Identifying Hate Speech on Online Platform in Bangladesh Using Machine Learning

Zarif Khan, Md. Shariar Fahim, Abraham Kaikobad

Abstract

With the increasing prevalence of online communication in Bangladesh, the identification and mitigation of hate speech on digital platforms have become imperative. This study proposes a machine learning-based approach to detect hate speech in Bengali language text data on online platforms. The model employs natural language processing techniques and leverages a dataset specifically curated for the linguistic nuances of Bangladesh. Our methodology includes preprocessing steps, feature extraction, and the development of a classification model trained on labeled data. To enhance model performance, we explore the effectiveness of various machine learning algorithms, considering the unique linguistic characteristics and cultural context of hate speech in Bangladesh. The proposed system aims to contribute to the ongoing efforts to create a safer online environment by automating the detection of hate speech, facilitating prompt interventions, and fostering responsible digital discourse. The results demonstrate the model's efficacy in identifying hate speech on online platforms, offering a valuable tool for content moderation and community well-being in the digital landscape of Bangladesh.

1. Introduction

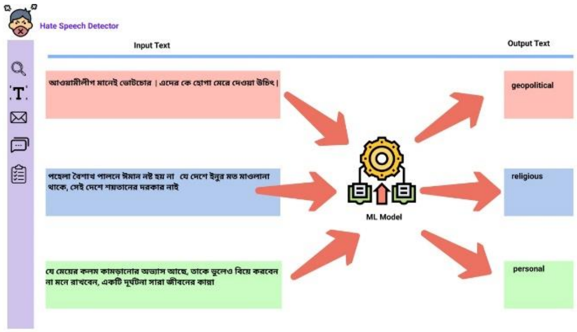
Social media has become an integral part of people's lives, providing an easy platform for expressing opinions and interacting with others. However, it has also become a breeding ground for harassment and hate, including sexism, racism, and political animosity. As of November 2017, Bangladesh had around 25-30 million Facebook users, with 72% being male and 38% female. Notably, Dhaka ranks second globally in terms of active Facebook users. Unfortunately, cyberbullying affects a significant portion of the population, with 73% of women and 49% of students in Bangladesh experiencing online harassment.

Machine learning can play a significant role in addressing the issue of hate speech and online harassment on social media platforms. Machine learning algorithms can be trained to automatically detect hate speech and offensive language. These algorithms can scan large volumes of content and flag potentially harmful posts, enabling faster and more efficient content moderation. It can analyze the context and sentiment of text-based content. By understanding the nuances of language, machine learning models can identify hate speech that might be disguised through sarcasm, slang, or coded language. Many researchers and organizations have actively worked on using machine learning techniques to detect hate speech and improve online content moderation. Extensive research has been conducted on abusive text detection in English, but there's a notable lack of research in the Bangla language. [2] In a dataset of nearly 12,000 instances from diverse social media platforms, SVM stood out as the top performer with the highest accuracy (88%) and F-score, surpassing other models. However, k-NN, random forest, and decision tree exhibited lower performance, while stemming adversely affected accuracy due to the limitations of the Bengali stemmer. [1] On the other research, utilizing a dataset of 6000 comments, including 2500 instances of hate speech, the research team employed algorithms such as Support Vector Machine (SVM), Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbor (KNN) to classify Bangla comments. Logistic Regression emerged as the top-performing technique, achieving an accuracy rate of 97.09%.

Bangla is the seventh most widely spoken native language globally, with approximately 205 million speakers, accounting for around 3.05% of the world's population. In Bangladesh, there are about 81.7 million internet users, among whom 30 million are active on social media, with 28 million accessing social platforms via mobile phones. Around 42 million Facebook users communicate in Bangla, constituting nearly 1.9% of the total Facebook user base. The use of the Bangla language is also growing steadily across various social media platforms. And also, most of the users of Bangladesh use ‘Banglish’ which refers to the mixing of Bengali (Bangla) language with English. In this paper, the following contributions are made:

- A dataset of more or less 10,000 Bengali comments were gathered. The comments were collected from various sources like Facebook and Instagram posts and comments and the datasets of other papers etc. These comments were then classified into five categories: personal, political, religious, geopolitical and gender abusive.

- In the dataset, there are five columns. Four columns are input and the label column is the output. There are two new features which are Bangla hate words and Banglish hate words so that the model will have a nice idea about the hate words.
- Also, we are launching a website for detecting hate speech detection which will find the hate word.



