Detection of Bangla Hate Comments and Cyberbullying in Social Media Using NLP and Transformer Models

Md Imdadul Haque Emon , Khondoker Nazia Iqbal , Md Humaion Kabir Mehedi ,
Mohammed Julfikar Ali Mahbub and Annajiat Alim Rasel

Department of Computer Science and Engineering
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{md.imdadul.haque.emon, khondoker.nazia.iqbal, humaion.kabir.mehedi,
mohammed.julfikar.ali.mahbub}@g.bracu.ac.bd
annajiat@bracu.ac.bd

***Abstract*—Hate speech and cyberbullying detection on social media is one of the most trending natural language processing tasks in the current scenario. With the frequent use of Bengali text usage in social media, it is now very urgent to find a robust and consistent way of detecting hate speech and comments for Bengali language. However, due to the lack of resources, a few work has been done in detection of cyberbullying in social media for the Bangla language. In our paper, we have applied different transformer models to detect cyberbullying in social media. We have used a Bangla text dataset [1] with 44,001 Bangla comments which are collected from Facebook posts. This dataset is labeled into five categories: sexual, threat, religious, troll, and not-bully. We have applied three transformer models: Bangla BERT, Bengali DistilBERT, and XLM-RoBERTa. At last, we have compared the performance with another previously done work on cyberbullying detection using machine learning which was implemented on the same dataset. Using transformer models, we achieved a very satisfactory score and the best performance we got was using the XML-RoBERTa model which achieved the highest accuracy of almost 86% and F1-score of 85%.**

***Index Terms*—NLP, Bangla BERT, Bengali DistilBERT, XLM-RoBERTa, Cyberbullying, Bangla hate comments**

I. Introduction

Today people are more attached to the virtual world compared to their real world. Sudden bloom of the smartphone use and internet connectivity has made people more connected than ever. But every blessing comes with its own consequences. And the most common consequence people commonly face is hate comments, harassments and cyberbullying in their personal social media account.

Online harassment and cyberbullying is now in a very alarming level. It has become a trend among people in social media, especially teenagers. They find it very cool to throw hate comment on people. According to a report [2], around 37% of people in the age range of 12-17 have faced bullying online, and approximately 30% are bullied more than once. About 95% of the teenagers of the U.S use mobile as a common medium to bully others. It has been seen that girls are more victims of cyberbullying than boys. According to a report published in Dhaka tribune [3], women are mostly harassed in social media and about 70% of the women are from the 15 to 25 years age group. So this is a huge problem ground to detect harassment and cyberbullying in Bengali language.

For being a a low resource language, number of works in this domain are very insignificant for Bangla. In this paper we used transformer models to detect hate comments and abuse in social media comments, specially Facebook and classify them into 5 categories: threat, sexual, troll, religious and not bully. We have used three transformer models and compared the performance of the transformer based approach against previously applied machine learning approaches on the same dataset.

In this paper, we have briefly summarized previously done works on this domain in Section II. In Section III we have provided the information about the dataset used in our research. After that, in Section IV we have described the methodology and preprocessing steps. Next, Section V includes the results part. And final Section has the conclusion and future work part.

II. Related Works

Different hate speech detection works have already been done for high-level languages like English. Only a small amount of work has been done for low-level languages like Bangla. A recent research work [4] presented a binary and multi class classification model to detect cyberbullying from social media comments in the Bangla language. They have built their own dataset, which contains around 44,001 user comments from different Facebook groups and pages. They have classified the dataset into five categories such as sexual, threat, religious, troll, and not-bully in order to

multi class classification. At first, they applied their binary classification method to detect if there's any bully text and achieved around 87.91% accuracy. They have also used a multi class approach for classifying the harassment comments. For multi-class classification, they achieved 79.29% accuracy. Finally, in order to improve the classification prediction result, they introduced an ensemble method and tried some different machine learning models like SVM, RF, KNN, and Naive Bayes. After applying the ensemble method, they achieved 85% accuracy using SVM. The author struggled with long training time for their model. Moreover, for long text prediction sometimes their model showed false positive result.

