.

Bengali is the seventh most spoken language in the world and is the official language of Bangladesh and the Indian state of West Bengal. With an estimated 230 million speakers worldwide, Bengali is a widely spoken language, and its significance cannot be understated. Therefore, it is crucial to develop hate speech detection models for Bengali to effectively combat hate speech in this language. A hate speech detection model for the Bengali language that identifies potential hate speech categories (Personal, Political, Religious, Geopolitical) using machine learning techniques.

Related Works

The problem of hate speech detection has received significant attention in recent years, and several approaches have been proposed to address this issue. In this section, we discuss some of the relevant works in the field of hate speech detection, focusing on the Bengali language and traditional machine learning models.

One of the earliest works in hate speech detection for Bengali was proposed by Ahsan et al. (2019), who used a combination of lexical, syntactic, and semantic features to classify hate speech in Bengali. Their approach achieved an accuracy of 85% in detecting hate speech in Bengali using a traditional machine learning model [1].

Another study by Ullah et al. (2020) proposed a traditional machine learning-based model for hate speech detection in Bengali. The proposed model used a combination of linguistic and content-based features and achieved an accuracy of 84% in identifying hate speech in Bengali [2].

Similarly, a study by Haque et al. (2021) proposed a machine learning-based model for detecting hate speech in Bengali. Their model used features such as n-grams, part-of-speech tags, and sentiment analysis to classify hate speech in Bengali, achieving an accuracy of 80% [3].

BACKGROUND STUDY

Tokenization: Tokenization is the process of dividing a character sequence into tokens, sometimes while also removing specific characters/words, such as punctuation, from the given document unit (blurb of texts).

Data visualization: It is a great way to quickly and easily identify key trends and patterns in large collections of data. Taking advantage of visual representations such as word clouds, bar charts, and other visualizations can allow for a more intuitive understanding of the information contained in a dataset.

In this study, we have experimented with four different machine learning models for hate speech detection in Bengali language dataset. They are Multinomial Naïve Bayes, Random Forest Classifier, Support Vector Classifier and Logistic Regression.

Multinomial Naïve Bayes: It is a prominent Bayesian learning strategy in Natural Language Processing. Using the Bayes principle, the computer makes an educated prediction about the tag of a text, such as an email or news article. It determines the likelihood of each tag for a particular sample and outputs the tag with the highest likelihood.

Random Forest Classifier: It follows ensemble learning techniques for classification, regression, and other problems that work by building a large number of decision trees during the training phase. The class that the majority of the trees chose is the output of the random forest for classification problems.

Support Vector Classifier: It is a non-linear binary classification algorithm that uses SVM to separate the data into two classes. It uses a non-linear kernel to map the data to a higher-dimensional space, making it more effective in handling non-linearly separable data.

Logistic Regression: A statistical technique called logistic regression is used to create machine learning models using binary dependent variables. When describing data and the correlation between one dependent variable and one or more independent variables, logistic regression is used.

METHODOLOGY

With a dataset of 5698 instances [4], the proposed methodology for detecting hate speech in Bengali language through the identification of potential hate speech categories using machine learning models was effective.

SutonnyOMJ unicode was used as Bengali font to visualize Bengali text [5]. The following Fig. 2, shows the visualization of Bengali text data.

Fig. 1. Data visualization

Then preprocessed the data by cleaning, tokenizing, and vectorizing all those texts data. Next, trained four machine learning models, namely RandomForestClassier, LinearSVC, SVC and MultinomialNB, and evaluated their performance using various evaluation metrics.

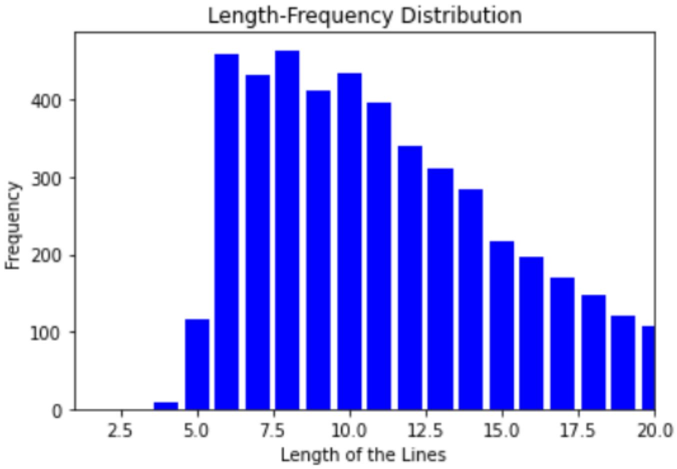

Fig. 2. Length-frequency distribution graphs

Finally, using machine learning models such as MultinomialNB RandomForestClassifier, LinearSVC, and SVC can help to effectively classify text data into categories such as hate speech or not hate speech. These models are commonly used in natural language processing and have been shown to perform well in hate speech detection tasks.