আওয়ামীলীগ মানেই ভোটচোর | এদের কে ঝোপা মেরে দেওয়া উচিৎ |

geopolitical

পহেলা বৈশাখ পালনে ঈমান নষ্ট হয় না যে দেশে ইনুর মত মাওলানা থাকে, সেই দেশে শয়তানের দরকার নাই

religious

যে মেয়ের কলম কামড়ানোর অভ্যাস আছে, তাকে ভুলেও বিয়ে করবেন না মনে রাখবেন, একটি দুর্ঘটনা সারা জীবনের কান্না

personal

Figure: This figure illustrates analyzing an input sentence to achieve an exact classification by ML model.

2. Related work

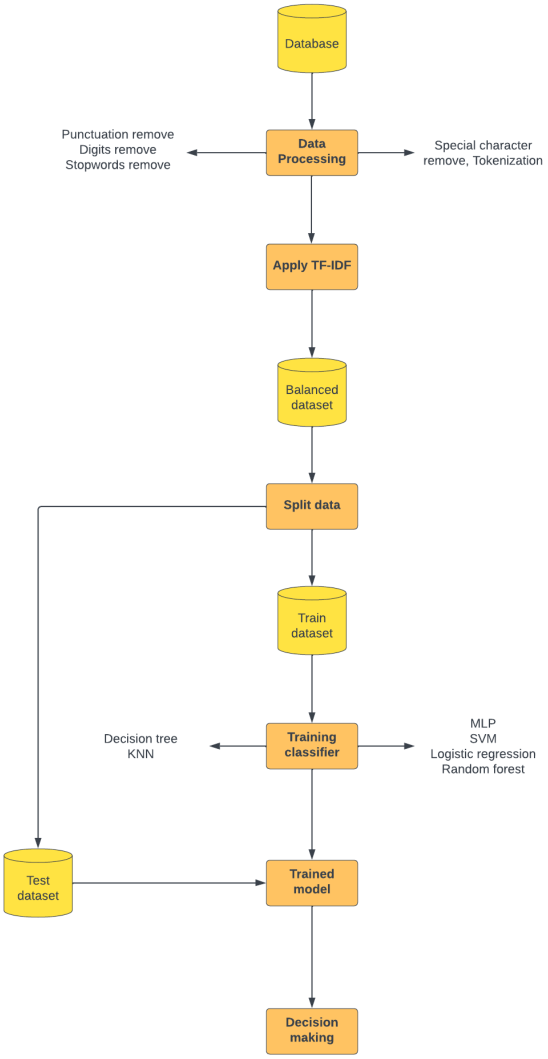
Many researchers have worked on this very topic of hate speech with other methods. As they have worked on English language hate speech but the methods are applied on Bangla language too. Arum Sucia Saksesi et al. used a recurrent neural network to detect hate speech. They made a combination of LSTM with RNN. To find the result of the LSTM hidden layer they have used SoftMax regularization. They partitioned the data with different ratios at different times.

N.D.Gitari et al. have used lexicon in their work. They got average results with the lexicon model. Their F-score was 70.83. Ricardo Martins et al. have used emotional words to classify hate speech. They have obtained an accuracy of 80.56% with a support vector machine. Nur Indah Pratiwi et al. have used Fast text approach to detect hate speech in Instagram comments. They have used word n-gram and char n-grams.

And also, some researchers have worked on Bangla hate speech as well. Asfi Hossain Choudury et al. have implemented NLP libraries such as NLTK (Natural Language Toolkit), and TF-IDF, as well as separate ML toolkits such as learning Sci-kit, Numpy, Matplotlib, and Pandas to achieve these objectives and collected a dataset of more than 6000 comments.

Shovon Ahammed et al. have built a new dataset to find malice in Bangla Language that contains hate speech on different categories such as religion, community, gender, race. They have used count vectorizer and TF-IDF. Tanvirul Islam et al. have a dataset of almost 12,000 instances has been collected from different social media. They have used tokenization technique.