Karim et al. used an method using ensemble and transformer, where the author proposed an explainable approach named DeepHateExplainer to detect hate speech in the Bangla language [5]. They have added 5000 texts to the existing Bengali Hate Speech dataset and used it in their paper. At first, they classified the Bengali text into four different classes: personal, political, geopolitical, and religious hatred. After that, they applied machine learning methods (SVM, KNN, LR), Neural networks (CNN, Bi-LSTM), and BERT variants (Bangla BERT, XML-RoBERT) in order to detect hate speech. At last, they compared the prediction result of different methods and found that DeepHateExplainer can detect hate texts with an f1-score of 88% when both ML and DNN baselines are performed together. Their proposed paper proved that a combination of several models can detect hate speech better than individual models. Because of having a limited number of labeled data while training their model, the author couldn't rule out the chance of over-fitting.

In another paper [6], the authors prepared three datasets of Bengali text consisting of Bangla hate speeches to do three different experiments: sentiment analysis, hate speech detection, and document classification. They introduced a word embedding model for Bangla language and named it "BengFastText". It contains data based on around 250 million articles. After that, based on the Multichannel Convolutional Long Short-Term Memory Network, they predicted the result of the three experiments as mentioned earlier. After that, they compared their prediction result with other models. Their experiment's outcome showed that BengFastText can detect the texts more correctly than other embedding methods like Word2Vec and GloVe. Using the BengFastText method, they achieved around 92.30% F1-scores in document classification and around 82.25%, and 90.45% F1-scores are achieved in sentiment analysis and hate speech detection.

In 2020 Makhadmeh et al. proposed a model for using NLP and ML approaches combinely to detect abusive comments from social media in the English language [7]. They have collected data from a neo-Nazi website [8], which contains around 10,568 sentences, and each sentence is around 20.39 words in length. They explore the dataset using their proposed method named as "A killer natural language processing optimization ensemble deep learning method (KNLPEDNN)". By using the approach, the dataset is classified into three different classes, which includes hate, offensive, and neutral languages. Their proposed method achieved a maximum accuracy of around 98.71% to predict hate speech from social media texts.

Another research by Akhter et al. [9] proposed an approach to detect social media abuse using different ML algorithms on Bangla language. They have collected data from different social media platforms like Facebook and Twitter. In order to extract the dataset from Facebook and Twitter, they have developed a java program. They have collected 1000 public Bangla comments from Facebook and 1,400 comments from Twitter. After that, they labeled the dataset into two categories: "bullied" and "not bullied". At last, they applied Machine Learning algorithms like SVM, Naive Bayes, KNN (1-Nearest), KNN (3-Nearest) to predict bullying. After comparing different algorithms, they found that the support vector machine's prediction accuracy is highest than other applied Machine Learning algorithms. They achieved around 97% accuracy while detecting bullying using the SVM approach.

Another proposed method for cyberbullying detection [10] has applied a pre-trained BERT model, which is a very popular transformer model. They have used two publicly available datasets, Formspring [11] and Wikipedia [12] talk pages, to train and evaluate the model. The first dataset contains around 12,773 posts, among which 776 are marked as a bully. On the other hand, the Wikipedia dataset has about 1,15,864 posts, with 13,590 of them labeled as bullies. They have used a hugging face library to work with BERT models. They have compared their models' accuracy with other models like CNN and RNN. From the Formspring and Wikipedia datasets, the BERT model achieved around 98% and 96% accuracy respectively.

III. Dataset

In our research paper, we have collected Bangla online comments dataset from the Mendeley Data website [1]. This dataset contains around 44,001 Bangla comments collected from Facebook posts. The dataset is labeled into five categories: sexual, threat, religious, troll, and not-bully. According to the dataset, 29,950 comments are for females, and 14,051 comments are for males. Besides, the dataset's comments are categorized into five different professional categories: actor, singer, social, politician, and sports. It contains emoticons, special characters, stopwords which are removed from the dataset while preprocessing. In the dataset, the number of reacts to each comments are also mentioned. In our proposed model, we have only used the five different categories of label to detect and classify cyberbullying.

