Classifying Offensive Speech of Bangla Text and Analysis using Explainable AI

Istinub Azad, Amena Akter Aporna, Nibraj Safwan Amlan,
Md Humaion Kabir Mehedi, Mohammad Julfikar Ali Mahbub and Annajiat Alim Rasel

Department of Computer Science and Engineering
Brac University
66 Mohakhali, Dhaka -1212, Bangladesh
{istinub.Azad, amena.akter.aporna, nibraj.safwan.amlan, humaion.kabir.mehedi, mohammad.julfikar.ali.mahbub
}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—The rapid rise of social networking websites and blogging sites not only provides freedom of expression or speech, but also allows people to express society-prohibited behaviors such as online harassment and cyberbully, which are known as offensive speech or hate speech. Despite the fact that, various research work has been done on detecting hate or abusive speech on social networking websites in the English language, the opportunities of research for detecting offensive or abusive speech in the Bengali language remain open, due to the computational resource constraints or the lack of standard-labeled datasets for accurate or effective Natural Language Processing (NLP) of Bangla language. In our paper, an Explainable AI approach is used for analysis as well as for detecting offensive comments or speech in Bengali language was proposed. Moreover, Convolutional Neural Network (CNN) model was used to extract and classify features. Since, the Neural Network is time consuming for extracting features from the dataset, our proposed approach allows people to save time and effort. In the dataset, we classified all user's comments from social media comment sections into four categories: Religious, Personal, Political, and Geopolitical. Our proposed model successfully detects Bangla offensive speeches from the dataset (Bengali Hate Speech Dataset) by evaluating Machine Learning algorithms like linear and tree-based models and Neural Networks like CNN, Bi-LSTM, Conv-LSTM, and SVM models. Moreover, we calculate scores for completeness and sufficiency to assess the quality of explanations in terms of fidelity, achieving the results with the accuracy of 78% score significantly outperforming ML and DNN baselines.**

***Index Terms*—Bangla Offensive Speech Classification, Explainable AI, NLP, CNN, DNN**

I. Introduction

In the present world, to know about a person is becoming easier day by day. Almost every person is connected through social media. By using social media people share their lifestyle daily activities. It is helping people to keep in touch with other people. Social media platforms like – Facebook, Google, YouTube, etc. are connecting people and adding new dimensions towards human life. These platforms or websites of social media are giving service by creating their own profiles, enabling the opportunity for interaction and to read what the other people post.

As social media is playing a great role where people are sharing their feelings, it is becoming a place of harassment for some people, groups, ethnicity, culture or nation. Social media platform is becoming the place where people are becoming the victim of cyberbullying, sexual predation, and self-harm practices incitement. People are spreading hatred among the social media. The victim or the target of this type of hatred are an individual or can be in the common people group. Nowadays, people enjoy attacking others through social media.

Bangla offensive speech detection in the social media or any other online platform is important because it helps to visualize and to find out various forms of abusive languages, like which texts or comments or speech are offensive and which texts or comments are non-offensive. Bangla offensive speech detection is also helpful for the Bengali speaking community to detect harassment, abusive texts, hate speeches and offensive comments in social media which are written in Bengali language. Detecting Bangla offensive speech in social media plays a significant role in detecting racism or racial discrimination. Moreover, detecting hate speeches in social networking websites, one can get acquainted with cyberbullying. It's essential to detect cyberbullying because it is a severe concern in today's age.

The offensive speech spreading is becoming a great concern for social media companies. They are investing a lot in order to get rid of this huge problem. Still their work regarding this field is not sufficient. Distinguishing this type of hate speech is becoming a necessary step for social media. At the present time hate speech in the English language is slowly becoming a familiar scenario. According to the research by Islam [1], 230 million people are speaking Bangla language in the countries in South Asia. But as a major language of the world Bangla, the resources for detecting offensive Bangla language are less. Moreover, the detection of Bengali offensive speech on social media websites faces a variety of challenges for the Bengali speaking community, due to the lack of computational resources or standard-labeled datasets for effective or accurate natural language processing (NLP) of the Bangla language. Also, the models and datasets are not sufficient enough to detect offensive Bangla comments on social media. In this paper, our main contribution is to make the benchmark dataset available and accessible so that