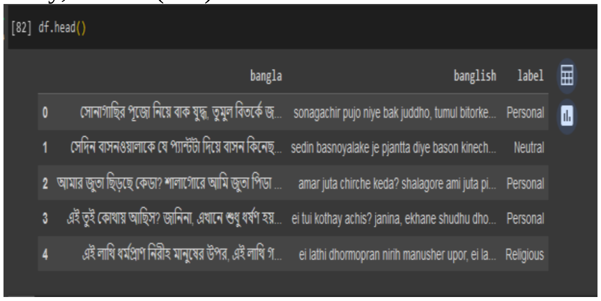
3. Methodology



Dataset Details:

We have grabbed more than 20,000 datasets from social media comments and other hate speech papers. We have only found bangla datasets in the papers. For our work we have to added new column which is ‘Banglish’. In our dataset, we have 3 columns and 12131 rows. The first two columns are ‘Bangla’ and ‘Banglish’ and they are both inputs. They are both numerical featuresas they can’t be categorized. And the output column is ‘Label’ and it is the categorical feature as it can be categorized into five labels. In our 12131 datasets, we have divided them into 5 categories

which are personal (4225), geopolitical (3761), religious (1714), political (1596) and
lastly, neutral (835).



```
[82] df.head()
```

	bangla	banglish	label
0	সোনাগাছির পুজো নিয়ে বাক যুদ্ধ, তুমুল বিতর্কে জ...	sonagachir pujo niye bak juddho, tumul bitorke...	Personal
1	সেদিন বাসনওয়ালাকে যে প্যান্টটা দিয়ে বাসন কিনেছ...	sedin basnoyalake je pjantta diye bason kinech...	Neutral
2	আমার জুতা ছিড়ছে কেডা? শালাগোরে আমি জুতা পিডা ...	amar juta chirche keda? shalagore ami juta pi...	Personal
3	এই তুই কোথায় আছিস? জানিনা, এখানে শুধু ধর্ষণ হয়...	ei tui kothay achis? janina, ekhane shudhu dho...	Personal
4	এই লাথি ধর্মপ্রাণ নিরীহ মানুষের উপর, এই লাথি গ...	ei lathi dhormopran nirih manusher upor, ei la...	Religious

EDA is the initial phase of data analysis where the main goal is to summarize the key characteristics, patterns, and relationships in a dataset. It involves utilizing descriptive and graphical statistical techniques to gain insights into the data. EDA helps in understanding the underlying structure of the data, identifying outliers, and formulating hypotheses for further analysis. Fig1 and Fig2 we can cleary get idea about the dataset in imblanced.and aslo we don’t have any null datas on dataset.So we process full dataset for data preprocessing.

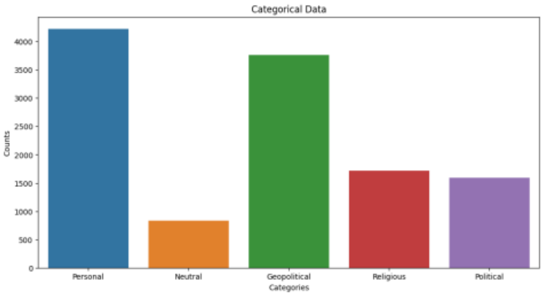


Fig1: Categorical Data visualization

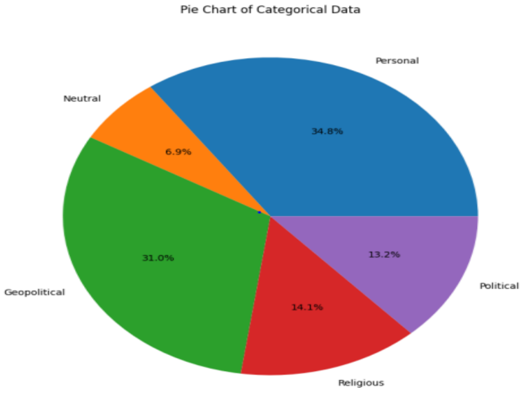


Fig2: Pie chart of categorical Data

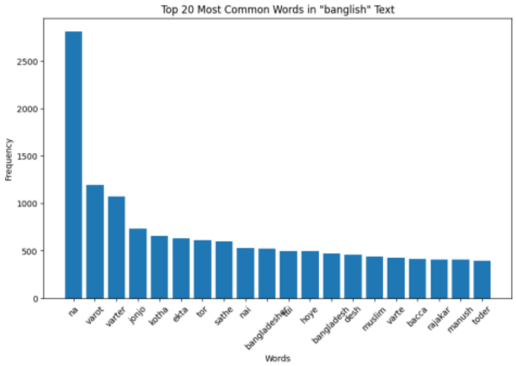


Fig3: Most Frequents words in Banglish Text

Pre-Processing:

For data processing we have first separated our data by bangla and banglish. Then we have created separately Bangla and Banglish text cleaning.