Some examples of the labels of our dataset are shown in below Table I:

TABLE I
Dataset Example

Label	Example
Sexual	খানকীর ছবি দেখলেই ঘিন্না লাগে
Threat	তরে কুট্টাইলবাম
Religious	নাস্তিকের বাচ্চা নাস্তিক! জাহান্নামের কিট
Troll	ফইন্নির ঘরের ফইন্নি
Not-Bully	আপনার জন্য ভালবাসা

IV. Methodology

The dataset we used was previously used in another research [4] for cyberbullying detection. They proposed a deep neural network (DNN) approach and also ensemble approach by using four supervised machine learning algorithms (Random Forest (RF), SVM, KNN and Naïve Bayes). We used three transformer models (Bangla BERT Base [13], Bengali DistilBERT and XLM-RoBERTa Base [14]) from Hugging Face [15] for detection and multi-class classification of hate comments and will compare the accuracy with previously applied approaches. Before training the models with text data we have done some pre-processing of the data. The workflow diagram is shown the Figure 1.

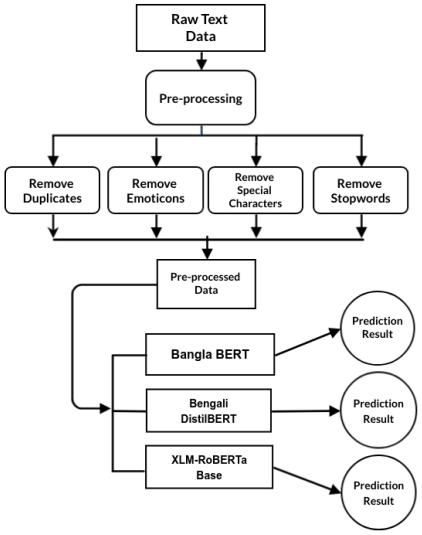

Fig. 1. Workflow Diagram

A. Preprocessing

The dataset had a total number of 44,001 comments. The comments consisted of Bengali text with a mixture of numeric, emoticons, special characters(punctuation and other symbols) and some of the comments had some English text. Also, there were some duplicate comments in the dataset. At first we removed the duplicate comments from the dataset. There were a total 434 duplicate comments. After removing the duplicates we were left with 43,567 comments. Then we removed all the special characters and also removed all the English text from the comments. All the emoticons were removed and now we are left with only Bengali text in the comments.

At last we removed the stopwords from the data. The stopwords were taken from a GitHub repository [16]. The author listed a total number of 398 stopwords for Bengali language and all of them were removed from the dataset.The dataset pre-processing example is shown in the below Table II.

TABLE II
Pre-processed Text Example

Before Pre-processing	After Pre-processing
অন্যরকম .. ভালো লাগলো ..☺	অন্যরকম ভালো লাগলো
অদেখা জিনিসে ত ওর believe নাই,এইটা কেমনে করলো..	অদেখা জিনিসে ত নাই এইটা কেমনে করলো
সাফা কবির কোন **** বাল... ???	সাফা কবির বাল
নাস্তিকের বাচ্চা নাস্তিক! জাহান্নামের কিট।	নাস্তিকের বাচ্চা নাস্তিক জাহান্নামের কিট

B. Transformer Models

Bangla BERT Base: Bangla BERT Base [13] is a pre-trained transformer model for Bengali language. This BERT based model is pre-trained with Bengali CommonCrawl corpus from OSCAR1 [1] and Bengali Wikipedia Dump [2] Dataset2. This model was pre-trained using mask language modelling.

Bengali DistilBERT: Bengali DistilBERT is another transformer model for Bengali language, and this model was pre-trained on almost six Gigabyte of monolingual training corpus. This is a very lightweight model and provides a very good accuracy for downstream works like POS-tagging and text-classification. Among the three models we applied, the training time of this model was the least.

XLM-RoBERTa-Base: XLM-RoBERTa [14] is a very famous transformer model which supports multilingual texts. This model is pre-trained on 2.5TB of filtered data which has 100 different languages. This model was pre-trained on raw text data only using Masked Language modelling. This model has been used in a lot of text-classification and downstream tasks before and it showed excellent accuracy in almost every task.