more research work into Bengali offensive speech detection can be done. We focus on distinguishing offensive speech for detecting harassment and abusive texts in Bangla language and classifying the offensive speech. We have also included thorough implementation of DNN in NLP to classify features. The amount of offensive speech in the Bengali language is increasing daily. The manual detection of Bangla offensive speech is quite a challenging task as it takes a long time and is labor-intensive. Since the neural network is time consuming for extracting features from the dataset, our proposed approach will allow people to save time and effort. Moreover, we are classifying the offensive speech of Bangla texts and analyzing it by using Explainable AI (XAI).

In our paper, We have summarized the previous activities which were carried out in the field of offensive speech distinguishing briefly in Section II. In Section III, datasets and its training have been described ,Section IV,the used methodologies in this paper have been described.Result and analysis is shown in Section V and lastly,the conclusion and our future plan is provided in Section VI.

II. Related Work

The growth of social media is quite visible. People are connecting with each other through social media. As in social media people get the freedom to share his/her life but it also enables people to spread hatred about the person. Spreading hatred is now becoming a serious concern around the world. People are attacking people by using social platforms. So, the detection of abusive speech has become a necessary task globally. There are lots of research works on this topic all around the world. Implementation of NLP in the English language is the most common among all as it is the international language. Scientists and researchers from different nations are working on languages other than English. In the research work by Fabio Dell'Orletta et al. (2009) [2] have worked on detecting hate speech in the Italian language where they selected Facebook as a platform for collecting data. In their proposed paper, they used SVM and LSTM models for hate speech detection. There is another work where the author Malmasi and Zampieri (2017) [3] detected hate speech in social media using sentiment analysis where they collected data from Twitter posts and tweets. In their proposed paper, they used different classifiers and majority class baselines. Similar works can be found where they carried out their research on abusive words and sentiment analysis, selecting social media platforms like Facebook and Twitter to collect necessary data. Most of the detection of hate or abusive speech works are carried out using ML, neural networks transformers, and so on. In Machine Learning (ML), the commonly used classifiers are Naive Bayes (NB), Support Vector machine (SVM), Linear Regression (LR), etc. Deep Neural Network architecture (DNN) is also used in the field of NLP for detecting offensive speech in social networking sites. In the research paper, the researcher Yin et al. (2017) [4] discussed about that, Natural Language Processing (NLP) has been transformed by Deep Neural Networks (DNNs). They also mentioned in their paper that, the two primary forms of DNN architecture, Convolutional Neural Network (CNN) (LeCun et al., 1998) [5] and Recurrent Neural Network (RNN) (Elman, 1990) [6], are widely used to perform a variety of NLP applications. RNN is good at modeling units in sequence proposed by the authors like Tang, Adel, Gupta et al. (2016) [7], while CNN is good at extracting position-invariant characteristics which is mentioned by the researcher like Dauphin et al. (2017) [8]. Due to the struggle of CNNs and RNNs, the state-of-the-art on many natural language processing tasks frequently shifts. The goal of this study is to provide fundamental guidelines for DNN selection by comparing CNN and RNN on a wide range of relevant natural language processing tasks.

In the research paper of Karim et al. (2020) [9] the author has talked about the hate speech detection in Bengali language. From different research works and papers on Bangla language, it is found that the resources for thorough and proper research are very limited. Thus, they created the dataset from different newspapers, image posters and the contents from the social media. For detection, they have used multiple approaches to train their data with ML, DNN and transformers and before that they have preprocessed the Bengali texts and classified them into different genres. A pre-trained model of Fast Bangla text is also used here. For data preprocessing they have used the hashtag normalization, stemming, emojis, duplicates, and tokenization. After the data preprocessing they have trained the baseline models. For training machine learning baseline models, they have used character n-grams and word uni-grams with the two term weighting which is the TF-IDF weighting. They have also trained transformed based models in their paper for getting better results.

