Gender-based Cyberbullying Detection for Under-resourced Bangla Language

Hasnain Karim Rabib[1], Mostafa Galib[1], Takia Mosharref Nobo[1], Tanjila Alam Sathi
Mohammed Saidul Islam, Abu Raihan Mostofa Kamal, and Md Azam Hossain[*]
Network and Data Analysis Group (NDAG), Department of Computer Science and Engineering
Islamic University of Technology (IUT), Gazipur-1704, Bangladesh
Email: {hasnainkarim, mostafagalib, takiamosharref, anjilaalam187, saidulislam, raihan.kamal, azam}@iut-dhaka.edu

2022 12th International Conference on Electrical and Computer Engineering (ICECE) | 979-8-3503-9879-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECE57408.2022.10088574

***Abstract*—The present study explores the detection of gender discrimination based cyberbullying in under-resourced Bangla language. While being spoken by 230 million people globally and being rich in diversity, the Bengali language lacks computational models and annotated resources for cyberbullying detection. To address this, we created GenDisc, a corpus of Bangla Facebook comments featuring gender-based cyberbullying. This study also presents a framework for cyberbullying detection. In our proposed approach, we used four different models to train a gender-based discriminatory text classifier, followed by an ensembling technique on those four models. Then we compared the individual prediction accuracies with the ensembled prediction accuracy. While training the dataset, we followed the stratified k-fold cross validation technique. We demonstrated that integrating gender-based discrimination variables improve a classifier's capacity to discriminate against cyberbullying. Our evaluations yielded an accuracy of 68% in gender-based speech detection during cross-validation tests.**

***Index Terms*—Cyberbullying, Gender discrimination, Transformer, BERT, Ensemble**

I. Introduction

Young people have fully embraced the Internet as a social and communication tool [1]. However, they have also promoted antisocial conduct and online abuse [2]. A survey [3] conducted in 2021 reveals that 14% of teenagers had encountered internet bullying at least once per week . One of the worrisome areas is speech depicting discrimination based on gender or online sexual harassment.

The main purpose of a reliable cyberbullying detection system on a social platform is to monitor, prevent, or at least minimize the incidence of cyberbullying. Moderators of online discussion boards cannot read every single post. Although cyberbullying is a well-researched topic in the social sciences, much of the literature on the topic is published in major languages such as English, whereas Bangla remains unexplored [4].

Bangla is a rich language spoken in Bangladesh, as well as the second most frequently spoken dialect in India and the seventh most widely spoken dialect in the world, with around 230 million native speakers. More than 90% of Bangladesh's 80.83 million Internet users [5] are on Facebook, where the majority is young, insecure, and anxious for security. Moreover, little research has been done on the Bangla language for social media monitoring platforms due to a lack of annotated corpora, named dictionaries, and morphological analyzers [6]. As a consequence, this research focuses on cyberbullying based on gender discrimination in Bangla.

- We have created a dataset of bullying expressions in Bangla language which consists of comments stemming from sexual harassment to gender discrimination.
- To the best of our knowledge, this is the first investigation in the Bangla language that concentrates on the gender discriminatory class of cyberbullying.
- Proposed study trained a gender-based discriminating text classifier using four models namely BERT, mBERT and DistilBERT and mDistilBERT. We demonstrated that our proposed framework nets satisfactory results in detecting gender based cyberbullying.
- We also used stratified k-folds cross validation to train our dataset. We used an ensembling strategy to determine the best accuracy after receiving predictions from the four models.

The rest of the paper is structured as follows: Section 2 reviews the related work on online bullying and Bangla hate speech detection. Section 3 describes our Dataset Preparation process: dataset creation, annotation, pre-processing and dataset augmentation. Section 4 contains our proposed framework where the processes of model selection and training, along with the experimental settings and results have been discussed. This is followed by the performance evaluation. Section 5 talks about the challenges and future scopes for work before concluding the paper.

II. Related Work

With the growth of social media there has been a surge in interest in this field of hate speech [7]. But the majority of this work has been done on abundantly resourceful languages like English [8]. Twitter has been a major platform to conduct these studies on hate speech because most of the people who speak English predominantly use this platform [9]. The absence of systematic text collecting methods, annotated corpora, named dictionaries, morphological analyzers, and overall research perspectives makes it difficult to delve into this area. There has been some recent research in the Bangla language, but

[1]Authors have equal contribution.
[*]Correspondence: azam@iut-dhaka.edu.

none, to the best of our knowledge, has focused on gender-based hate speech detection in Bangla.

