Islamic University of Technology (IUT)

Department of Computer Science and Engineering (CSE)

A model agnostic explainable approach for detecting Cyberbullying in Bangla language using transformer based models

Authors

Takia Mosharref Nobo, 170041008

Mostafa Galib, 170041028

Hasnain Karim Rabib, 170041040

Supervisor

Dr. Md. Azam Hossain

Assistant Professor, Department of CSE, IUT

A thesis submitted to the Department of CSE

in partial fulfillment of the requirements for the degree of B.Sc.

Engineering in CSE

Academic Year: 2020-2021

May 10, 2022

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by Takia Mosharref Nobo, Mostafa Galib and Hasnain Karim Rabib under the supervision of Dr. Md. Azam Hossain, Assistant Professor of Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Gazipur, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Authors:

_____

Takia Mosharref Nobo

Student ID: 170041008

_____

Mostafa Galib

Student ID: 170041028

_____

Hasnain Karim Rabib

Student ID: 170041040

Supervisor:

_____

Dr. Md. Azam Hossain

Assistant Professor

Department of Computer Science and Engineering

Islamic University of Technology

Acknowledgement

We would like to express our gratitude towards IUT authority for granting us the fund and providing assistance required to implement our proposed system. We are indebted to our supervisor, Dr. Md. Azam Hossain for providing us with insightful knowledge and guiding us at every stage of our journey. Finally, we would like to express our heartiest appreciation towards our family members for their continuous support, motivation, suggestions and help, without which we could not have achieved the scale of implementation that we have achieved.

Abstract

Almost every facet of social communication has changed as a result of the exponential growth of social media platforms usage. Meanwhile, evidence is accumulating that the rising usage of social networks in the digital realm has given rise to an unsettling problem that has resurrected in new contexts: cyberbullying. The majority of current cyberbullying detection research focuses on English texts. On the other hand, while being spoken by 230 million people globally and being rich in diversity, the Bengali language is under-resourced for natural language processing (NLP). Recently, there has been an alarming surge in the number of incidences of gender-based discrimination or sexual harassment expressed on social media sites. In this study, we presented the cyberbullying detection under different categories in low-resourced Bangla language using transformer based models. We created our own dataset on gender discrimination and appended it to another open-source Bangla dataset with 4 classes. In our proposed approach, we used five different models to train our augmented dataset, followed by an ensembling technique on those five models.Then we make the models explainable using model agnostic approaches. Finally, we compared the individual prediction accuracies with the ensembled prediction accuracies. While training the dataset, we followed the stratified k-fold cross validation technique. Our evaluations yield up to an Accuracy of 75% in cyberbullying detection on emsembling.

Keywords— Cyberbullying, Transformer-based, Explainability, Bangla-text

Contents

List of Tables

List of Figures

1 Introduction

1.1 Overview

Young people have completely embraced the internet as a tool for socializing and communication [1]. Micro-blogging systems and social networking sites have risen tremendously in recent years, allowing their users to express themselves [2]. At the same time, they have encouraged anti-social behavior [3], online harassment, and, in particular, sexual discrimination [4] [5]. The confidentiality and mobility provided by such mediums have bred and spread cyberbullying [6]. In fact, the increased usage of social networks in the digital realm has given rise to cyberbullying reappearing in new circumstances [1]. Bullying was once thought to be a face-to-face interaction between children and adolescents in schoolyards, but it has now spread into cyberspace piling up over a range of categories. It is defined as an aggressive, intentional act committed by a group or individual against a victim through the use of electronic forms of communication (e.g., email and chat rooms) on a repeated or ongoing basis. In fact, a survey in 2021 says 14 percent of youths experienced online bullying at least once a week or more and that more than 2/3 people said online bullying is a serious problem for youths today [7]. One of the troublesome categories is the speech depicting gender-based discrimination or online sexual harassment.

When compared to physical bullying, cyberbullying can have more profound and long-lasting consequences. Online materials have a larger readership and disseminate quickly. There is also the force of the printed word, as well as the persistence and endurance of internet materials. The intended victim and onlookers can read what the bully has said over and over again in the instance of cyberbullying through text. Bullying has been linked to sadness, low self-esteem, and even suicide among youths [8].

While the World wide web originated as mostly an English phenomenon, it today contains texts in hundreds of languages. Languages are becoming obsolete at an astounding rate in the everyday world and may lose more than 50% of their linguistic variety by 2100. Some languages are under-resourced since they are low-density languages in terms of the number of people who speak them. The number of live languages used throughout the world is around 7,099, and this figure is constantly changing. One-third of these

languages are currently endangered, with fewer than 1,000 speakers left in many cases. In the meantime, just 23 languages are spoken by more than half of the world's population. Bangla is a rich and diverse language spoken in Bangladesh which is the second most widely spoken dialect in India, and the seventh most widely spoken language in the world, with about 230 million native speakers.bangla.

Although the perpetrators of cyberbullying should be held accountable, once proven guilty of such acts people have had to go to jail [9]which is why we need to be able to explain the decisions made by machines because any wrong decision can have serious consequences and repercussions. The explainability aspect is where we have tried making it explainable regardless of the model used. Lime and Shap have been used in this regard.

- We have made our own dataset of bullying expressions in Bangla language which consists of comments stemming from sexual harassment to gender discrimination. We further augmented it to one publicly available dataset.
- We have made the decisions made by the transformer models explainable since it is crucial for characterizing model accuracy, fairness, transparency and outcomes in AI powered decision making. Our main goal was to make it model agnostic so as to offer a generalist method using techniques like LIME and SHAP.

1.2 Problem statement

The primary goal of a good cyberbullying detection technique in a social platform is to prevent or at least reduce cyberbullying instances. It can be used to aid and simplify the work of monitoring internet environments. It is common to have a moderator, particularly in forums, however, they are unable to read all of the entries in these fora due to the large number of them. From a sociological standpoint, cyberbullying is a well-studied subject, yet most of the studies have been done using English texts. Even more so, it gets difficult since people are using Romanized Bangla texts besides just Bangla texts.

Now, there are impending effects of cyberbullying regardless of the language used for example, developing social anxiety, depression, self-harm or even eating disorder. Thus it has become imperative to bring about solutions to this concern. Therefore, our problem statement stands as such that hardly any work making cyberbullying detection explainable has been done, with an even fewer attempt made especially in Bangla language.

1.3 Motivation and Scope of Research

More than 90% of Bangladesh's 80.83 million Internet users[10] use Facebook, with the overwhelming number being young, insecure, and anxious for protection. However, due to a lack of a substantial number of annotated corpora, named dictionaries, and morphological analyzers[11], there has been relatively little research on Bangla text for social media monitoring systems, which necessitates in-depth investigation from Bangladesh's standpoint. In this day and age of social media, cyberbullying is a sensitive topic with far-reaching ramifications[12]. Bangla was not given the opportunity to be properly investigated because it was a language with limited resources. Limited datasets and a paucity of research on cyberbullying in Bangla make it an excellent option for further study.

To simply put the motivation of our work, we can say there are scopes-

- Firstly, cyberbullying has become a major issue in our social life.
- Secondly, there is a lot of scope for research in this field
- Thirdly, making a contribution in the under-resource Bangla language

- Finally, making the ML models explainable ie in a way that makes sense to a human being in an acceptable level

1.4 Challenges in Cyberbullying Detection

The fight over hate speech regulation is still going on. It is still unclear if legal measures or other approaches (like counter-speech or education) are the best responses. The obvious impact of hate speech makes its discovery vital, regardless of the means of combating it. The volume of content created online, especially on social media, as well as the psychological strain of manual moderation underscore the need for automatic detection of provocative and hostile content. This has become vital now more than ever, but there are certain challenges of cyberbullying detection that can be categorized into the following ways-

Evaluation Criteria

The problem of distinguishing hateful and/or objectionable speech automatically, especially in social media, has many layers. Some of these issues can be traced back to keyword-based techniques' inadequacies. In addition, many expressions are not inherently offensive, but they might be when used in the wrong context. One suggestion for reducing bias is to actively prepare annotators for it.