III. Dataset

The dataset used here is in Bangla Language that is the texts and the string value is in Bangla. The resources of Bangla Literature in case of NLP based research is very limited. Although we are using an existing dataset originally generated by Md. Rezaul Karim and his team in the motive of Hate Speech detection in Bangla Literature we are planning to collect further and precise data to perform the task with better accuracy. The dataset is created from the collected data from different posts, comments and tweets from social media. In this work, the datasets are categorized into four sectors depending on the classification type and they are -

- Personal, (Labelled as 0)
- Political, (Labelled as 1)
- Religious, (Labelled as 2)
- Geopolitical, (Labelled as 3)

As we know, abusive and hate speech can refer to any specific sectors or nations or personal and taking this basis the dataset is labeled and classified accordingly. However, there are some speech or comments that can be put in any category based on the linguistic analysis from the specialists. Here the datasets had been collected using the bootstrap approach.

Many sentences are present here with some common words. These might be directed towards an individual or generalized targeting a community or group which are considered accordingly. A single word might refer to multiple categories. To handle that, a group of words are selected to differentiate it and used for the analysis in which it is used at what percentage. The details are present in the section of result and analysis.

The Data is preprocessed in such a way that extra connecting words, that is the stop words, extra and unnecessary spaces and other symbols are removed by tokenization with the help of lexicons. As the data are collected from social media, in the contents there are lots of unnecessary symbols and emojis that are not related to our work. These are removed and cleaned through the required processes.

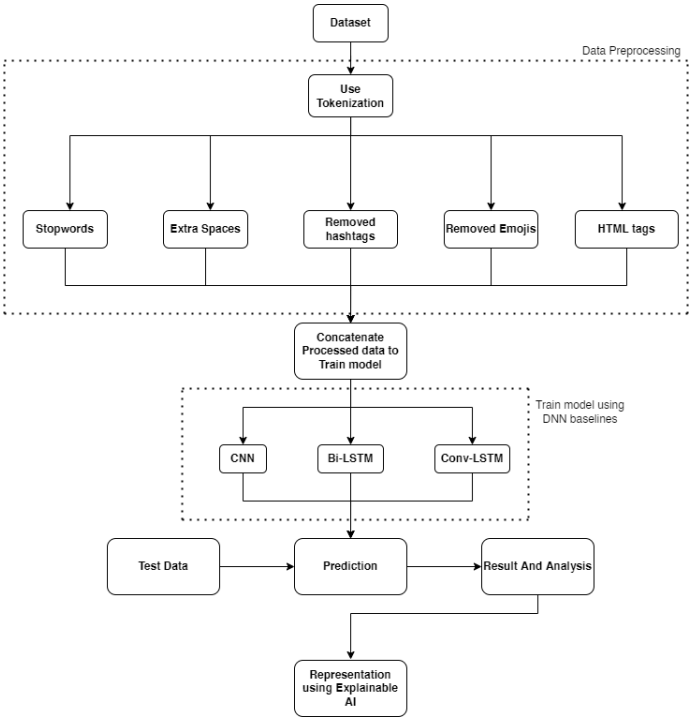

Fig. 1. Workflow Diagram

IV. Methodology

Multiple kinds of approaches are implemented for segmentation and classification of texts for Natural language processing for multiple cases. The most common text classification and segmentation methods are Convolutional Neural Network (CNN), Bi-LSTM, Conv-LSTM, and in the case of Machine Learning approach, Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM) and so on models can be used. In our research paper, we thoroughly implemented the Deep Neural Network (DNN) architecture. We trained the dataset using DNN with the baselines of CNN, Bi-LSTM and Conv-LSTM. In addition, we also tested and trained using ML approaches where we used NB, SVM, RF.

A. Convolutional Neural Network (CNN)

The deep learning method significantly constructed to process the data collected in groups or layers for deep analysis and also for the thorough classification of texts in the form of neural nodes is known as Convolutional Neural Network (CNN). It can also be used for one and three dimensions. In deep neural network architecture, it is known as the most popular artificial neural network. It is used in data analysis, language processing, and other classification problems. In the case of Natural language processing its implementation is much popular because of its high accuracy feature. It analyzes text data and is very good at extracting features from the dataset. These hidden layers of the model are called the convolutional layers. There is one input layer in CNN along with multiple convolutional and non-convolutional layers and finally with the output layer. The neurons of these layers are connected with the nearby same weighted neurons. The two primary operations of CNN are convolution and pooling. Convolution is the process of filtering inputs and results in a feature map. These feature maps for different inputs are stacked together, and it provides the output. The pooling operation mainly reduces the number of parameters. This operation is performed in each feature map.