Model Training Parameters: We wanted to train all the models using the same parameters so that the comparison of the test-score becomes more logical. So we trained all of these transformer models using a learning rate of $2e^{-5}$ in 10 epochs. All of the models were trained

[1]https://oscar-corpus.com/
[2]https://dumps.wikimedia.org/bnwiki/latest/

using a batch size of 12. Also all of the data was divided using a ratio of 70:20:10 for train, validation and test. So 70% data was used for training while 20% of the data was used for validation and last 10% of the data was used for further testing the accuracy.

V. Result and Analysis

All of the transformer models scored a satisfactory result. The Bangla BERT model scored 83% f1-score , 83% precision and 82% recall score. For the test data it scored 82.64% accuracy. On the other hand, the Bengali DistilBERT model scored a similar score to Bangla BERT, 83% f1-score, 83% precision and 82% recall score while scoring 82.73% accuracy for the test data. At last, the XLM-RoBERT model gave a most promising score: 85% f1, 84% precision and 84% recall. Also this model scored 85.42% accuracy for the test data. All the results are shown in the following table with a comparison of the accuracy score from "Cyberbullying Detection Using Deep Neural Network from Social Media Comments in Bangla Language" paper [4].

The Accuracy (Acc.), Precision (Pr.), Recall (Re.) and the F1-score (F1) of our applied transformer models and the previous ensemble approach are shown in Table III.

TABLE III
Accuracy Score Comparison

Approach	Model	Acc.	Pr.	Re.	F1
Previous Approach	RF	0.84	0.84	0.84	0.84
	SVM	0.85	0.85	0.85	0.84
	KNN	0.84	0.85	0.84	0.84
	Naïve Bayes	0.79	0.78	0.79	0.78
Transformers	Bangla BERT	0.83	0.83	0.82	0.83
	Bengali DistilBERT	0.83	0.83	0.82	0.83
	XLM-RoBERTa	**0.85**	**0.84**	**0.84**	**0.85**

From the table, it can be observed that the XLM-RoBERT model scored the highest F1-score among all of the approaches. Here the training vs validation accuracy and loss per epoch for the XLM-RoBERTa model is shown in Figure 2 and Figure 3.

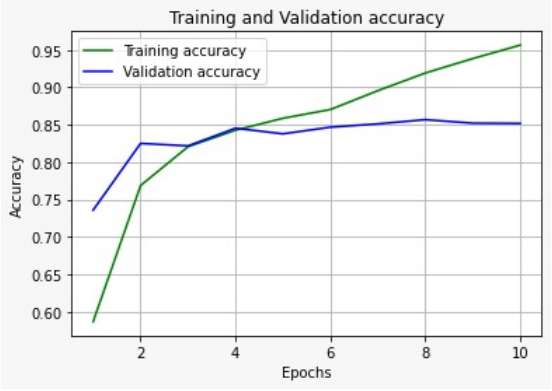

Fig. 2. Epoch vs Accuracy for training and validation(XLM-RoBERTa)

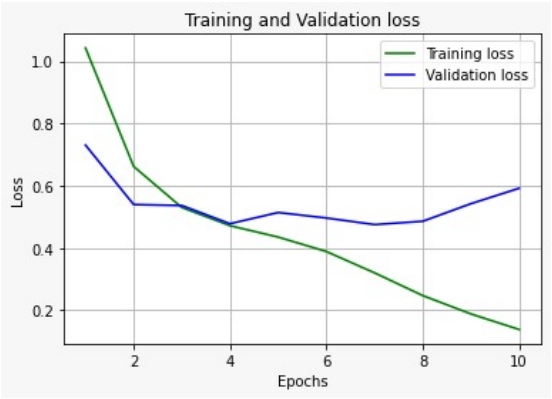

Fig. 3. Epoch vs Loss for training and validation(XLM-RoBERTa)