Bangla Text Cleaning:

The function clean_bangla_text plays a vital role in preparing Bangla text for natural language processing tasks. It begins by breaking down the input text into individual words through tokenization, and ready for the operations. Rigorous removal of stopwords follows, eliminating commonly used but less informative words, enhancing the text's relevance. The function then filters out punctuation marks and digits, focusing on the main context of text. The cleaned words are intelligently reassembled into a coherent string, ensuring consistent spacing. Finally, excess whitespace is trimmed, yielding a refined and preprocessed version of the original Bangla text. This meticulous preprocessing, involving tokenization, stopword removal, and character filtering, contributes to creating a more meaningful and streamlined representation of Bangla text, poised for effective utilization in various natural language processing applications.

Banglish Text Cleaning:

The function clean_banglish_text is designed to preprocess Banglish text, a combination of Bengali and English, using a set of transformations. Initially, the function removes punctuation and digits from the input text, ensuring that only alphabetic characters are retained. Following this, it reads a list of ignored words from an external file, representing terms that should be excluded during processing. The text is then tokenized into words, and a case-insensitive comparison is made to eliminate words present in the list of ignored terms. The resulting processed text consists of lowercase tokens, stripped of unwanted characters and irrelevant words. Finally, the function returns the cleaned

Banglish text as a single string, ready for subsequent natural language processing tasks. This preprocessing aids in enhancing the quality of the text for tasks such as sentiment analysis or classification, where irrelevant symbols and words can impact the accuracy of the analysis.

Feature Engineering:

Feature engineering involves transforming raw data into a format that enhances a machine learning model's performance. It includes creating new features, selecting relevant ones, and optimizing existing features to improve the model's ability to capture patterns.

TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF is a numerical statistic that reflects the importance of a word in a document relative to a collection of documents. It is widely used in natural language processing to represent the significance of words in a text corpus.

TF-IDF equation:

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

$$tf_{i,j} = \text{number of occurrences of } i \text{ in } j$$
$$df_i = \text{number of documents containing } i$$
$$N = \text{total number of documents}$$

Data Balancing:

Data balancing techniques address class imbalance in machine learning datasets, where some classes have significantly fewer instances than others. This is crucial for preventing models from being biased towards the majority class. Under Sampling: Under sampling involves reducing the number of instances in the majority class to match the minority class, equalizing class distribution and addressing the imbalance issue.

Personal	Neutral	Political	Geo Political	Religious
835	835	835	835	835

Oversampling: SMOTE is a technique used to address class imbalance in a dataset, particularly when the minority class is underrepresented. It works by generating synthetic samples in the feature space of the minority class. The method involves selecting a data point from the minority class and creating synthetic instances along the line segments connecting it to its nearest neighbors. This process enhances the representation of the minority class, providing a more balanced distribution. SMOTE is widely used to improve the performance of machine learning models, especially in scenarios where imbalanced classes can lead to biased predictions.

Personal	Neutral	Political	Geo Political	Religious
4225	4225	4225	4225	4225

Tomek Links: Tomek Links are pairs of instances from different classes that are close to each other. Removing these links can enhance the separation between classes and improve the performance of a machine learning model.

Personal	Neutral	Political	Geo Political	Religious
4139	835	1570	3738	1693

From these 3 techniques, we checked 3 techniques on our dataset. But we get best results on oversampling(smote) method.

Model:

So far, we used 10 models for our system to get precise and accurate prediction. Here we are showing only 5 best models according our system.

Support Vector Machine (SVM): SVM is a powerful supervised machine learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates different classes in the feature space.

Random Forest: Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes for classification tasks or the average prediction for regression tasks. It enhances accuracy and controls overfitting.

Logistic Regression: Despite its name, logistic regression is a linear model for binary classification that predicts the probability of an instance belonging to a particular class. It is widely used for its simplicity, interpretability, and effectiveness.

Decision Tree: Decision trees are versatile models used for both classification and regression tasks. They recursively split the data based on feature conditions to form a tree-like structure, making them intuitive and easy to understand.

MLP (Multilayer Perceptron): MLP is a type of artificial neural network with multiple layers of nodes, consisting of an input layer, hidden layers, and an output layer. It is particularly effective for complex tasks such as image recognition and natural language processing.