Consistent data availability

An issue arises as a result of this, namely the availability (or lack thereof) of consistently labeled data. One aspect contributing to the problem is the lack of a globally acknowledged definition of hate speech (a statement on which many newspapers agree), let alone one that is useful. A United Nations report can be used as a definition, but we would argue that it fails to meet the criteria of being a universally acknowledged productive definition on multiple points.

Imbalanced data

Another issue to address here is data that is unbalanced. While the propagation of hostile and offensive content on social media is a big concern, it is nevertheless true that this makes up a small percentage of all content. Hate speech corpora are also affected by this

imbalance.

Cross-domain data source

Not every social media platform has a robust interaction mechanism. Furthermore, it is potentially delusional to think that social media platforms such as Facebook (for which, in a perfect world, all of the aforementioned sources of information are available) will continue to be the main communication medium. Millennials have recently gravitated for more direct means of communication like WhatsApp, Snapchat, or media-focused means like Instagram, and TikTok. This shift implies more personal and less well off environments in which data can be accessed (creating even more scarcity), and that further advancement in the field necessitates a critical assessment of current use of available features, as well as ways of improving overall cross-domain generalization.

1.4.1 Research Challenges for Bangla Language

- Bangla being the under resourced language, there has not been enough work done in the cyberbullying detection in this language. As a result there is very little resource to go by when looking to perform research in this sector.
- There has not been any work done in the gender discrimination aspect of cyberbullying in Bangla language. Given a lot of hate online tends to be very gender specific this was a missed opportunity on which we tried to capitalize.
- There are not any standard datasets in Bangla which contains enough data regarding cyberbullying speech. This poses a great challenge for doing research in this sector. The datasets also do not encompass a wide variety of classes. Even though some datasets may have classes pertaining to political or personal, there is not any which covers the gender discrimination of cyberbullying.
- Performance may vary due to linguistic distinctions between English and non-English contents, as well as the study population's social and emotional behavior.

1.5 Thesis Outline

In Chapter 1 we have discussed our overall study in a precise and comprehensive manner. Chapter 2 and 3 deals with the necessary literature review and related work for our study and its development so far. We have mentioned the skeleton of our proposed approach methodology and also diagrams to provide a visual insight of the working procedure of our work for **GenDisc** in chapter 4 and for **CyberBullDetector** in chapter 5. Lastly, Chapter 6 contains the conclusion and future work. The final segment of this study contains all the references used.

2 Background Study

2.1 What is Cyberbullying

Cyberbullying is bullying done through the use of digital technologies. It can happen on social networking sites, messaging platforms, gaming platforms, and mobile phones. It is repeated behavior aimed at frightening, antagonizing, or ridiculing those who are attacked.[13] [14] Examples include:

- spreading lies about or posting embarrassing photos or videos on social media platforms
- sending hurtful,abusive or threatening messages, images or videos via messaging platforms
- impersonating someone and sending mean messages to others on their behalf or through fake accounts

Face-to-face bullying and cyberbullying can often happen alongside each other. But cyberbullying leaves a digital footprint - a record that can prove useful and provide evidence to help stop the abuse. Cyberbullying is a term that is frequently used to characterize a wide range of online abuse, including harassment, doxing, reputation attacks, and so on. By originating or engaging in online hate campaigns, the offender uses technology such as computers, consoles, cell phones, and/or any other device with internet or social media access to harass, stalk, or abuse another person. Although the majority of media attention portrays cyberbullying as a social media problem, it is also a major issue in the video game industry.

Cyberbully victims frequently have no idea who is behind the profiles that are harassing them. When random strangers become informed of cyberbullying, they fall into a 'mob mentality,' contributing to and reinforcing the bullying instead of aiding the victim. Online abuse is no longer limited to a single demographic; anyone can become a victim of cyberbullying in some fashion.

2.2 Types of Cyberbullying

There are many ways[15] that someone can fall victim to or experience cyberbullying when using technology and the internet:

- **Harassment** – When someone is harassed online, they receive a series of nasty messages or attempts to contact them from a single individual or a group of people. People can be harassed on social media, their cell phone (texting and calling), and their email. The majority of the contact the victim receives will be malicious or threatening.

- **Doxing** – Doxing occurs when someone or a group of people publishes another person's personal information, such as their residential address, cell phone number, or place of employment, on social networking sites or internet forums without their permission. Doxing can make the victim uneasy and have a negative impact on their mental health.

- **Cyberstalking** – Cyberstalking is similar to harassment in that it involves the offender making repeated attempts to contact the victim; but, unlike harassment, people are more likely to cyberstalk another person because they have strong feelings for that person, whether positive or negative. A cyberstalker is more likely to extend their stalking to the real world.

- **Swatting** – When someone calls 911 to report harmful happenings at a specific address, this is known as swatting. When armed security units come into their home or office building, some strike others with the goal of generating terror and dread. Swatting is more common in online gaming communities.

- **Corporate attacks** – In the corporate sector, attacks can be employed to flood a website with material in order to take it down and render it useless. Corporate attacks can erode public trust, harming businesses' reputations and, in some cases, causing them to fail.

- **Account hacking** – Hackers can gain access to a victim's social media profiles and send abusive or destructive messages. This is especially harmful to brands and public persons.

- **False profiles** – Fraud social media profiles can be created with the goal to harm a person's or a company's reputation. This is simply accomplished by using publicly available photographs of the target and making the profile appear as genuine as possible.

- **Slut shaming** – Slut shaming occurs when someone is dubbed a "slut" for something they have done in the past or even for how they dress. When somebody is sexting some other person and their photographs or conversations become public, this type of cyberbullying is common. Slut shame is more common among young individuals and teenagers, but it can happen to anyone.

2.3 Reasons why people bully others online

There are many reasons[15] that someone might choose to cyberbully another person. Some of the most common reasons are:

- **They have been cyberbullied themselves** – Because they have experienced cyberbullying, someone may choose to cyberbully another person. They may believe it is OK to treat others in this manner, or they may believe it is the only way to express their own sorrow.

- **To fit in** –Someone who witnesses another person being cyberbullied by a group of people may believe that by joining, they would 'fit in' or make new friends.

- **Home life** – It's possible that the culprit is having a rough home life and is projecting his or her anger and frustration onto someone else. Most of the time, this occurs because the cyberbully has no one to talk to about their problems.

- **Power** – Someone may decide to cyberbully so that they feel powerful and in control of a situation.

- **Jealously** – One of the most common motivations for cyberbullying, especially among teenagers and young people, is jealousy. Teenage years can be stressful since young people are developing themselves and may be self-conscious about their appearance. Because they are insecure, they may compare themselves to their classmates, which can lead to cyberbullying and harassment based on jealousy.

- **Cyberbullying and video games** – Over the last few years, online gaming has exploded. As a result of this boom, more online gamers are reporting toxicity and abuse while gaming. Online gamers can utilize a microphone to interact with other players, which can be used to encourage teamwork, develop friendships, and improve the overall gaming experience. Some players take advantage of this technology and utilize it to verbally or text/message abuse other gamers.