In the work [10], authors proposed the use of machine learning (ML) algorithms and the inclusion of user information for cyberbullying detection on Bangla text. It shows that the impact of user-specific information such as location, age, and gender can further improve the classification accuracy of Bangla cyberbullying detection systems. Chakraborty at al. [5] proposed to build an automatic system using Machine Learning and NLP techniques to identify threats and abusive languages. They considered both Unicode emoticons and Unicode Bengali characters as valid input in our proposed system. Besides MNB and SVM algorithms, the work also implemented Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM).

In a recent work, Paul et al. [11] proposed a novel application of BERT (Bidirectional Encoder Representations from Transformers) for cyberbullying identification. BERT has achieved remarkable results in many language understanding tasks. In the study [12], authors examined the use of BERT for cyberbullying detection on various datasets and attempted to explain its performance by analyzing its attention weights and gradient-based feature importance scores for textual and linguistic features.

For the under-resourced Bengali language, Karim et al. [4] provided DeepHateExplainer, an explainable solution to hate speech identification. With an F1 score of 88%, DeepHate-Explainer is able to recognize a wide range of hate speech, beating various ML and deep neural network (DNN) baselines.

III. Dataset Preparation

In this section, we discuss the methodology behind the creation of the GenDisc (Gender-Based Discriminatory) dataset.

A. GenDisc Dataset Creation

Source selection: The most commonly used social media platforms are Facebook, Youtube, Tiktok, Twitter etc. But for our dataset, we selected Facebook, because it is the most widely used social network platform in Bangladesh and the biggest hub for abusive and hateful comments and speeches.

Data sample selection criteria: For gender specific data samples, we looked for comments which were targeted towards both men and women and showed hints of **discrimination, abuse, harassment, and victimization**

Scraping: We have chosen Instant Data Scraper[1] for collecting our data samples to create our GenDosc dataset. It is a free to use online scraping tool, which is easy to work with; especially for collecting data from Facebook.

Annotation: Annotation can be done either through an automatic process or manually with the help of expert linguists. As Bangla is not very common in the NLP domain, hardly any automatic process works on Bangla and so we had to move to the manual annotation process.

Annotation Criteria: Our dataset had data samples which had direct hints of gender discrimination, but there were also ample amount of data samples which were hard to put into a specific category. And in that case, we needed experts on Bangla linguists who have better sense and context about the language. They looked for direct representation of gender based hate, discrimination, and harassment in the comments and labelled the data samples accordingly. The samples which contained gender based discrimination were labelled as '1' and the opposites were labelled as '0'. Figure 1 illustrates some sample of our dataset. GenDisc has around 2600 data samples. Table I shows data distribution in our dataset.

text	label
দিতিও সানি লিওন বাংলাদেশে কি দারুণ	1
তর মত নষ্ট মেয়ে পার্টিতে জারা আসে তারাও নষ্ট মানুষ	1
অনেক কিছু বলতে ইচ্ছা করে ভাই কিন্তু বলতে পারিনা	0

Fig. 1. GenDisc: Data Sample

TABLE I
GenDisc: Data Distribution

Type	Data Samples	(%) in Dataset	(%) in Train + Val	(%) in Test
Non-Discriminatory	1421	55.40	53.13	52.93
Discriminatory	1146	44.60	46.87	47.07
Total	2567	100	100	100

B. Data Pre-processing

The web-scraped dataset has undergone a variety of conventional preprocessing processes. The major objective of pre-processing was to remove redundant, unnecessary words, emojis, and characters from the dataset, followed by the tasks of stop-word removal and data sample duplication reduction.

We got rid of non-Bangla, meaningless and numerical (0-9) words which we deemed redundant. The characters or symbols, mainly the punctuation marks, were removed. Moreover, parenthesis and meaningless single characters (if there were any) were also removed. Emoticons,emojis or any graphical items in the text were also removed. There is a long list of commonly used stop-words in Bangla language. A customized collection of some common Bangla stop-words was collected from a public repository on github; which was free to use and open for extension and eventually it was appended with more of significant stop-words in Bangla language. Each and every data sample in our dataset was filtered for finding stop-words and then were removed. Our dataset had some overlapping or duplicate data samples, which we had to get rid of as they might add up to the biases in the dataset and will result in the enhancement of the context of a specific category of the dataset.

C. Data Augmentation:

To achieve better and more satisfactory results on our dataset, we decided to augment our dataset and the method that we used was random swapping (RS). In RS, two words are randomly selected from the sentence and their orders are swapped.

[1]https://chrome.google.com/webstore/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah

IV. Proposed framework

In the NLP area, innovative research has recently been performed to develop better architectures and models, which have fundamentally resulted in exceptional increases in the performance of task-specific NLP applications. Consequently, the world came across the 'Transformer' [13] architecture. Transformer architecture is a model which has a set of encoders and decoders as its building blocks and adopts the mechanism of self-attention to give equal importance to each part of the input data.