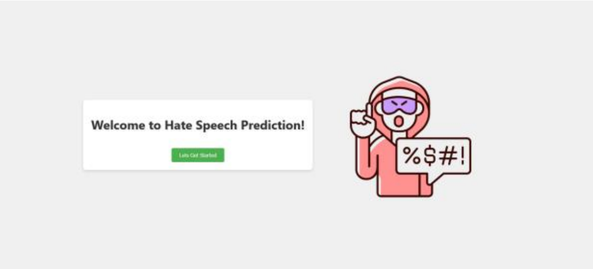
We have used several numbers of models including Naïve bayes, SVM, Random Forest, Logistic regression, Decision tree, KNN etc. We tested the dataset into two ways. Firstly, we tested the imbalanced dataset where Random Forest has done very good with the accuracy of 82%. SVM and MLP models have done well too with 81%. KNN has done very poor with 48%. And in the balanced dataset, MLP model has performed the best with 92% accuracy. RF and SVM models have performed well too with the 91% and 90%. KNN has done poor with 61% accuracy.

App:

Our website is designed to predict the type of hate speech in user-generated content. The user interface is intuitive and straightforward, ensuring ease of use for visitors.

Homepage:

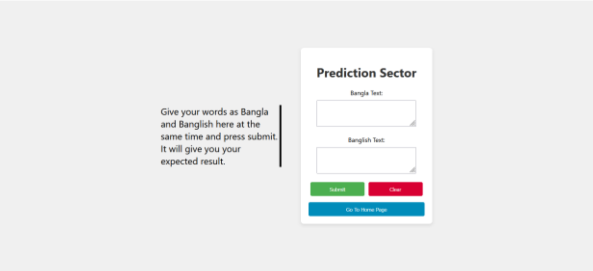
Upon landing on the homepage, users are greeted with a clean and simple design. The focal point is a prominently displayed button that directs users to the prediction page.



Welcome to Hate Speech Prediction!

Prediction Page:

Upon clicking the button, users are taken to the prediction page. This page features two input fields—one for Bangla and another for Banglish, catering to both language preferences. Users can input text in either language to analyze its potential for hate speech.

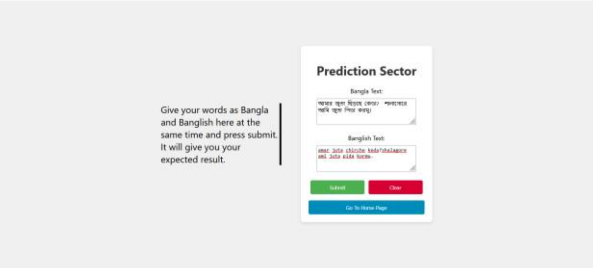


Give your words as Bangla and Banglish here at the same time and press submit. It will give you your expected result.

Prediction Sector

Submission Process:

The prediction page includes a submit button, allowing users to submit their input for analysis. Additionally, a clear button is provided to reset the input fields in case users want to start over.



Give your words as Bangla and Banglish here at the same time and press submit. It will give you your expected result.

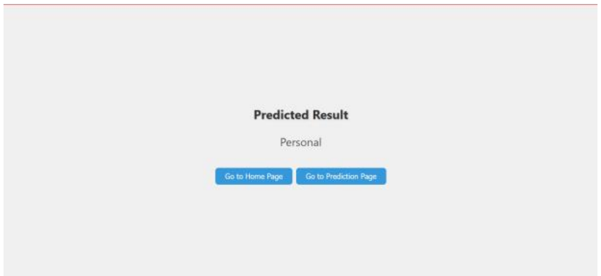
Prediction Sector

Bangla Text:
আমার জুতা ছিড়েছে কেডা? শালারে আমি জুতা পিটা করমু।

Banglish Text:
amar juta chirche keda?shalagore ami juta pida korm-

Result Page:

Once the user submits the input, the website processes the data and redirects to the result page. Here, users receive a comprehensive analysis of the input, revealing the predicted category of
hate speech.



Predicted Result

Personal

4. Result

Two datasets have been used to check which one performs the best. The first one is the balanced dataset where all the five labels are equal and the second one is imbalanced dataset where the labels have biasness. In the imbalanced dataset, the random forest model has worked pretty well with the accuracy of 82%. The SVM and MLP models have done pretty well too with the accuracy rate of 81% which is so close to random forest. The worst performer of this dataset is KNN model which has 48%. And in the balanced dataset using smote, the multi-layer perceptron (MLP) model has performed the best among all the other model with 92% accuracy. The SVM and random forest models have performed as well with 90% and 91% accuracy. Here, KNN model performed bad with
61% accuracy.

