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, Bangladesh

Article Info

Article history:

Received Sep 10, 2020
Revised May 28, 2021
Accepted Jun 20, 2021

Keywords:

Abusive text detection
Bangla text
Social media
Supervised learning
TF-IDF

ABSTRACT

The use of Bangla abusive texts has been accelerated with the progressive use of social media. Through this platform, one can spread the hatred or negativity in a viral form. Plenty of research has been done on detecting abusive text in the English language. Bangla abusive text detection has not been done to a great extent. In this experimental study, we have applied three distinct approaches to a comprehensive dataset to obtain a better outcome. In the first study, a large dataset collected from Facebook and YouTube has been utilized to detect abusive texts. After extensive pre-processing and feature extraction, a set of consciously selected supervised machine learning classifiers i.e. multinomial Naïve Bayes (MNB), multi layer perceptron (MLP), support vector machine (SVM), decision tree, random forrest, stochastic gradient descent (SGD), ridge, perceptron and k-nearest neighbors (k-NN) has been applied to determine the best result. The second experiment is conducted by constructing a balanced dataset by random under sampling the majority class and finally, a Bengali stemmer is employed on the dataset and then the final experiment is conducted. In all three experiments, SVM with the full dataset obtained the highest accuracy of 88%.

Corresponding Author:

Tanvirul Islam
Department of Computer Science and Engineering
Daffodil International University
102/1, Sukrabad, Mirpur Road, Dhaka 1207, Bangladesh
Email: tanvirul15-6117@diu.edu.bd

1. INTRODUCTION

The 21st century has been blessed by social media such as Facebook, YouTube, and Twitter. Which has accelerated modern communication with a handful of services. It was reported that 4.021 billion people accessed the internet and 3.196 billion people used social media in 2018. The rate of internet and Facebook users increases by 7 percent and 13 percent each year, respectively [1]. Two hundred five million native Bangla speakers make it the 7th most spoken native language in the world by population [2]. Bangla is becoming very popular and it is being used by 42 million people on different social sites as per the digital report 2018 [1].

This rapid growth in social networking sites has also raised online harassment, hate speech, cyber oppression, online nuisance, blackmailing, and many other cyberbullying. The perpetrators share negative images, abusive comments and messages for harassment. There is a rapid growth of cyberbullying and cybercrime in Bangladesh [3]. 49% of the Bangladeshi students are affected by bullying [4] and there are 73% of women face cybercrime offline or online [5]. The negative impacts of cyberbullying on children have been stated in [6]. Research [7] showed that youth who experienced traditional bullying or cyberbullying had more suicidal thoughts and were more likely to attempt suicide than others. That's why a robust and effective automatic system is a must to detect threats and abusive languages to stop cyberbullying.

The variation in the language and the changing nature of social sites have made it challenging to identify abusive text from a particular language. Comprehensive approaches have been developed to get a satisfactory result in detecting offensive or abusive text. In paper [8], they investigated several learning models, e.g., SVM, RF, radial basis function (RBF), multinomial Naïve Bayes, polynomial and sigmoid kernel. They used unigram, bigram and trigram string property for features extraction, and the evaluation shows that the SVM linear kernel with trigram term frequency-inverse document frequency (TF-IDF) features performs the best. NB classifier has been employed in paper [9], [10] to detect abusive and malicious Bangla text. They used a training dataset collected from Facebook and YouTube comments and employed the algorithm to classify the text. In paper [11], different classification algorithms have been approached to identify cyberbullying on Bangla text where SVM outperforms after cross-validations. [12] proposed a root level algorithm for detecting abusive text and used unigram string features for a better result. Lee *et al.* in [13], an abusive text detection system is introduced utilizing unsupervised learning of abusive words based on word2vec's skip-gram and the cosine similarity. The proposed method shows an f-score of 86.93% in malicious word detection, 85.00% for online community comments and 92.09% for Twitter tweets.

In recent years, opinion mining or sentiment analysis has also become a point of focus in neuro-linguistic programming (NLP). In paper [14], emotion from Bangla text is extracted using a comprehensive set of techniques. The authors classify sentences with a three-class (neutral, positive, and negative) and a five-class (neutral, positive, negative, strongly positive, and strongly negative) sentiment label using a deep learning-based model. In their literature [15], several supervised machine learning algorithm has been proposed to study automated Bangla article classification using a large dataset, which contains almost 3,76,226 articles. A recent study [16] is conducted in classifying abusive text content written in the Indonesian language. The data are labeled into three, i.e., abusive, not abusive and abusive but not offensive. Islam *et al.* introduced a bangla blog article classification system in [17]. Abusive language detection of Twitter text content [18] is implemented utilizing the bidirectional recurrent neural network (BiRNN) method. Besides, results are compared with convolutional neural network (CNN) and RNN, which demonstrated that the proposed BiRNN outperformed CNN.