2.4 Psychological analysis of cyberbullying

Regardless of age, the multiple psychological ramifications can be devastating to victims, and it appears that no one is immune to the agony it produces. Children and teenagers, on the other hand, are particularly vulnerable and susceptible since they are still learning to regulate their emotions and responses to social interactions.[16]

Cyberbullying can result in crippling dread, low self-esteem, social isolation, and poor academic achievement. It can also make it difficult to create healthy relationships, and victims may experience severe post-traumatic stress, anxiety, and depression symptoms.

Young victims are roughly twice as prone as their classmates to ponder suicide. Many young victims injure themselves by slashing, head banging, or even punching themselves. They are also more likely to turn to substance usage to cope with their psychological distress. Between 2007 and 2016, the rate of cyberbullying among teenagers roughly doubled. According to a 2018 research, 59 percent of American teenagers have been bullied or harassed online. That's an incredible figure.

According to studies, the most common cause of cyberbullying is the breakdown of personal relationships as a result of breakups or unsolved problems. Certain groups are more vulnerable than others and are regularly attacked. Shy and socially uncomfortable students, overweight children, and children from low-income households are among them. Name-calling, propagating false stories, transmitting sexually explicit photographs and messages, cyberstalking, physical threats, and illegal sharing of personal images and information without consent are all examples of online abuse.

2.5 Cyberbullying in different languages

Many studies have provided ways for identifying cyberbullying in the English language, but just a handful have done so for other languages. The majority of commonly used methods for automatically detecting cyberbullying rely on English text.

However, Asian countries such as Bangladesh, India, China, Japan, and South Korea have a huge number of mobile device users. In Bangladesh, for example, there are 52.58 million internet users who are particularly active on social media platforms and use Bangla language or mixed Bangla-English. Because of the overwhelming volume, an automatic cyberbullying detection mechanism in other languages is required.

3 Literature Review

For a long time, researchers in domains such as data mining, information retrieval, and natural language processing (NLP) have been focused on the automatic detection of hate speech. There has been a spike in interest in this sector as a result of the growth of social media and social platforms. We'll look at the latest studies on detecting hate speech in social media text content in this part.

3.1 Recent work on English Language

S. Paul and S. Saha. 2020. CyberBERT: BERT for cyberbullying identification Summary:

Sayanta Paul et al. presented a novel use of BERT for detecting cyberbullying. [17] Cyberbullying is a recurring act that is aggressive in nature that is carried out on social networking sites such as Facebook, Instagram, Twitter, and others. BERT (Bidirectional Encoder Representations from Transformers), a cutting-edge pre-training language model, has achieved outstanding results in a variety of language comprehension tests. They introduced a novel use of BERT for cyberbullying detection in this research. In three real-world corpora, a simple classification model utilizing BERT was able to produce state-of-the-art results: Formspring (12k posts), Twitter (16k posts), and Wikipedia (100k posts). In compared to slot-gated or attention-based deep neural network models, experimental results show that their suggested model delivers significant gains over prior studies.

Limitations:

They compared their results with the state-of-the-art works constrained to their dataset only which although is adequate but it is not enough and there is a possibility that there might be over-fitting here which they did not address.

Fatma Elsafoury et al. 2021. Does BERT pay attention to cyber-bullying Summary:

Fatma Elsafoury et al. investigated the application of BERT for cyberbullying detection across a variety of datasets, attempting to explain its success by examining its attention

weights and gradient-based feature importance scores for textual and linguistic aspects. [18] They investigated the usage of BERT for detecting Cyberbullying on a variety of datasets and attempted to explain its success by looking at its attention weights and gradient-based feature attention scores for textual and linguistic features. The results reveal that attention weights are unrelated to feature importance scores and so do not explain the model's performance.

Contribution:

BERT outperformed other commonly used DL models on multiple cyberbullying-related datasets.

Limitations:

The results demonstrated that attention weights do not explain fine-tuned BERT's performance, and that its effectiveness is attributable to the datasets' reliance on syntactical biases.

L. Bacco et al. 2021. Explainable sentiment analysis: A hierarchical transformer-based extractive summarization approach Summary:

Luca Bacco offered two distinct transformer-based approaches [19] for performing sentiment analysis while extracting the most significant (in terms of the model conclusion) lines to generate a summary as the output explanation. **Contribution:**

1st method placed 2 transformers in cascade and leveraged the attention weights. 2nd method employed a single transformer to classify the single sentences and then combined the probability scores of each to perform the classification Both proposed models achieved good classification results, not so far from the SOTA works on IMDB dataset.

Limitations:

The explainability is not explored at a fine granular level rather only a summary which shows that there is more scope to explore as far as explainability is concerned.

3.2 Recent work on Arabic Language

- Guellil et.al(2021) [20] used both machine learning and deep learning techniques in their research. We employed the SGD Classifier (SGD), RandomForest (RF), Logistic Regression (LR), and Gaussian NB (GNB). Deep learning algorithms such as CNN and LSTM were used. For feature extraction, Word2vec and Fast-Text

were utilized. The balanced corpus did well.

- Alsafari et. al(2020) [21] created an Arabic corpus using Twitter data and labeled the dataset using a "three-level labeling approach." Later on, they applied a variety of classification models and feature extraction approaches.

- Aljarah et.al(2020) compiled a collection of tweets on topics such as racism, journalism, sports orientation, terrorism, and Islam. Then, using the Decision Tree algorithm, Naive Bayes, Support Vector Machine (SVM), and Random Forest (RF), the labelled dataset was trained. In this dataset, the Random Forest classifier worked admirably.

- Faris et. al(2020) [22] gathered data from Twitter and used a word embedding approach to identify different traits. For this challenge, CNN and LSTM were integrated. In terms of accuracy, this method produced the greatest results when categorizing tweets as hate or non-hate.

3.3 Recent work on Indonesian Language

- Alfina et. al [23] in a study employed machine learning algorithms to construct a dataset that encompassed hateful speech in multiple categories such as sex, religion, and others. The categorization was accomplished utilizing techniques such as Nave Bayes, RandomForest Decision Tree, Logistic Regression, and Support Vector Machine, and the performance of multiple approaches for automated hateful text identification was examined. With a 93.5 percent accuracy, the Random Forest Decision Tree algorithm performed well.

- In another study pratiwi et. al [24] utilized FastText as the classifier in an experiment on Instagram comments for identifying nasty messages. The precision was 65.7 percent.

- Fauzi et. al [25] developed a new ensemble algorithm for inappropriate text detection. K-Nearest Neighbours, Support Vector Machines, Naive Bayes, Maximum Entropy, and Random Forest were the first five classification algorithms used. On the data set, two ensemble approaches (soft voting and hard voting) were applied to improve accuracy. The analysis revealed that the ensemble technique might

improve accuracy, with soft voting achieving the highest accuracy (F1score (unbalanced dataset)=79.8% and F1 score (balanced dataset) =84.7%).

3.4 Recent work on Hindi Language

Among under-resourced languages, Hindi and some other languages of the South Asian region are also in works by quite a few researchers. Since a lot of the existing studies mostly focus only on cyberbullying detection in the English language, recent work emerges on other languages like Hindi.

Aditya Bohra et al. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection

Aditya et.al were the first to detect hate speech in Hindi-English tweets. [26]. They gathered 4575 Hindi-English tweets, which were annotated by two linguists and validated using Cohen's Kappa coefficient to calculate inter annotation agreement. After that, features including punctuation count, emoticon count, word n-gram, character n-gram, and word2vec of lexicon words were retrieved from the data. These characteristics combined to form a fat matrix. To lower the size of the feature derived fat matrix, they applied a dimensionality reduction technique like chi square. SVM and Random forest were employed for classification, with 71.7 percent and 66.7 percent accuracy, respectively.