Evaluation Parameter:

Accuracy: Accuracy is a measure of the overall correctness of a classification model. It is calculated as the ratio of correctly predicted instances to the total instances.

Recall: Recall, also known as sensitivity or true positive rate, measures the ability of a model to capture all relevant instances of a class. It is calculated as the ratio of true positives to the sum of true positives and false negatives.

Precision: Precision measures the accuracy of positive predictions made by a model. It is calculated as the ratio of true positives to the sum of true positives and false positives.

F1-Score: F1-score is the harmonic mean of precision and recall, providing a balanced metric for model evaluation.

Equation:

$$\begin{aligned} precision &= \frac{TP}{TP+FP} \\ recall &= \frac{TP}{TP+FN} \\ F1 &= \frac{2 \times precision \times recall}{precision + recall} \\ accuracy &= \frac{TP+TN}{TP+FN+TN+FP} \\ specificity &= \frac{TN}{TN+FP} \end{aligned}$$

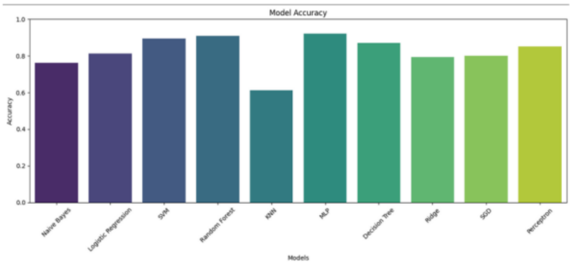
Performance of imbalanced dataset:

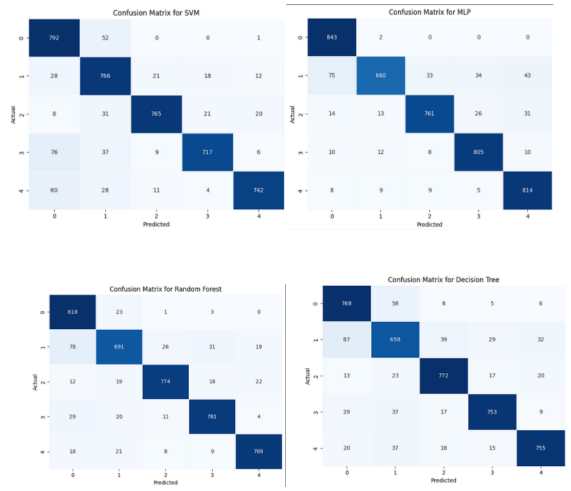
	NB	LR	SVM	RF	KNN	DT	SGD	MLP	Ridge
Accuracy	0.73	0.78	0.81	0.82	0.48	0.79	0.78	0.81	0.78
Recall	0.73	0.78	0.82	0.82	0.57	0.79	0.78	0.81	0.78
Precision	0.73	0.78	0.81	0.83	0.48	0.79	0.78	0.82	0.78
F1-Score	0.72	0.77	0.79	0.82	0.46	0.79	0.77	0.82	0.77

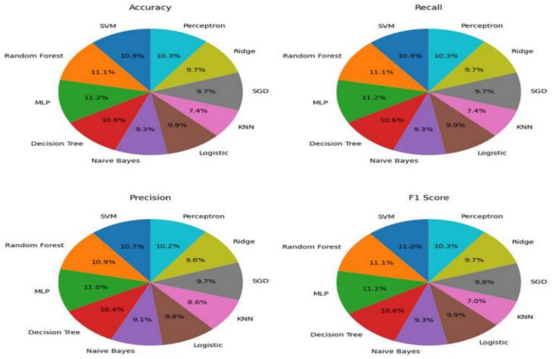
Performance of balanced dataset:

	NB	LR	SVM	RF	KNN	DT	SGD	MLP	Ridge
Accuracy	0.76	0.82	0.90	0.91	0.61	0.87	0.79	0.92	0.79
Recall	0.76	0.81	0.90	0.91	0.72	0.88	0.80	0.92	0.81
Precision	0.76	0.81	0.90	0.91	0.61	0.88	0.79	0.92	0.79
F1-Score	0.76	0.81	0.90	0.91	0.57	0.87	0.79	0.92	0.79