Islam *et al*. in [19] also showed a comparison between NB, SGD, and SVM for Bangla text document classification. Paper [20] performs abusive content detection on Yahoo finance and news data, with a model combining various syntactic and linguistic features in the text, considered character unigram and bigram level, and tested on Amazon data. Several approaches have been introduced in abusive text detection and most of them are in English languages. Although researchers have rigorously analyzed sentiment from the Bangla language, very few works have been displayed to identify abuses and hate speech with a minimal size dataset. Hence, detecting abusive texts from the Bangla language needs a lot of attention. This experimental research has shown the in-depth exploration of machine learning (ML) algorithms to detect Bangla abusive texts with a comprehensive set of data and compare the performances. We stored about 12,000 data in our corpus from the discussion threads of different controversial Facebook celebrities ' pages and YouTube videos. After pre-processing and feature extraction, diverse techniques, i.e., MNB, MLP, SVM, decision tree, random forest, SGD, ridge, perceptron, and k-NN, have been used to classify the data.

2. RESEARCH METHOD

The block diagram of the entire approach implemented for solving Bangla abusive text detection is delineated in Figure 1. As shown in Figure 1, the process has been started by collecting data from social media. We have accumulated comments from popular pages and channels of Facebook and YouTube under the privacy policy of Facebook and YouTube. Moreover, the data are unstructured and contain plenty of noise. Consequently, several pre-processing techniques have been applied to the dataset. Then the dataset has split into a test set and train set in an 80:20 ratio. Finally, several machine learning techniques are used and performance is evaluated to find the best approaches.

2.1. Description of dataset

Controversial Facebook pages and YouTube channels are the best sources of potential abusive text data. Those pages and channels have thousands of likes, comments, and interactions each day. We have selected some of the most controversial and viral Facebook pages and YouTube channels of Bangladesh's current time like Naila Nayem, Hero Alom, Shefuda, Ripon Video, Model Arif Khan and a few more. Those pages and channels belong to controversial Facebook celebrities, actors, and vulgar content creators. Public comments are scrapped without commenter's information to ensure privacy utilizing graph application program interface (API) and "Youtube Comment Scraper" for Facebook and YouTube, respectively. An irrelevant comment like sticker/emoticon, other language comments except Bangla has been removed from the dataset. As training need tagged data, the dataset is tagged manually as abusive or non-abusive. The distribution of our dataset is given in Table 1.