Santosh, T. Y. S. S., and K. V. S. Aravind. 2019. Hate Speech Detection in Hindi-English Code-Mixed Social Media Text

Santosh et.al [27] completed the second study using this data. They tested two deep learning models: a sub-word level LSTM model and a hierarchical LSTM model with attention based on ph onemic sub-words. The accuracy of the LSTM model at the sub-word level was 69.8%. The accuracy of a hierarchical LSTM model with attention based on phonemic sub-words was 66.6 percent. They also compared the performance of their model to that of previous research.

Shrikant Tarwani et al. 2019. Cyberbullying Detection in Hindi-English Code-Mixed Language Using Sentiment Classification

The primary goal of this research [28] was to create a method for detecting cyberbullying

in the Hindi-English code-mixed language (Hinglish), which is widely used in India. Because the Hinglish dataset was unavailable, the authors constructed the Hinglish Cyberbullying Comments (HCC) labeled dataset, which includes comments from social media sites like Instagram and YouTube. They also created eight different sentiment categorization machine learning models to detect cyberbullying situations automatically. These models are evaluated using performance criteria such as accuracy, precision, recall, and f1 score. Finally, a hybrid model is created using the top performers of these eight baseline classifiers, which have an accuracy of 80.26 percent and a f1-score of 82.96 percent.

3.5 Recent work on Under-resourced Bangla Language

Recently, a few studies have been done on Bangla. The absence of systematic text collecting methods, annotated corpora, name dictionaries, morphological analyzers, and overall research perspectives makes it difficult to delve into this area.

Shahin Akhter et al. 2018. Social media bullying detection using machine learning on Bangla text

In this work [29], authors proposed the use of Machine Learning 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. A set of Bangla text has been collected from available social media platforms and labeled as either bullied or not bullied for training different machine learning- based classification models. Cross-validation results of the models indicate that a support vector machine-based algorithm achieves superior performance on Bangla text with a detection accuracy of 97%. However, it lacked the use of a larger dataset.

Puja Chakraborty and Md Hanif Seddiqui. 2019. Threat and abusive language detection on social media in Bengali language

Chakraborty et al [10] proposed to build an automatic system using Machine Learning and Natural Language Processing 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). Among three algorithms, SVM with linear kernel performed best with 78% accuracy. However, they noticed that the result of the SVM classifier with RBF kernel fluctuated to a large degree so they have a scope of improvement.

Alvi Ishmam and Sadia Sharmin. 2019. Hateful Speech Detection in Public Facebook Pagesfor the Bengali Language

Ishmam et.al [30] labelled 5,126 comments and used multi label annotation scheme. The created corpus was the biggest and initial contribution to this area in the Bengali language.Various algorithms comparing the per-formance were employed.Random Forest performed wellwith the accuracy of 52.20%.

Md Rezaul Karim et al. 2020. DeepHateExplainer: Explainable Hate Speech Detection in Under-resourced Bengali Language

For the under-resourced Bengali language, Karim et al. [12]provide DeepHateExplainer, an explainable solution to hate speech identification. With an F1 score of 88%, DeepHateExplainer is able to recognize a wide range of hate speech, beating various ML and DNN baselines. However, the approach is limited by a scarcity of labeled data available during training. As a result, there is a considerable likelihood of overfitting.

Kumar et al. 2021. Aggressive and Offensive Language Identification in Hindi, Bangla, and English: AComparative Study

Kumar et.al [31] developed classifiers for hateful languge identificatin code mixed dataset. The dataset is available in HASOC-2020. They have used BERT and SVM for classification. Exclusive divisions of BERT includes ALBERT and DistilBERT for growing the classifiers.The highest accuracy performed with F-score between 0.70 and 0.80.

3.6 Datasets on Bangla Hate Speech text

There aren't many publicly available datasets on Bangla hate speech. But in the recent times, there have been some research on Bangla hate speech and so a small number of datasets have been created and are public. A few of them have been mentioned in the table 1.

Table 1: Datasets on Bangla Hate Speech text

Paper	**Dataset Size + Text Type + Labels**	**Performance**	
[12] DeepHateExplainer: Explainable Hate Speech Detection in Under -resourced Bengali Language [2020]	6115 Bangla Text 4 labels	**Model**	**Score (Pr.)**
		Bangla_BERT	86.00
		mBERT-cased	84.00
		XML-RoBERTa	88.00
		mBERT-uncased	85.00
[32] Hate Speech detection in the Bengali language: A dataset and its baseline evaluation [2020]	30000 Bangla Text 7 labels	**Model**	**Score (Acc.)**
		SVM	87.50
		Word2Vec + LSTM	83.85
		Word2Vec + Bi-LSTM	81.52
		FastText + LSTM	84.30
		FastText + Bi-LSTM	86.55
		BengFastText + LSTM	81.00
		BengFastText +Bi-LSTM	80.44
[33] Bangla hate speech detection on social media using attention-based recurrent neural network [2020]	7425 Bangla Text 7 labels	**Model**	**Score (Acc.)**
		CNN + LSTM	74.00
		CNN + GRU	74.00
		CNN + attention	77.00
[34] Multilingual Offensive Language Identification for Low-resource Languages [2020]	4000 Bangla Text 3 labels	**Model**	**Score (F1)**
		XLM - R (TL)	84.00
		Risch and Krestel	82.00
		BERT - m (TL)	82.00
		XLM - R	82.00
		BERT - m	81.00
[35] Abusive content detection in transliterated Bengali-English social media corpus [2021]	3000 Transliterated Bangla Text 2 labels	**Model**	**Score (F1)**
		SVM	82.70
		LR	82.30
		Bi-LSTM	79.00
		RF	77.00

4 GenDisc: A Gender Discrimination based Cyberbullying Detection System in Under-resourced Bangla Language

4.1 Methodology

To confirm the practicality of our dataset and the model, we have performed a task of gender-based discrimination text detection. Our methods including dataset creation, preprocessing, model selection, tokenization, data splitting, training, ensembling and evaluation are discussed below -

4.1.1 GenDisc Dataset Creation:

Source selection:

The most commonly used social media platforms are Facebook, Youtube, Tiktok, Twitter etc. But for our dataset, we solely looked into Facebook, because it is indubitably the biggest hub for abusive and hateful comments and speeches. Moreover, as we are dealing with Bangla language only, it would be wise to go for Facebook, as Facebook has close to 47 million users from Bangladesh whereas Youtube has around 34 million, which is much lower compared to what Facebook has.

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 linked to gender.

Scraping:

There are various online tools for web scraping. Also, customized web scrapers can also be coded in Java or Python from scratch. But we have opted for the former as we have chosen Instant Data Scraper for collecting our data samples and creating our novel dataset. It is a free to use online scraping tool, which is easy to work with; especially for collecting data from Facebook.

Annotation:

One of the major tasks in dataset creation is annotating the data samples. It 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 ample amount of data samples which were hard to put into a specific category. And in that case, we needed expert 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 opposite were labelled as '0'.

Table 2: GenDisc: Data Distribution

Type	**Number of data samples**	**Percentage**
Non-Discriminatory	1421	55.40%
Discriminatory	1146	44.60%
Total	2567	100%

4.1.2 Pre-processing:

Our novel web-scraped dataset has undergone a series of standard steps of preprocessing to ensure that meaningful and useful words are being fed to the models while they are being trained. Getting rid of redundant, unnecessary and futile words as well as emojis and characters was the primary task of the whole data cleaning process, followed by the tasks of stopwords removal and data sample duplicity elimination in the dataset.