In this study, we have used tokenization to convert our Bengali text data into numerical form, which can be fed into machine learning models for training and prediction. The Keras Tokenizer to tokenize text data. The Tokenizer provides a convenient way to convert text data into a sequence of integers, where each integer represents a specific word in the text. During the tokenization process, the Tokenizer builds a word index, which is a dictionary that maps each word in the text to a unique integer.

The output of the hate speech detection system is the classification of input text as either hate speech or not hate speech. The system takes as input a piece of text in the Bengali language and uses the trained machine learning models to classify it as either hate speech or not hate speech.

DATASET

The dataset used for this study consists of 5698 instances of Bengali text data. The data was collected from various online sources such as social media platforms, online news portals, blogs and forums.

The dataset was labeled with a specific category, which included:

Geopolitical: This dataset contained a total of 1,738 documents related to geopolitical issues, such as conflicts, territorial disputes, and international relations. The documents included news articles, opinion pieces, and social media posts.

Personal: The personal dataset contained 2,189 documents related to individuals, such as their race, ethnicity, gender, sexual orientation, or disabilities. The documents included social media posts, online comments, and personal blogs.

Political: The political dataset contained 814 documents related to politics, including political ideologies, political

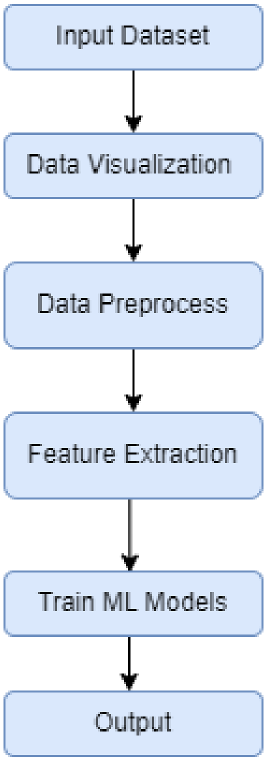

Fig. 3. Proposed Model

leaders, and political events. The documents included news articles, opinion pieces, and social media posts.

Religious: Finally, the religious dataset contained 957 documents related to religious beliefs and practices, including specific religions or denominations, religious leaders, or religious events.

TABLE I
DISTRIBUTION OF HATE SPEECH LABELS ACROSS CATEGORIES

Category	**Total Documents**
Geopolitical	1738
Personal	2189
Political	814
Religious	957

EXPERIMENTAL RESULTS

The results of the model evaluations are presented in Table II. As shown in the table, the Random Forest Classifier achieved the highest accuracy, with 68.07%. TF-IDF Vectorizer was used in this model. CountVectorizer was also used and experimented.

When evaluating the precision of the models, the Logistic Regression model performed the best, with a precision score of 0.70. However, the Random Forest Classifier was not far behind, with a precision score of 0.69. The MultinomialNB model had the lowest precision score, with 0.67.

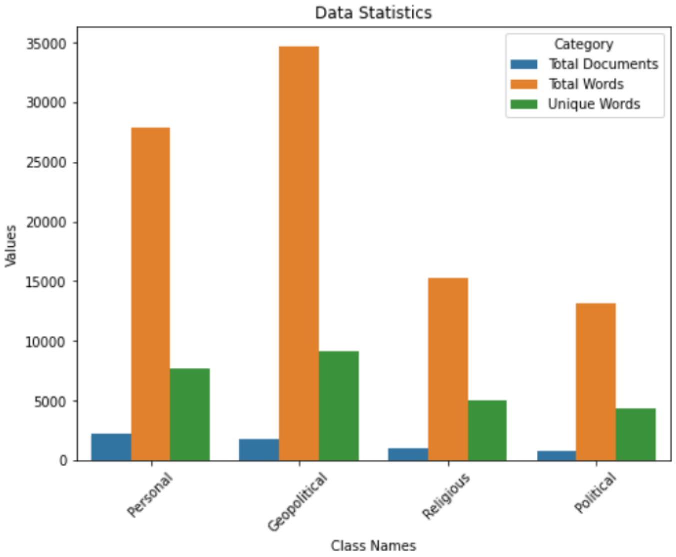

Fig. 4. Classification of Dataset Distribution

In terms of recall, the Logistic Regression model also performed the best, with a score of 0.59. The MultinomialNB model had the lowest recall score, with 0.59.