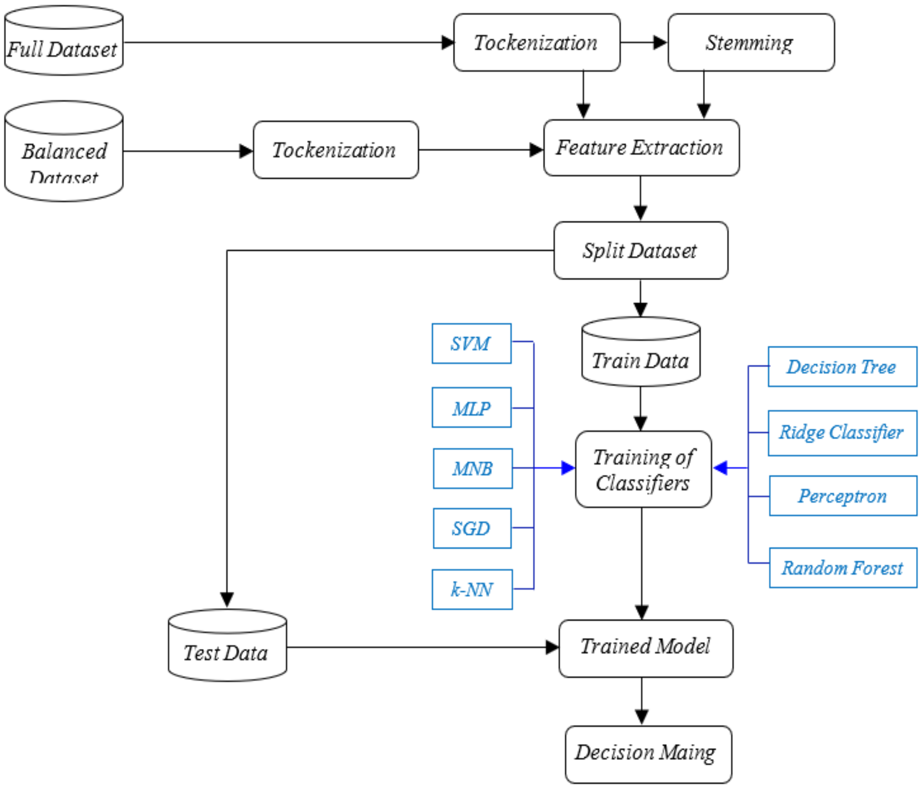

Figure 1. Block diagram of proposed model

Table 1. Description of dataset

Abusive	Non Abusive	Total
4880	7148	12028

Some of the classifiers e.g., Naive Bayes, random forest are sensitive to the proportion of different classes. Consequently, these classifiers tend to bias with the most significant class frequency, which may lead to ambiguous accuracy. Moreover, some models are shown to be less insensitive to class imbalances. Hanse, it has been decided to experiment with a balanced dataset besides our collected imbalanced dataset. There are several techniques for dealing with imbalanced dataset i.e., undersampling, oversampling, synthetic data generation and cost sensitive learning. Accordingly, random undersampling has been chosen in this study. After random undersampling, the distribution of the balanced dataset is following:

Table 2. Balanced dataset

Abusive	Non Abusive	Total
4880	4880	9760

2.2. Data pre-processing

2.2.1. Tokenization

For better classification, it is necessary to go through tokenization and pre-process before applying any feature extraction technique. Tokenization is the system that intends to break the content report into tokens delimited by a newline or tab or white space etc. In this experiment, the token is split based on some delimiter like punctuation, space, newline, emoji, and special character. All the special characters, punctuation, English and Bangla numerals, extra whitespace are eliminated in this stage. Word of other languages except Bangla and single-character words are also removed. Figure 2 portrays the pseudocode of the tokenization process.

```
for i = 0 to i = length (dataset)
        list_of_word = dataset [i] split into (r"\.\s|\?\s|\!\s|\n")
        for j = 0 to j = length(list_of_word)
            marged_document = marged_document+ list_of_word [j] + ” ”
            dataset[i] = marged_document
        end
        marged_document = ””
end
```

Figure 2. Bangla dataset tokenization

2.2.2. Stemming

It is essential to determine the root word of any word for an enhanced classification in a highly inflectional language like Bangla. The idea is to erase inflections from a word to obtain its stem word. Several stemmers exist for the Bangla language including, a rule based bengali stemmer [21], a corpus based unsupervised Bangla word stemming using N-gram language model [22], Designing a Bangla Stemmer using rule based approach [23], N-gram Statistical Stemmer for Bangla Corpus [24]. Papers [25] illustrate that stemming and stopword removal can improve the performance of automated text classification. We have employed [21] a lightweight rule-based stemmer in this research. The sample output of the stemmer used is given in Table 3.

Table 3. Sample output of used stemmer

Input Word	Output Word
মাসের	মাস (Month)
অফিসে	অফিস (Office)
থাকাটাই	থাকা (Stay)
এমনটা	এমন (Thus)
এসেই	এস (Come)
চাকরিটির	চাকরি (Service)

2.3. Feature extraction

The feature extraction method aims to reduce the dimensionality of the corpus by eliminating inappropriate features for classification. Several statistical approaches can be used for feature extraction. TF-IDF have been chosen in this experiment. TF-IDF considers two scores, TF, and IDF. TF counts the frequency of a term within the document, and IDF scales down the term that frequently occurs within multiple documents, as those are less important terms.

2.4. Training of classifiers

A comprehensive study has been performed to find a proper model for Bangla text classification. MNB, MLP, SVM, decision tree, random forest, SGD, ridge classifier, perceptron, and *k*-NN has been considered among the available learning models from the literature review. Each of the datasets is split into an 80:20 ratio as a training set and a test set. All the mentioned classifiers are applied in each of the datasets. Anaconda software platform and python 3 programming language have been used for implementation.