Getting rid of unnecessary words, characters and emojis:

Categories of unnecessary or redundant words:

1. Non-bangla word
2. Meaningless word
3. Numericals: Any 0-9 number were eliminated

Categories of unnecessary characters:

Removal of symbols:

The characters or symbols, mainly the punctuation marks, were removed. Moreover, parenthesis and meaningless single characters (if there were any) were also removed.

Removal of emojis:

Emoticons or emojis or any graphical items in the text were removed

Removal of stopwords:

Stopwords are essentially those words which do not provide useful information for making a decision regarding classifying a text. There is a long list of commonly used stopwords in Bangla language.

Customized Collection:

A customized collection of some common Bangla stopwords 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 stopwords in Bangla language.

Filtering:

Each and every data sample in our dataset was scanned for finding stopwords i.e. any word was found which is present in our list of stopwords, it was removed.

Duplicity elimination in the data samples:

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.

4.1.3 Data Augmentation:

As our dataset got shrunk to a quite smaller size after going through data preprocessing and data cleaning, it was not big and good enough for achieving decent or satisfactory results in our model experiments. As a result, we had to find ways to get around this issue and thus we ended up augmenting our dataset. There are many ways to augment datasets, but our method was n-word swapping.

N-word swapping:

The way in which this technique works is it randomly selects 'n' number of words from the data sample and switches their positions and in result of which we get a new data sample which supposedly has the same words and the same class, but a different arrangement of words or context. This technique helped us to double the size of our dataset and ensured that there was no change in the biases.

For our dataset, we selected n=2 for this technique.

Specialty of our dataset:

1. Hardly any Bangla dataset on gender based discrimination texts can be found online. So, our dataset includes the gender bias class that could be one of the firsts of its kind, if not the first i.e Novel.
2. Datasets created targeting a specific issue/task are very rare in Bangla language.
3. All the samples have been made anonymous

4.1.4 Models

With the advent of modern and advanced science, newer technologies have been emerging or are being proposed quite frequently nowadays for accomplishing a variety of some specific Artificial Intelligence based tasks in the fields of Computer vision, Natural language processing etc.. In the Nlp domain, in recent times, some revolutionary research has been conducted to come up with better architectures and models, which essentially have brought exemplary improvements in the performances of the task-specific nlp systems. Consequently, the world came across the 'Transformer' architecture.

Transformers

Transformer, a model which has a set of encoders and decoders as its building block and adopts the mechanism of self-attention to give equal importance to each part of the input data. Eventually, after the transformer model was introduced, various newer transformer-based models were proposed from time to time i.e. transformer xl, BERT etc. Each of these models touched State-of-the-art results in certain Nlp tasks.

4.1.5 Model selection

In the case of our task of binary text classification, we wanted to conduct a comparative or ablation study between a number of transformer based models to see how the models perform on our novel dataset in order to justify the relevance and acceptance of our Gendisc dataset. Moreover, as our dataset is a complete Bangla dataset, we also opted for multilingual models along with the vanilla models. Hence, we have chosen 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 four individual models.

BERT

BERT (Bidirectional Encoder Representations from Transformers) [36], a language representation model, is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. BERT doesn't contain a decoder; it is only the encoder part of the transformer. It was pretrained on close to 3300 million words from the internet (Wikipedia and Book corpus)

The pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering, text classification and language inference, without substantial task-specific architecture modifications. BERT obtains new state-of-the-art results on eleven natural language processing tasks.

For text classification, BERT uses [cls] token which is a special classification token and the tokens in the last hidden state of BERT are used for the classification task. Bert also uses positional and segment embeddings along with the tokens to get more context from the data. BERT base has around 110 million trainable parameters.

Table 3: GenDisc: Summary of selected models

Model	Layers of transformer block	Hidden States	Trainable parameters	Attention heads
BERT - base cased	12	768	110	12
mBERT base cased	12	768	178	12
DsitilBERT base cased	12	768	66	12
mDistilBERT base cased	6	768	134	12

Multilingual - BERT

Multilingual BERT is the same as BERT, but it was pretrained on text from multiple languages. It has been pretrained on Wikipedia words and the shared vocabulary across 104 languages. It performs better than BERT in multilingual tasks.

On our Bangla dataset, multilingual BERT is believed to perform better than vanilla BERT.

DistilBERT

DistilBERT [37] is a variation of BERT. It has been created based on BERT architecture. It is fast, cheap and lightweight. It has 40 percent less parameters than BERT, but it preserves 95 percent of BERT's performance. It performs well on small datasets.

Our Gendisc dataset is relatively small and so DistilBERT is a good option to validate our dataset.

Multilingual - DistilBERT

Multilingual DistilBERT is similar to DistilBERT, but it has been pre-trained on texts from multiple languages just like multilingual BERT.

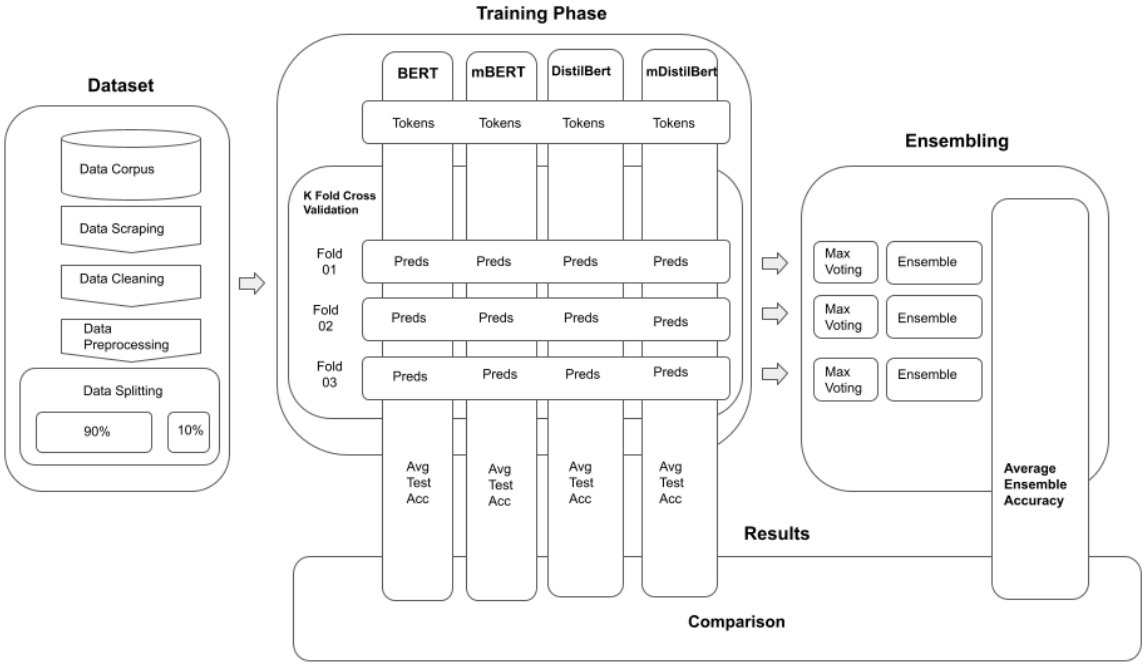

Figure 1: Gendisc: Methodology

4.1.6 Training

After preprocessing the dataset and selecting the models to work with, we moved towards the training phase.