B. Long Short Term Memory (LSTM)

Long Short Term Memory (LSTM) is a neural network system which can keep track of the sequence in prediction problems. It is basically a complex context of deep learning which is basically used for machine translation,speech recognition,anomaly detection etc. Moreover, it helps to detect the words which are dependent on previous words.It basically keeps track of the words which are used. In LSTM there are basically two states. One state is called the short term memory and another state is called the long term memory.Long term state is basically represented as c and the short term memory is represented as h.

- Bi-directional Long Short Term Memory (BI-LSTM): Bi-directional Long Short Term Memory (BI-LSTM) is basically a two layered neural network system [10]. It is a sequence based model which is the advanced form of LSTM (Long Short Term Memory). In this model the input can flow in both the directions which makes it an advance version of the LSTM. In LSTM, the input flows in one direction, it can either flow in the forward direction or in the backward direction. However, in BI-LSTM, it can flow in both the direction means that the input can flow in backward also it can flow in forward. BI-LSTM is usually applied where we need to use something related to sequence. For the tasks of speech recognition and text classification, BI-LSTM network model can be used.

C. Support Vector Machine (SVM)

The Support Vector Machine, which is also called SVM, is an effective classification approach. Support vector machines (SVMs) are incredible yet adaptable directed AI calculations utilized for classification and regression. Several tasks related

to Natural Language Processing (NLP) have benefited from adopting Support Vector Machines (SVM). The SVM method can be used for various Information Extraction (IE) tasks to achieve better performance. This method has high classifying accuracy and excellent performance in case of any fault generalization. For the Information Extraction task, SVM follows some steps like First of all, SVM converts the problem into multiple steps for classification tasks. Then transforms the problem of those steps into a series of collective binary classification issues. After that, for each binary classification, an SVM classifier is being trained; and finally, the outputs of the classifiers are merged to get a suitable result of the original problem. SVM can be helpful for a classification hyperplane in feature space and the classifier's generalization for which SVM can perform excellently in a good range of NLP tasks. In SVM, the feature vector is formed deliberately from the text using many linguistic properties. The feature vector is mapped into higher dimensional space using the so-called kernel function in many situations. The SVM approach [11] is mainly used for finding an optimal decision boundary for classifying n-dimensional space. For this reason, future categories might easily include more data points. Moreover, SVM determines the extreme points and vectors of hyperplanes to reflect the optimal decision-making boundary. So, for the IE task, SVM significantly provides a good result in NLP applications.

V. Result and Analysis

Here, we discuss the result analysis of the model and a comparative explanation of DNN baselines with ML. Although we preferred the approach of Deep Neural Network, we also implemented ML and presented the results. The following table shows the precision based on the model used.

A. Figures and Tables

TABLE I
Performance Analysis

Method	Classifiers	Precision
DNN Baseliners	CNN	0.73
	Bi-LSTM	0.75
	LSTM	0.78
ML Baseliners	SVM	0.67

B. Explainable AI

After setting up the training model and testing it we represented our data through a special set of framework known as the Explainable AI. Here, the analysed results are interpreted through Artificial Intelligence in the form of graphs and values. It also helps to debug and predict the models so that we can improve it later for further experiments and work in the future. For this procedure we demonstrated the results and analysis generated by implementing the models of DNN and ML. Here we presented global and local explanations. For the former, linguist analysis is used to identify a list of the highest and lowest essential words for all the classes.

Individual example explanations are provided by emphasizing the most significant phrases. We used the leave-one-out experiment to assess word-level relevance for the proper explanation of locales. It was used to implement the backward approach. First, we choose a sample hate statement from the test set at random. Then, for the two most likely classes, we provide prediction probabilities for all of them, followed by an explanation of word-level significance.