3. EXPERIMENTAL EVALUATION AND RESULTS

Three different datasets have been used to investigate which approach performs better. The first experiment was conducted with the full dataset we have collected where the sample of abusive and non-abusive are not equalized. The performance of different learning models using that dataset is given in Table 4.

SVM obtained the highest accuracy and highest f-score among other algorithms. SGD is also performing very close to SVM, with an accuracy and f-score of 0.87. Ridge classifier and MLP are also performing well, with the accuracy of 0.86 and 0.85, respectively. k-NN and random forest have the poorest performance. As some classifiers perform better with a balanced dataset, the second experiment has been conducted using the balanced dataset generated by random under sampling of the most frequent class of the full dataset. The performance of 2nd experiment is given in Table 5.

The performance of MNB and random forest were improved while experimenting with the balanced dataset. The performance of random forest has a massive change from 0.60 to 0.72. Decision tree and perceptron are performing the same as full dataset. The performance of MLP, SVM, SDG, ridge classifier, and k-NN is dropped. In this experiment, MNB and SVM obtained the highest accuracy.

And lastly, another experiment was conducted by stemming the full dataset (as full dataset have better performance with most of the classifiers) to find if the stemming can improve the performance. A lightweight rule-based stemmer was employed to find the root word of each full dataset word. The performance after stemming is given in Table 6.

Table 4. Performance of different algorithm on full dataset

Algorithm	1st Experiment (Full Dataset)			
	Accuracy	Precision	Recall	F_1 Score
MNB	0.83	0.85	0.79	0.80
MLP	0.85	0.85	0.85	0.85
SVM	0.88	0.88	0.88	0.88
Decision Tree	0.73	0.76	0.72	0.72
Random Forest	0.60	0.36	0.60	0.45
SGD	0.87	0.87	0.87	0.87
Ridge	0.86	0.86	0.86	0.86
Perceptronon	0.80	0.79	0.80	0.79
k-NN	0.55	0.71	0.55	0.51

Table 5. Performance of different algorithm on balanced dataset

Algorithm	2st Experiment (Balanced Dataset)			
	Accuracy	Precision	Recall	F_1 Score
MNB	0.85	0.85	0.85	0.85
MLP	0.68	0.78	0.68	0.64
SVM	0.85	0.85	0.85	0.85
Decision Tree	0.73	0.74	0.73	0.72
Random Forest	0.72	0.74	0.72	0.71
SGD	0.83	0.83	0.83	0.83
Ridge	0.84	0.83	0.83	0.83
Perceptronon	0.80	0.80	0.80	0.80
k-NN	0.53	0.63	0.55	0.46

Table 6. Performance of different algorithm after stemming

Algorithm	3st Experiment (Stemmed Dataset)			
	Accuracy	Precision	Recall	F_1 Score
MNB	0.81	0.82	0.80	0.79
MLP	0.84	0.84	0.84	0.84
SVM	0.85	0.85	0.85	0.85
Decision Tree	0.72	0.75	0.72	0.72
Random Forest	0.60	0.76	0.60	0.45
SGD	0.84	0.84	0.84	0.84
Ridge	0.84	0.83	0.83	0.83
Perceptronon	0.80	0.79	0.79	0.79
k-NN	0.62	0.71	0.62	0.61

After stemming the dataset, the performance of most of the classifiers has slightly decreased except k-NN, random forest and perceptron. Only k-NN has improved performance. Again, SVM performed with the highest accuracy. MLP, SDG and ridge classifier also performed decent, very close to SVM. Still, now, no Bengali stemmer can perform accurately. The stemmer we have used has an accuracy of 83%. In the rest of the 17% case, stemmer is producing many meaningless, invalid word as like, তার (his/her) → ত, মেয়ে (daughter) → মা, ভাই (Brother) → ভা, আছে (have) → আ, তোরা (you) → তো. Those are negatively impacting the performance of classifiers. The comparison of performance among all the classifiers of all the three experiment are illustrated in Figure 3 and Figure 4.

With all the experiments, it is evident that SVM with the full dataset obtained the highest accuracy, while its performance has decreased with under sampling and stemming the dataset. Moreover, the majority of the classifiers are performing better with the full dataset. MNB and random forest show sensitivity with class frequency; consequently, they perform better with the balanced dataset. Only k-NN has a better performance after stemming while the performance of other classifiers is diluted.