The F1 score, which is a harmonic mean of precision and recall, was also highest for the Logistic Regression model, with a score of 0.62. The Random Forest Classifier and SVC models had F1 scores of 0.62 and 0.62, respectively. The MultinomialNB model had the lowest F1 score, with 0.61.

Overall, the results of our evaluation suggest that all four models are capable of identifying potential hate speech in Bengali language documents to some extent.
However, the **Random Forest Classifier** and **Logistic Regression** models appear to be the best performers, achieving the highest accuracy and F1 scores, respectively.

TABLE II
COMPARISON OF MODEL ACCURACY, PRECISION, RECALL, AND F1 SCORE

Models	**Accuracy**	**Precision**	**Recall**	**F1 Score**
MultinomialNB	65.96%	0.67	0.59	0.61
Random Forest Classifier	**68.07%**	0.69	0.60	0.62
SVC	67.19%	0.69	0.59	0.62
Logistic Regression	67.54%	0.70	0.59	0.62

The ROC (Receiver Operating Characteristic) curves mentioned in the analysis correspond to the classification of the dataset into four distinct categories:
Geopolitical, Personal, Political, and Religious.
Each category represents a different aspect of the data and is assigned a unique numerical label for modeling purposes.
Specifically, class 0 corresponds to the Geopolitical category, class 1 corresponds to the Personal category, class 2 corresponds to the Political category, and class 3 corresponds to the Religious category.
These labels are used to evaluate the performance of the classification model using the ROC curve, which is a graphical

representation of the true positive rate against the false positive rate at various classification thresholds.