The weight of the embedding layer is initialized using fastText embeddings for each DNN baseline model. As can be seen, each model outperforms or performs similarly to the ML baseline models. Conv-LSTM, in particular, outperforms the other DNN baselines which is higher than Bi-LSTM.

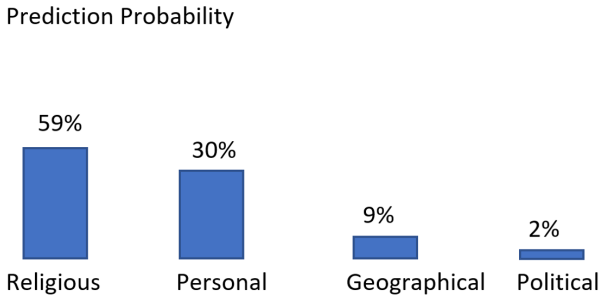

Fig. 2. Relevance test based on Prediction

As we mentioned in the data-set section III that, a word or some common terms are repeated in multiple sentences as well as it falls into multiple categories. A thorough analysis is performed in our paperwork and representation is displayed in Fig-3. The data-set named as Bangla Hate speech that we used for model implementation and analysis, contains hate speech of Religious and Personal category most according to the predictions. The model predicted and detected Religious hate speech which is $59\,\%$. Then, there is hate speech in the category of Personal is $30\,\%$. The least probability we got is the category of Geography which is $9\,\%$ and Political is $2\,\%$. There are some words which are present in multiple categories. These words' predictions are analyzed accordingly. The words are compared in such a way that these are related to a category and how much different it is from a classification. The values are calculated as probability values. The words are compared here on the basis of Religious and Personal classes.

VI. Conclusion and Future Work

In the proposed paper we have implemented Deep Neural network (DNN) baseline models which are used for detecting abusive speech in social media for Bangla Language. It is basically an under-resourced Language. For understanding the debugging we have implemented the explainable AI. Applying this framework of A.I. will help to debug the models for

proper analysis. This analysis and results will contribute in future works in the most efficient ways. In our paper we have implemented the AI explainable model in a way that it has improved detection & classification. After applying the methods we got the maximum precision of 78%. Our paper has some drawbacks as well. Here, we implemented an existing data-set which is limited, but planning to generate more suitable data for the task. In our proposed system we used DNN and the baselines for the classification thoroughly. However, there are many other models which might produce better results which were not applied here. As we know the resources are not enough and we just have used the baseliners so in the future we are going to continue our work. We are going to work to get better precision. Implementing other approaches will help us to compare between the existing approaches and the approaches we are going to implement in our future work. Thus the result will be more accurate.

REFERENCES

[1] Islam, M. S. (2009, June). Research on Bangla language processing in Bangladesh: progress and challenges. In 8th international language & development conference (pp. 23-25).

[2] Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17) (pp. 86-95).

[3] Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media. arXiv preprint arXiv:1712.06427.

[4] Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. arXiv preprint arXiv:1702.01923.

[5] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick ´ Haffner. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324.

[6] Jeffrey L. Elman. 1990. Finding structure in time. Cognitive Science 14(2):179–211.

[7] Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schutze. 2016. Combining recurrent and convolutional neural networks for relation classification. In Proceedings of NAACL HLT. pages 534–539.

[8] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2016. Language modeling with gated convolutional networks. arXiv preprint arXiv:1612.08083 .

[9] Karim, M. R., Dey, S. K., Islam, T., Sarker, S., Menon, M. H., Hossain, K., ... & Decker, S. (2021, October). DeepHateExplainer: Explainable Hate Speech Detection in Under-resourced Bengali Language. In 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 1-10). IEEE.

[10] Isnain, A. R., Sihabuddin, A., & Suyanto, Y. (2020). Bidirectional Long Short Term Memory Method and Word2vec Extraction Approach for Hate Speech Detection. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 14(2), 169-178.

[11] Li, Y., Bontcheva, K., & Cunningham, H. (2009). Adapting SVM for Natural Language Learning: A Case Study Involving Information Extraction. Natural Language Engineering, 15(2), 241-271.