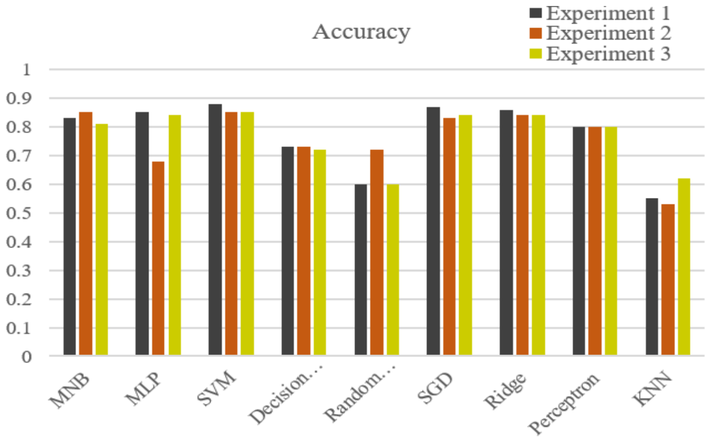

Figure 3. Accuracy comparisn of different classifiers

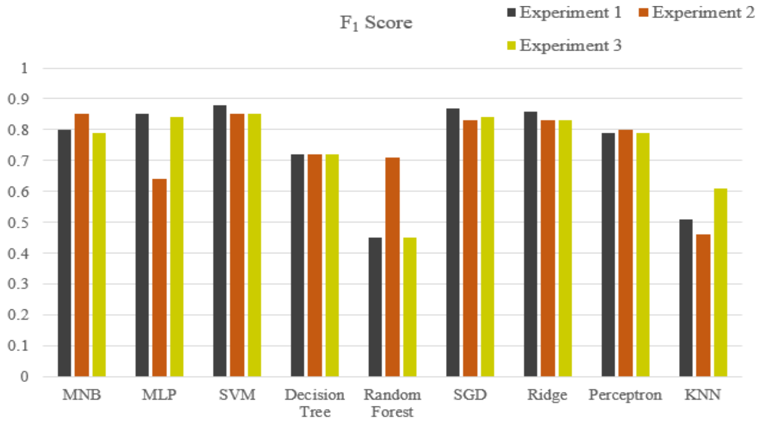

Figure 4. F_1 score comparisn of different classifiers

4. CONCLUSION AND FUTURE SCOPE

Significant research has been done for abusive test detection on language like English, but insufficient work has been done on Bangla. In this research, a dataset of almost 12,000 instances has been collected from different social media. After extensive pre-processing, several learning models, including MNB, SVM, ridge classifier, perceptron, k-NN, random forest and decision tree, have been experimented on this dataset to find the best classifier. It was observed that SVM has the highest accuracy and f-score while MLP, SDG, and ridge classifier are also performing very close to it. On the other hand, k-NN, random forest, and decision tree have a poor performance in this classification task. Stemming is decreasing performance because the accuracy of Bengali stemmer is not satisfactory. As future work, neural network-based models like deep neural network, convolutional neural network, and long short-term memory may be considered. Still, there has scope to enhance our approach. A various machine learning algorithm is being used in NLP while minimal work on Bangla is done because of insufficient resources and monitoring.

REFERENCES

[1] H. Currey, G. Leeding, and M. Nazir, “Digital in 2018: World's internet users pass the 4 billion mark,” *We Are Social*, January 2018. [Online]. Available: https://wearesocial.com/blog/2018/01/global-digital-report-2018.

[2] “List of languages by number of native speakers,” *Wikipedia*, October 2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers.

[3] S. Correspondent, “Cybercrime cases on the rise,” *Prothomalo*. [Online]. Available: https://en.prothomalo.com/bangladesh/Cybercrimes-on-the-rise-due-to-section-57.

[4] Unb, “‘49% Bangladeshi school pupils face cyberbullying',” *The Daily Star*, February 2016. [Online]. Available: https://www.thedailystar.net/bytes/%E2%80%9849-bangladeshi-school-pupils-face-cyberbullying%E2%80%99-287209.

[5] bdnews24.com Senior Correspondent, “73 percent women subject to cyber-crime in Bangladesh,” *bdnews24.com*. [Online]. Available: https://bdnews24.com/bangladesh/2017/03/09/73-percent-women-subject-to-cyber-crime-in-bangladesh.