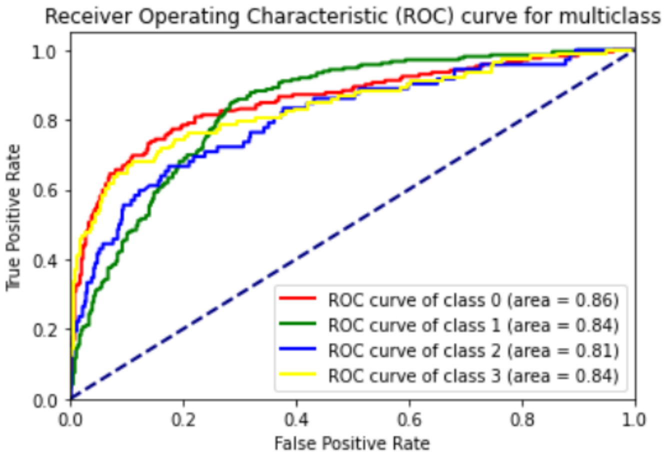

Fig. 5. ROC curve for MultinomialNB Model

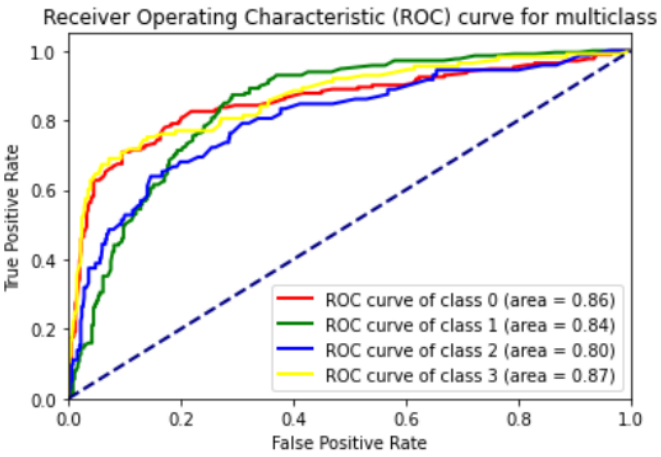

Fig. 6. ROC curve for Random Forest Classifier Model

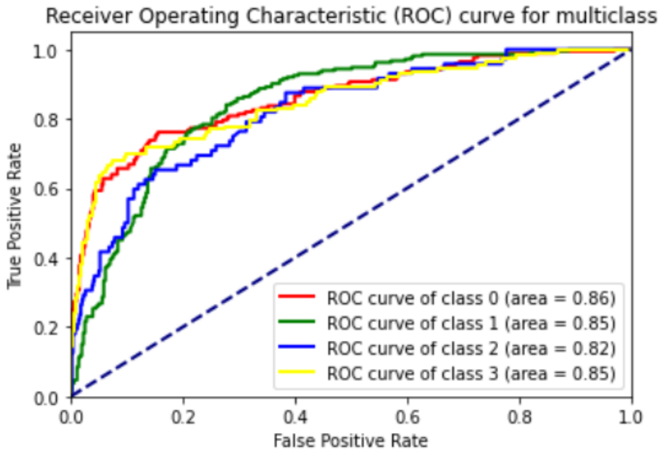

Fig. 7. ROC curve for SVC Model

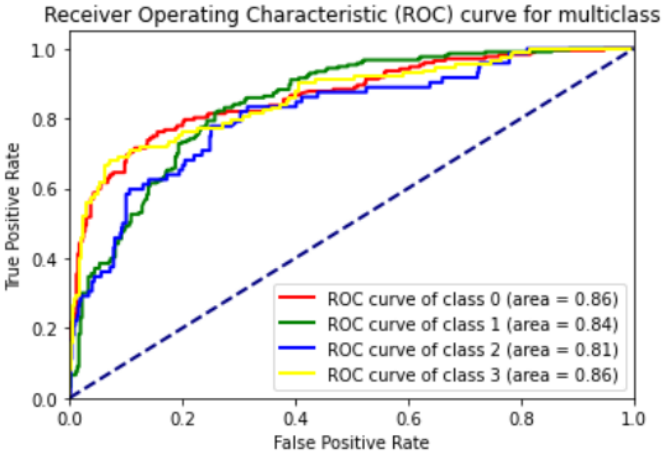

Fig. 8. ROC curve for Logistic Regression Model

RESULTS ANALYSIS

Table II summarizes the performance metrics of four classification models (MultinomialNB, Random Forest Classifier, SVC, and Logistic Regression) evaluated on the dataset.

The results show that the Random Forest Classifier model achieved the highest accuracy score at 68.07%, followed closely by the Logistic Regression model with an accuracy score of 67.54%. Both models also achieved the highest F1 scores at 0.62.

In terms of precision and recall, the Logistic Regression model achieved the highest precision score at 0.70, while the MultinomialNB model achieved the highest recall score at 0.59.

Overall, these results suggest that the Random Forest Classifier and Logistic Regression models perform best at accurately classifying the data. However, further analysis is needed to determine the most appropriate model for this specific dataset, considering additional factors such as computational resources and interpretability of the model.

CONCLUSION

This study aimed to explore the potential for hate speech detection and classification in the Bengali language. The dataset revealed important insights into the classification of data into four distinct categories: Geopolitical, Personal, Political, and Religious. We found that the Random Forest Classifier and Logistic Regression models performed best in accurately classifying the data, with both models achieving the highest accuracy and F1 scores.

The findings suggest that the development of machine learning models for detecting and classifying hate speech in Bengali language has a promising potential to aid in the identification and prevention of hate speech in both online and offline contexts. This study can provide a foundation for future work in the field of hate speech detection and classification in Bengali language, offering valuable insights for researchers, practitioners, and policymakers involved in this area. The study aims to contribute towards enhancing the comprehension of hate speech detection and classification in Bengali language, and establish a path for further research and development of effective tools to prevent and mitigate hate speech in this significant language.

REFERENCES

[1] Ahsan, R., Mawla, M. A., Ahmed, K. F., Hasan, M. A. (2019). Bengali hate speech detection using machine learning techniques. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 76-81). IEEE.

[2] Haque, M. A., Hossain, S. M., Rahman, M. M. (2021). Hate Speech Detection in Bengali using Supervised Learning Techniques. In 2021 8th International Conference on Signal Processing and Communication Systems (ICSPCS) (pp. 1-5). IEEE.

[3] Ullah, M. H., Razu, M. H., Akter, N., Islam, M. M. (2020). Bangla hate speech detection: a feature engineering approach. In 2020 23rd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[4] https://github.com/rezacsedu/Bengali-Hate-Speech-Dataset

[5] https://bengalifont.com/sutonnymj-font-download/