In this work, we proposed an ensemble method to detect gender based cyberbullying. As our dataset is a complete Bangla dataset, we opted for multilingual models along with the vanilla models. Hence, we have selected the BERT model, the multilingual-BERT model, the DistilBert model and the multilingual Distilbert model. In the end, we implemented an ensemble technique on the individual predictions of the aforementioned four models which produced better results than the four individual models. Figure 2 depicts the proposed framework for gender-based cyberbullying detection.

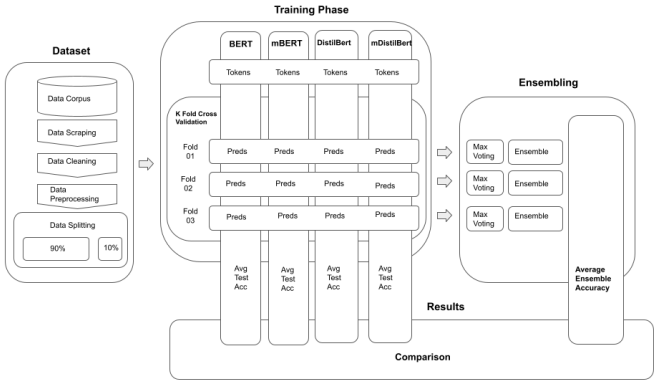

Fig. 2. Proposed Framework Architecture and Methodology

A. Model Training

After data preprocessing and model selection, we moved on to the training phase. For the specific task of binary text classification, we trained our pretrained models on our new dataset. We divided the whole dataset into two parts (train and test) while keeping a constant ratio (90:10).

To make our models perform better in case of unseen data, we have implemented the **K-fold cross validation** technique on our dataset. Therefore, in each fold of the data, a validation set was also included, and the ratio was 80:10:10. This procedure iterates k times, each time reserving a distinct subset for testing purposes. We set the number of folds to 3, or k=3, and the number of epochs per fold to 15, or epochs=15. In each epoch, the training accuracy and loss, as well as the validation accuracy and loss, were calculated. Figure 4 shows the training accuracy curves for all the four models. For classification, BERT and its variants require special CLS Tokens. Each model has its own tokenizer.

The tokenizers were used to generate unique token ids and attention mask for each word in the input dataset; which were eventually fed to our model while training. Here, the weight initialization was done with random numbers instead of all '0's or all '1's. Since we have used the models from the imported transformer library from hugging face, we did not need to explicitly call the weight initializer function, because the function is invoked by default. We used the Adam Optimizer which combines the best properties of the two algorithms: AdaGrad and RMSProp. Since the task is a binary text classification, we have used the Sigmoid Activation function. For calculating the loss and adjusting the weights, we used the Cross Entropy Loss function. We had to use small learning rates in order to make the loss converge to a point, as our dataset was comparatively small in size.

B. Experimental Settings

We conducted K-fold cross validation technique; where k=3. So, we trained each of our models three times with a different set of training and validation data folds in each training. The table 3 shows the experimental parameters used.

For hyper-parameter tuning, we conducted hundreds of experiments with a different set of learning rate, batch size and number of epochs for each model. But the best results were found for the specific sets of parameters shown in the table II.

TABLE II
Experimental Settings

Model	Batch Size	L.R.	Epochs	Max Length
Bert	16	5e-6	25	150
mBert	16	5e-6	25	150
DistilBert	16	1.4e-6	25	150
mDistilBert	16	5e-5	25	150

Ensemble technique is a machine learning approach to combine multiple models in the prediction process. Here we take the four trained models from each fold and combine the test data predictions of each of those models to make an ensemble prediction following the Max voting method. In max-voting, each base model makes a prediction and votes for each sample. Only the sample class with the highest votes is included in the final predictive class as shown in Figure 5. For each epoch, we evaluated the model by calculating the loss and accuracy of training and validation dataset.

C. Results And Discussion

text	predicted label
ওর চেয়ে পতিতলার মেয়েরাও ভালো	1
কম পানির মাছ বেশি পানিতে পরলে যা হয় আর কী	0

Fig. 3. GenDisc: Model outcome sample

After the models were trained, each of them were tested on the test data. The models predicted whether the instances of the test data have the sense of gender based discrimination or not as shown in Figure 3. For each model, we calculated the score in each fold. Afterwards, the average of the scores of all the three folds is taken as the cross validation score for that specific model; which is the final accuracy of that model.