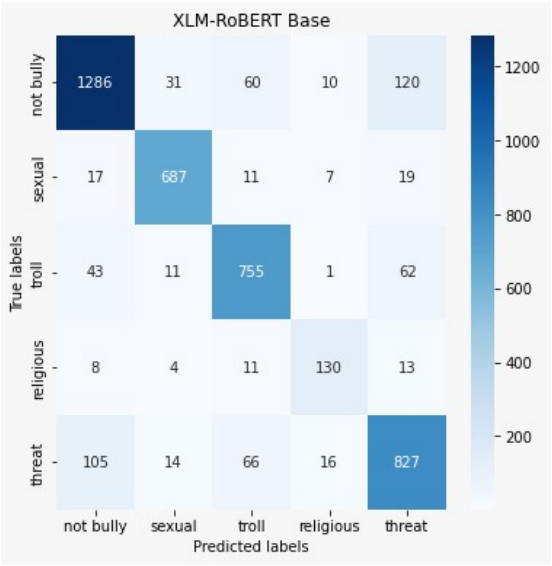

Fig. 4. Confusion Matrix of XLM-RoBERTa

The confusion matrix of Figure 4 shows the prediction labels for cyberbullying detection using XLM-RoBERTa base.

VI. Conclusion

Our research portraits that using transformers approach shows a better F1-score against the traditional ensemble learning approach to detect and classify Bangla hate comments and bullies in social media platforms, specially Facebook. With time, the toxicity is increasing and more people are joining everyday to make the task more complex. Transformers can be a very good solution for hate speech detection and classification. The task is still quite challenging because of the low resource availability of Bengali language. This paper can be a dedication for building a more accurate and robust transformer model to filter out the underlying bully and hatred of Bengali social media comments. In future, we look forward to implement transformer models for other downstream tasks in Bengali language.

References

[1] M. F. Ahmed, Z. Mahmud, Z. T. Biash, A. A. N. Ryen, A. Hossain, and F. B. Ashraf, *Bangla online comments dataset*, 2021. [Online]. Available: https://data.mendeley.com/datasets/9xjx8twk8p/1.

[2] *11 facts about cyberbullying*. [Online]. Available: https://www.dosomething.org/us/facts/11-facts-about-cyber-bullying (visited on 12/24/2021).

[3] *70% of women facing cyber harassment are 15-25 years in age*$_2$019. [Online]. Available: https://archive.dhakatribune.com/bangladesh/dhaka/2019/09/24/70-of-women-facing-cyber-harassment-are-15-25-years-in-age (visited on 12/20/2021).

[4] ——, "Cyberbullying detection using deep neural network from social media comments in bangla language," *arXiv preprint arXiv:2106.04506*, 2021.

[5] M. R. Karim, S. K. Dey, T. Islam, *et al.*, "Deephate-explainer: Explainable hate speech detection in under-resourced bengali language," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2021, pp. 1–10.

[6] M. R. Karim, B. R. Chakravarthi, J. P. McCrae, and M. Cochez, "Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2020, pp. 390–399.

[7] Z. Al-Makhadmeh and A. Tolba, "Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach," *Computing*, vol. 102, no. 2, pp. 501–522, 2020.

[8] N. Caren, K. Jowers, and S. Gaby, "A social movement online community: Stormfront and the white nationalist movement," 2012.

[9] S. Akhter *et al.*, "Social media bullying detection using machine learning on bangla text," in *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*, IEEE, 2018, pp. 385–388.

[10] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained bert model," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, 2020, pp. 1096–1100.

[11] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th International Conference on Machine Learning and Applications and Workshops*, vol. 2, 2011, pp. 241–244. DOI: 10.1109/ICMLA.2011.152.

[12] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17, Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399, ISBN: 9781450349130. DOI: 10.1145/3038912.3052591. [Online]. Available: https://doi.org/10.1145/3038912.3052591.

[13] S. Sarker, *Banglabert: Bengali mask language model for bengali language understading*, 2020. [Online]. Available: https://github.com/sagorbrur/bangla-bert.

[14] A. Conneau, K. Khandelwal, N. Goyal, *et al.*, "Unsupervised cross-lingual representation learning at scale," *CoRR*, vol. abs/1911.02116, 2019. arXiv: 1911.02116. [Online]. Available: http://arxiv.org/abs/1911.02116.

[15] *Hugging face, the ai community building the future.* [Online]. Available: https://huggingface.co/ (visited on 12/15/2021).

[16] G. Diaz, *Stopwords bengali (bn)*, 2016. [Online]. Available: https://github.com/stopwords-iso/stopwords-bn.git (visited on 12/23/2021).