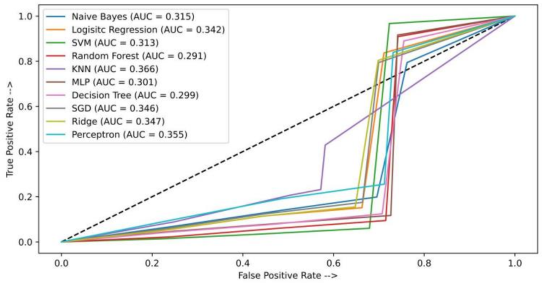
Comparison Graph:







AUC-ROC Curve:



5. Conclusion

We have made a new dataset with ‘banglish’ label. We have used TF-IDF method to vectorize the sentences using ngram. Also, our dataset is imbalance so we apply oversampling smote method to balance the dataset. Then we trained into multiple models on balanced dataset. We have also observed results depends on balanced and imbalance dataset. In the balanced dataset, MLP model has performed very well with 92% accuracy and in the imbalanced dataset, random forest model has performed well with 82% accuracy.

6. Reference

1. Arum Sucia Saksesi, Muhammad Nasrun and Casi Setianingsih, “Analysis Text of Hate Speech Detection Using Recurrent Neural Network,” International Conference on Control, Electronics, Renewable Energy and Communications, pp. 242-248, 2018.

2. N. D. Gitari, Z. Zuping, H. Damien, and J. Long, “A Lexicon-based Approach for Hate Speech Detection,” Int. J. Multimed. Ubiquitous Eng., vol. 10, no. 4, pp. 215–230, 2015.

3. Ricardo Martins, Marco Gomes, Jos´e Jo˜ao Almeida, Paulo Novais and Pedro Henriques, “Hate speech classification in social media using emotional analysis,” 7th Brazilian Conference on Intelligent Systems, pp. 61–66, 2018.

4. Nur Indah Pratiwi, Indra Budi, and Ika Alfina, “Hate Speech Detection on Indonesian Instagram Comments using FastText Approach,” International Conference on Advanced Computer Science and Information Systems, pp. 447–450, 2018.

5. Papers with Code - Multimodal Hate Speech Detection from Bengali Memes and Texts. (n.d.). Multimodal Hate Speech Detection From Bengali Memes and Texts | Papers With Code. https://paperswithcode.com/paper/multimodal-hate-speech-detection-from-bengali

6. Papers with Code - Multimodal Hate Speech Detection from Bengali Memes and Texts. (n.d.). Multimodal Hate Speech Detection From Bengali Memes and Texts | Papers With Code. https://paperswithcode.com/paper/multimodal-hate-speech-detection-from-bengali

7. BANGLA HATE SPECCH DETECTION USING MACHINE LEARNING BY Asfi Hossain Choudury

8. An evolutionary approach to comparative analysis of detecting Bangla abusive text Tanvirul Islam, Nadim Ahmed, Subhenur Latif Department of Computer Science and Engineering, Daffodil International University, Dhaka, Banglades

9. Paz, María Antonia, Julio Montero-Díaz, and Alicia Moreno-Delgado. "Hate speech: A systematized review." Sage Open 10.4 (2020): 2158244020973022.

10. MacAvaney, Sean, et al. "Hate speech detection: Challenges and solutions." PloS one 14.8 (2019): e0221152.

11. Fortuna, Paula, and Sérgio Nunes. "A survey on automatic detection of hate speech in text." ACM Computing Surveys (CSUR) 51.4 (2018): 1-30.

12. Das, Amit Kumar, et al. "Bangla hate speech detection on social media using attention-based recurrent neural network." Journal of Intelligent Systems 30.1 (2021): 578-591.

13. Romim, Nauros, et al. "Hate speech detection in the bengali language: A dataset and its baseline evaluation." Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020. Springer Singapore, 2021.

14. Aporna, Amena Akter, et al. "Classifying offensive speech of bangla text and analysis using explainable AI." International Conference on Advances in Computing and Data Sciences. Cham: Springer International Publishing, 2022.

15. Karim, Md Rezaul, et al. "Multimodal hate speech detection from bengali memes and texts." International Conference on Speech and Language Technologies for Low-resource Languages. Cham: Springer International Publishing, 2022.