For each fold, we get a different set of scores from all the four models and the ensemble model, which have been

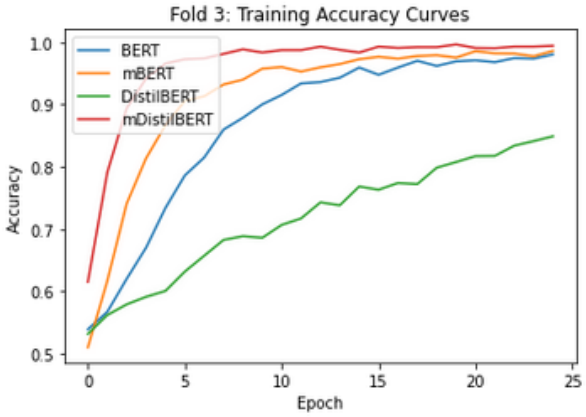

Fig. 4. Accuracy per epoch for fold 3

compared in table III. The best accuracy was achieved by the ensemble method that we have proposed.

TABLE III
GENDISC: PERFORMANCE COMPARISON

	Test Accuracy (%)				
	DistilBERT	mDistilBERT	BERT	mBERT	**Ensemble**
Fold 1	64.61	63.31	65.26	63.96	**68.50**
Fold 2	65.58	66.56	66.56	65.91	**69.16**
Fold 3	63.31	**68.51**	64.29	63.96	66.56
Avg	64.5	66.12	65.37	64.61	**68.07**

D. Performance and evaluation metrics

We have evaluated the models on several performance measure metrics to get a more vivid and convenient idea about the performance of the model. We used the performance metrics - Accuracy, Mcc Score, Precision, Recall and F1. In the table IV, we can see the above mentioned metric scores of our ensemble approach in each of the three folds.

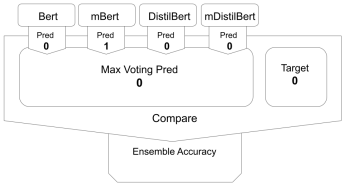

Fig. 5. GenDisc: Ensemble Process

TABLE IV
GENDISC: ENSEMBLE ACCURACY FOR EACH FOLD

Fold	**Ensemble Acc**	**Mcc Score**	**Precision**	**Recall**	**F_score**
1	68.50	36.82	68.52	68.51	68.51
2	69.16	38.18	69.19	69.16	69.17
3	66.56	33.10	66.68	66.56	66.59
Avg	**68.07**	**36.03**	**68.13**	**68.07**	**68.09**

It can be noted that the mBERT has performed worse than its vanilla model. Average score of mBERT was 64.61 whereas vanilla BERT had 65.37 and the inverse happened in case of DistilBERT and mDistilBERT. Our ensemble model achieved a better result than what all the other individual models had achieved. It maintained the highest accuracy in all the folds with an average accuracy of 68.07.

V. CONCLUSION AND FUTURE SCOPES

In this research, we have developed a gender-specific cyberbullying detection system for the under-resourced Bangla language. We have created the GenDisc dataset, and it currently consists of around 2600 data samples. In order to train a gender-based discriminating text classifier, we first constructed a framework that used four different models, and then we applied an ensembling approach to those four model. Our experiments have produced an accuracy of 68 percent.

This study paves the way for future research to examine the identification of cyberbullying in the Bangla language. In this day and age, every decision made by any model should not simply be a black box, which is why explainability is such an important factor to take into account while working on future work. The model used to make the prediction should be expalinable since cyberbullying can have long-lasting effects on both the person being bullied and the bully if they are caught.

REFERENCES

[1] R. M. Kowalski and G. W. Giumetti, "Bullying in the digital age," in *Cybercrime and its victims*. Routledge, 2017, pp. 167–186.
[2] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018.
[3] T. G. David Fidjeland, "Youths call for continued guidance to tackle online bullying amid increased internet use," 2021.
[4] M. R. Karim, S. K. Dey, T. Islam, S. Sarker, M. H. Menon, K. Hossain, M. A. Hossain, and S. Decker, "Deephateexplainer: 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.
[5] P. Chakraborty and M. H. Seddiqui, "Threat and abusive language detection on social media in bengali language," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*. IEEE, 2019, pp. 1–6.
[6] S. Tomkins, L. Getoor, Y. Chen, and Y. Zhang, "A socio-linguistic model for cyberbullying detection," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 53–60.
[7] R. Cao, R. K.-W. Lee, and T.-A. Hoang, "Deephate: Hate speech detection via multi-faceted text representations," in *12th ACM Conference on Web Science*, 2020, pp. 11–20.
[8] T. X. Moy, M. Raheem, and R. Logeswaran, "Hate speech detection in english and non-english languages: A review of techniques and challenges," *Technology*, 2021.
[9] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE access*, vol. 6, pp. 13 825–13 835, 2018.
[10] 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.
[11] S. Paul and S. Saha, "Cyberbert: Bert for cyberbullying identification," *Multimedia Systems*, pp. 1–8, 2020.
[12] F. Elsafoury, S. Katsigiannis, S. R. Wilson, and N. Ramzan, "Does bert pay attention to cyberbullying?" in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1900–1904.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.