Transfer learning:

We took our pretrained models and trained them on our novel dataset for a specific task of binary text classification.

Data Splitting:

With the help of experts, we annotated the dataset into labels containing 1's and 0's. Here basically 1's represent comments that are downright offensive and insinuate gender discrimination and vice versa for the 0's. We split the whole dataset into two parts (train and test) maintaining a ratio of (90:10) by using the train-test-split function. The first

portion is the train data and the latter one is the test data.

To make our models perform better in case of unseen data, we have implemented the K-fold cross-validation technique on our dataset. Cross validation makes the model use a different fold of validation set for each iteration.

K-fold cross validation:

In this technique, the data-set has been split into k numbers of subsets (known as folds). Then we perform training on all the subsets but leave one subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purposes each time. We set the number of folds to 3 i.e k=3 and the number of epochs for each fold to 15 i.e epochs = 15. In each epoch, we calculated the training accuracy and loss along with validation accuracy and loss.

CLS Tokenization:

BERT and its variations use special [cls] tokens for classification. Each model has its own tokenizer that we had to import from the transformer library from the hugging face. The tokenizers were used to generate unique token ids and the attention mask for each word in the input dataset; which were eventually fed to our model while training.

Weight initialization:

The weights are initialized 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 it is a constructor function and is called by default when the model is created.

Adam Optimizer:

We used the optimizer algorithm "Adam" which combines the best properties of the two algorithms - AdaGrad and RMSProp.

Sigmoid Activation function:

Since it is a binary text classification, we have used Sigmoid function. Because Sigmoid function is a mathematical function having a characteristic "S"-shaped curve or sigmoid

curve. The input to the function is transformed into a value between 0.0 and 1.0

Cross Entropy Loss function:

For calculating the loss and adjusting the weights, we used the Cross Entropy loss function. Cross-entropy is a measure from the field of information theory, building upon entropy and generally calculating the difference between two probability distributions.

Learning rate:

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.

4.1.7 Experimental Setup

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.

Hyper-parameter tuning:

We conducted uncountable number 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 4.

Table 4: GenDisc: Experimental Settings

Model	**Batch Size**	**L.R.**	**Epochs**	**Max Length**
Bert	16	5e-6	15	150
mBert	16	5e-6	15	150
DistilBert	16	1.4e-6	15	150
mDistilBert	16	1.4e-6	15	150

The number of epochs for each training was set to 15 i.e. epochs = 20. The batch size was set to 16 i.e. batch size = 16, because the models could not manage to learn from

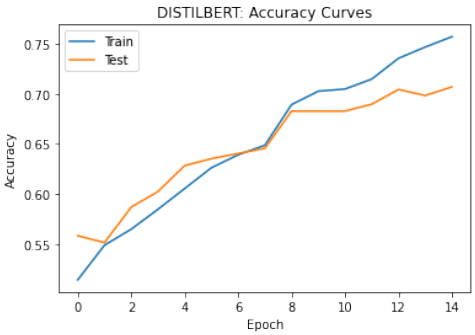

Figure 2: GenDisc: DistilBERT - Train and Validation
Accuracy per epoch for fold 2

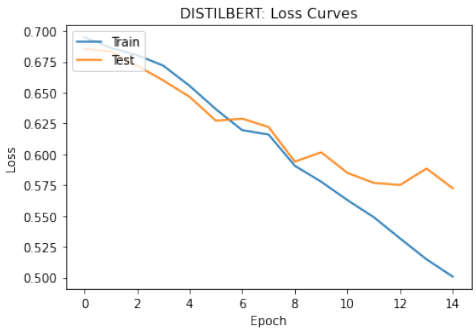

Figure 3: GenDisc: DistilBERT - Train and Validation Loss per epoch for fold 2

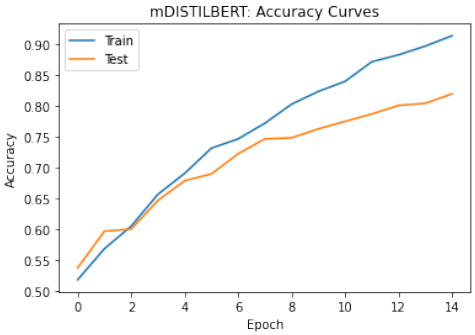

Figure 4: GenDisc: mDistilBERT - Train and Validation
Accuracy per epoch for fold 2

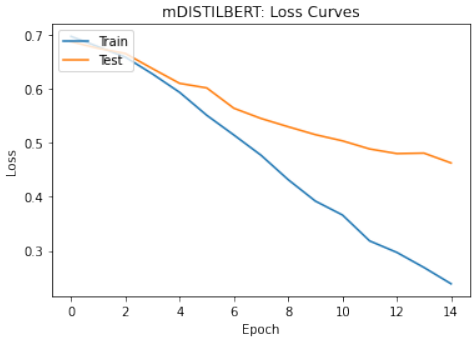

Figure 5: GenDisc: mDistilBERT - Train and Validation Loss per epoch for fold 2

larger batch sizes as it was taking up a lot of memory. Going through our dataset, it was seen that the text or data sample with the maximum length had a length less than 150. So, we set the max length for both tokenization and training to 150 i.e. max length = 150. The learning rates for BERT, mBERT, DistilBERT and mDistilBERT were taken as 5e-5, 5e-5, 1.35e-6 and 1.35e-6 respectively.

Evaluation

While training, each and every epoch, we evaluated the model by calculating the loss and accuracy of training and validation dataset; which helped to update the weights and get a smaller loss in the next epoch.

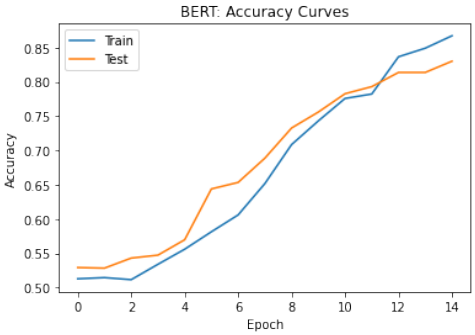

Figure 6: GenDisc: BERT - Train and Validation
Accuracy per epoch for fold 2

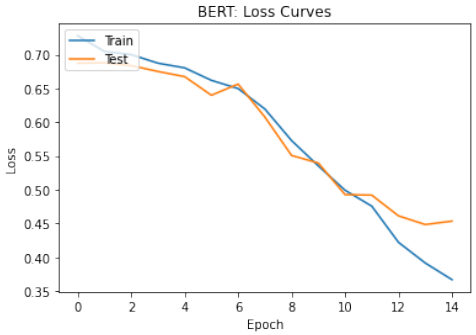

Figure 7: GenDisc: BERT - Train and Validation Loss per epoch for fold 2

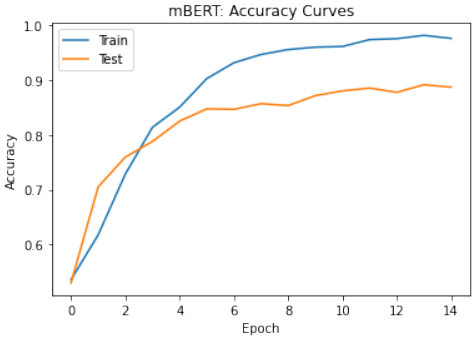

Figure 8: GenDisc: mBERT - Train and Validation
Accuracy per epoch for fold 2

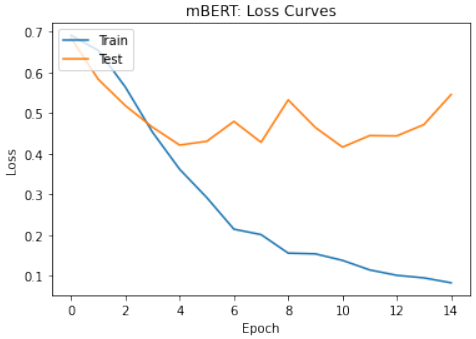

Figure 9: GenDisc: mBERT - Train and Validation Loss per epoch for fold 2

4.1.8 Results And Discussion

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. For each fold, we get a different set of predictions from all the four models.