[6] C. Nixon, "Current perspectives: the impact of cyberbullying on adolescent health," *Adolescent Health, Medicine and Therapeutics*, vol. 5, pp. 143-158, August 2014, doi: 10.2147/ahmt.s36456.

[7] S. Hinduja and J. Patchin, "Bullying, Cyberbullying, and Suicide," *Archives of Suicide Research*, vol. 14, no. 3, pp. 206-221, 2010, doi: 10.1080/13811118.2010.494133.

[8] S. C. Eshan and M. S. Hasan, "An application of machine learning to detect abusive Bengali text," *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 2017, pp. 1-6, doi: 10.1109/ICCITECHN.2017.8281787.

[9] M. A. Awal, M. S. Rahman, and J. Rabbi, "Detecting Abusive Comments in Discussion Threads Using Naïve Bayes," *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 2018, pp. 163-167, doi: 10.1109/ICISET.2018.8745565.

[10] T. Islam, S. Latif, and N. Ahmed, "Using Social Networks to Detect Malicious Bangla Text Content," *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 2019, pp. 1-4, doi: 10.1109/ICASERT.2019.8934841.

[11] Abdhullah-Al-Mamun and S. Akhter, "Social media bullying detection using machine learning on Bangla text," *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*, 2018, pp. 385-388, doi: 10.1109/ICECE.2018.8636797.

[12] M. G. Hussain, T. A. Mahmud, and W. Akthar, "An Approach to Detect Abusive Bangla Text," *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, 2018, pp. 1-5, doi: 10.1109/CIET.2018.8660863.

[13] Ho-Suk Lee, Hong-Rae Lee, Jun-U Park, and Yo-Sub Han, "An abusive text detection system based on enhanced abusive and non-abusive word lists," *Decision Support Systems*, vol. 113, pp. 22-31, September 2018, doi: 10.1016/j.dss.2018.06.009.

[14] N. I. Tripto and M. E. Ali, "Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1-6, doi: 10.1109/ICBSLP.2018.8554875.

[15] M. T. Alam and M. M. Islam, "BARD: Bangla Article Classification Using a New Comprehensive Dataset," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1-5, doi: 10.1109/ICBSLP.2018.8554382.

[16] M. O. Ibrohim and I. Budi, "A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media," *Procedia Computer Science*, vol. 135, pp. 222-229, 2018, doi: 10.1016/j.procs.2018.08.169.

[17] T. Islam, A. Prince, M. Khan, M. Jabiullah and M. Habib, "An in-depth exploration of Bangla blog post classification", Bulletin of Electrical Engineering and Informatics, vol. 10, no. 2, pp. 742-749, April 2021, doi: 10.11591/eei.v10i2.2873.

[18] D. Davis, R. Murali and R. Babu, “Abusive Language Detection and Characterization of Twitter Behavior” *ArXiv, abs/2009.14261*, September 2020.

[19] Md. S. Islam, F. E. Md. Jubayer, and S. I. Ahmed, “A Comparative Study on Different Types of Approaches to Bengali document Categorization,” *International Conference on Engineering Research Innovation and Education*, p. 6, Jan. 2017.

[20] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive Language Detection in Online User Content," *Proceedings of the 25th International Conference on World Wide Web-WWW '16*, April 2016, pp. 145-153, doi: 10.1145/2872427.2883062.

[21] M. R. Mahmud, M. Afrin, M. A. Razzaque, E. Miller, and J. Iwashige, "A rule based bengali stemmer," *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2014, pp. 2750-2756, doi: 10.1109/ICACCI.2014.6968484.

[22] T. T. Urmi, J. J. Jammy, and S. Ismail, "A corpus based unsupervised Bangla word stemming using N-gram language model," *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, 2016, pp. 824-828, doi: 10.1109/ICIEV.2016.7760117.

[23] M. S. Salim Shakib, T. Ahmed and K. M. Azharul Hasan, "Designing a Bangla Stemmer using rule based approach," *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2019, pp. 1-4, doi: 10.1109/ICBSLP47725.2019.201533.

[24] R. Sadia, M. Rahman and M. Seddiqui, "N-gram Statistical Stemmer for Bangla Corpus," *arXiv:1912.11612*, December 2019.

[25] M. Çağataylı and E. Çelebi, "The Effect of Stemming and Stop-Word-Removal on Automatic Text Classification in Turkish Language", *International Conference on Neural Information Processing*, pp. 168-176, 2015. Available: 10.1007/978-3-319-26532-2_19