Table 5: GenDisc: Test Accuracy for each fold

	Test Accuracy(%)			
	Bert	**mBert**	**DistilBert**	**mDistilBert**
Fold 0	60.71	66.88	66.88	67.21
Fold 1	62.88	66.8	62.99	63.31
Fold 2	64.61	68.51	64.94	66.23
Average	62.66	**67.42**	64.94	65.58

Cross validation score

For each model, we calculate the accuracy or 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.

Here the ratio of correct predictions to the total number of samples is taken as the score.

Observation

Here, we can see that the two multilingual modes have performed better than their vanilla models i.e average score of mBERT was 67.42 whereas vanilla BERT had 62.66. Similarly, mDistilBERT had an accuracy or score of 65.58 while the vanilla DistilBERT's score was close to 65. But overall, the best performance was given by the multilingual-BERT model.

4.1.9 Ensembling

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.

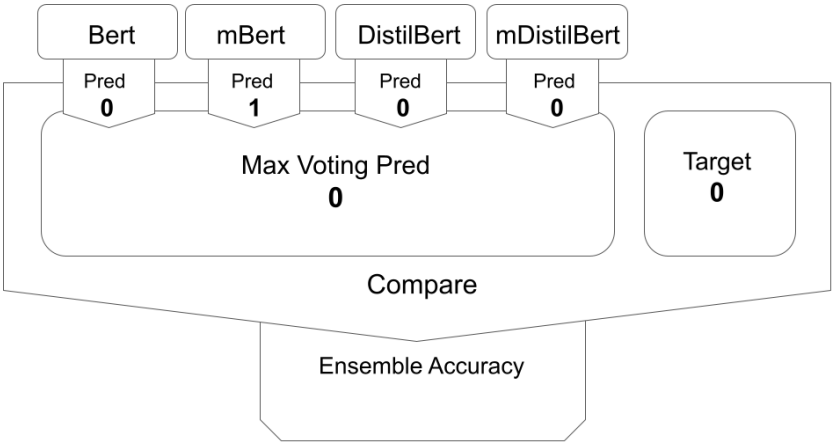

Figure 10: GenDisc: Ensemble Process

Table 6: GenDisc: Ensemble Accuracy for each fold

Fold	Ensemble Accuracy	Mcc Score	Precision	Recall	F_score
0	66.55	32.68	66.62	66.55	66.58
1	67.85	35.00	67.77	67.85	67.80
2	67.20	33.62	67.10	67.20	67.12
Average	**67.20**	**33.76**	**67.16**	**67.20**	**67.17**

Max voting method:

Max-voting is one of the simplest ways of combining predictions from multiple machine learning algorithms. 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.

4.1.10 Performance evaluation metrics

Evaluating on a single metric does not give the clear picture about the model's performance. So, we have evaluated the model on several performance measure metrics to get a more vivid and convenient idea about the performance of the model. Four metrics were used -

Accuracy

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall

It is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1 score

F1 Score is the weighted average of Precision and Recall.

Table 7: GenDisc: Performance Comparison

Fold	**BERT**	**mBERT**	**DistilBERT**	**mDistilBERT**	**Ensemble**
0	60.71	66.88	66.88	**67.21**	66.55
1	62.88	66.8	62.99	63.31	**67.85**
2	64.61	**68.51**	64.94	66.23	67.20
Average	**62.66**	**67.42**	**64.94**	**65.58**	**67.20**

4.2 Conclusion and Future Scopes

In this research, we have provided gender-based classification and detection under cyberbullying for under-resourced Bangla language using four different models to train a gender-based discriminatory text classifier, followed by an ensembling technique on those four models. Our dataset focusing on gender-bias or sexual harassment brings novelty in this field of study. However, our work has some future scopes.

4.2.1 Explainability

Implementing a way to making all the decisions taken by the Machine Learning model Explainable could be a major undertaking. It would be better to use a model agnostic explainable approach to better suit all the different kinds of models. Making it explainable is a major thing to consider because in this day and age, any decision taken by any model should not just be a black box. Since cyberbullying can have lasting effects on both the person and the offender if caught, the model used for the prediction should be able to explain as to how or why the model predicted that result. Thus, we need to be able to interpret the decisions made by the model and trace back the outputs to the inputs.

4.2.2 Further Augmentation of the dataset

Our dataset has only around close to 2600 data samples, which is a decent number but the accuracy and the results could change or get better if we there were more data samples. So, in future, more and more data samples can be added to our GenDisc dataset.

Using the dataset for other relevant tasks

We have used the dataset for the task of binary classification. But maybe it can be used for other relevant tasks as well. For example, gender discrimination is one kind of cyberbullying. So, our dataset can be merged with other Cyberbullying related Bangla datasets to accomplish a specific task.

5 CyberbullDetector: A model agnostic explainable approach to detect Cyberbullying in Bangla language using transformers based models

We have utilized our novel dataset 'GenDisc' in this section of our research. We have merged this with another dataset on Bangla hate speech called from the paper 'DeepHateExplainer' in order to perform the task of hate speech detection.

5.1 Our Proposed Approach

5.1.1 Dataset

In the paper 'DeepHateExplainer', the authors proposed a dataset on Bangla Hate speech for the task of hate speech detection.

DeepHateExplainer dataset:

In this dataset, there are four categories of hate speeches -

Training data distribution

Table 8: DeepHateExplainer: Training data distribution

Category	**Data samples**
Personal	2547
Geopolitical	1690
Religious	908
Political	727

They trained the three models XLM-Roberta, Bangla-bert and Bert-uncased on their dataset; and then performed an ensemble tehnque on their predictions. They were able to achieve an accuracy of 87 percent on the ensemble model.

Scope of further work in DeepHateExplainer

1. Here, their dataset had only four categories of hate speec whereas there are other important categories of hate speech such as gender discrimination, child abuse etc. which they were not able to propose or add in their dataset. So, there is a scope of introducing more categories into their dataset and eventually enlarging the same.
2. They have primarily used multilingual and cross-lingual models for training. So, there is a scope of training other transformer architecture based models on their dataset to check whether they perform better or worse.
3. They have followed the Layer-wise relevance propagation technique for making explainability, which is not model agnostic. So, there are scopes to work with model agnostic ways explainability technique.

Our contributions

1. We have tried to focus on almost all the scopes of further research or work in the 'DeepHateExplainer'.
2. We have added one more category called 'Gender' in their dataset.
3. We have trained a set of different transformer architecture based models. We have trained the five different models separately on the new dataset - vanilla BERT-cased, multilingual-BERT- cased, DistilBERT-cased, multilingual-DistilBERT-cased and XL-Net. Moreover, we have also performed an ensemble technique on the predictions of these five models.
4. We have followed the model agnostic explainability approach called 'SHAP' for making our models explainable.

CyberBullDetector dataset

From the 'GenDisc' dataset, we have taken all the data samples which are labeled as gender discriminatory and added them all to the 'DeepHateExplainer' dataset thus added a new category called 'Gender' in that dataset. And we called this new merged datset 'CyberBullDetector' dataset.

Table 9: CyberBullDetector: Training Dataset Distribution

Category	**Data samples**
Personal	2547
Geopolitical	1690
Religious	908
Political	727
Gender	795

After merging, pre-processing and cleaning, the training data distribution looked as follows -

The test data of 'DeepHateExplainer' had 1000 data samples. Hence, we added the 'Gender' category (200 data samples) to the test data as well.

5.1.2 Methodology

To compare the performances of different models on the new merged 'CyberBullDetector' dataset, we have performed the task of cyberbullying text detection. Our methods including pre-processing, model selection, tokenization, data splitting, training, ensembling, explainability and evaluation are discussed below -

5.1.3 Pre-processing

This task is similar to what we have done in the case of 'GenDisc' dataset pre-processing.

1. We removed the stop-word.
2. We removed the duplicity of samples.
3. We augmented the dataset following the process of 'n-word swapping'

5.1.4 Models

Along with the four models i.e BERT, multilingual-BERT, DistilBERT and multilingual-DistilBERT that we have used in the case of binary classification task in 'GenDisc', we

have also used 'XLNet' for the Cyberbullying classification task. 'XLNet' is an auto-regressive model which supposedly performs better than BERT in around 20 tasks such as sentiment analysis, question answering etc.

XLNet model

XLNet [38] is a generalized auto-regressive pre-training method.

It enables the learning of bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order.Due to its auto-regressive formulation, it overcomes the limitations of BERT which is neglecting the dependency between the masked positions and suffering from a pre-train-fine-tune discrepancy.

Furthermore, XLNet takes up ideas from Transformer-XL while pre-training, which is a state-of-the-art auto-regressive model.

Auto-regressive model

It specifies that the output variable depends linearly on its own previous values and on a stochastic term.

Since. 'XLNet' is supposed to perform better the BERT and its variations, it is a good option for us to train on our dataset and to compare its performance with the other variations of BERT model.

5.1.5 Training

After pre-processing our dataset and selecting the appropriate models, we trained those five models on the 'CyberBullDetector' dataset. Here too we conducted a K-fold cross-validation technique.

K-fold Cross-validation

We took k=2, and so the training data was split into 3 folds. So, all the five models were trained for 2 times; taking a different fold as the validation set in each training phase. Cross-validation gives us a better wy to asses the performance of our models, as in each training phase, the models are getting a different set of data for validating. Here, after each training, the model is tested on the test data and the accuracy is calculated. The average of the scores from each training for a model is taken as the cross-validation score for that model.

Tokenization

For the four BERT variation models, we used their respective tokenizers and it was the same case for XLNet as well. All the tokenizers were imported from the transformers library from hugging face.

Optimizer

The optimizer algorithm used for training the models was 'Adam' optimizer. It is better than AdaGrad and RMSProp as it combines the properties of both the algorithms.

Activation function

Since, it is not a binary classification problem, rather it is a multi-class classification problem, it is wiser to use the Softmax activation function. It is a mathematical function that converts a vector of numbers into a vector of probabilities, where the probabilities of each value are proportional to the relative scale of each value in the vector. It basically scales numbers/logits into probabilities.

Loss function

Here too, we used the cross-entropy loss function. The cross entropy function uses the following equation for multi-class classification respectively.

$$L = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{4}$$

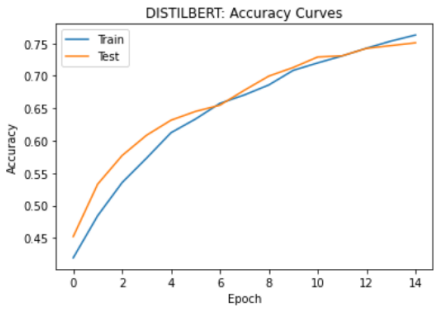

Figure 11: CyberBullDetector: DistilBERT - Train and Validation Accuracy per epoch for fold 2

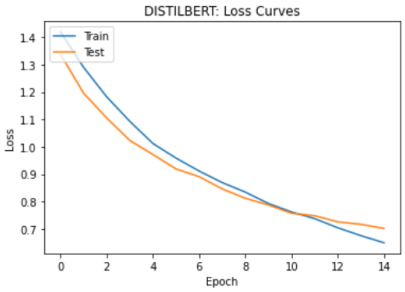

Figure 12: CyberBullDetector: DistilBERT - Train and Validation Loss per epoch for fold 2

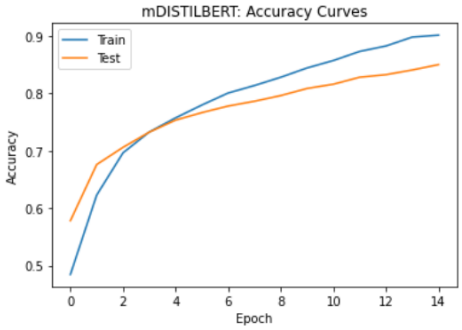

Figure 13: CyberBullDetector: mDistilBERT - Train and Validation
Accuracy per epoch for fold 2

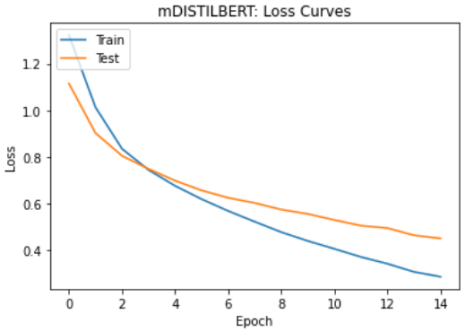

Figure 14: CyberBullDetector: mDistilBERT - Train and Validation Loss per epoch for fold 2

5.1.6 Experimental Settings

We took k=2 for the cross validation technique. And the number f epochs for each model was 15, while the batch size was 16. Each model was initialized with a different learning rate. A large number of experiments were conducted with different sets of learning rates and batch sizes. But the best results were found for a specific set of these hyperparameters.

5.1.7 Evaluation

While training, in each and every epoch, we evaluated the model's performance using the validation fold. And after the training phase, we evaluated the model using the test data. And for each fold, training and loss curves for each of the models were generated.

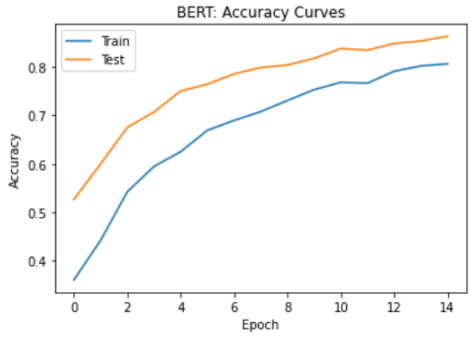

Figure 15: CyberBullDetector: BERT - Train and Validation Accuracy per epoch for fold 2

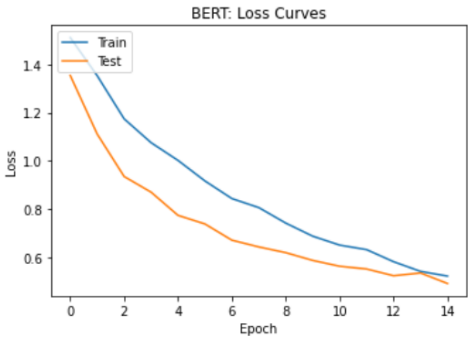

Figure 16: CyberBullDetector: BERT - Train and Validation Loss per epoch for fold 2

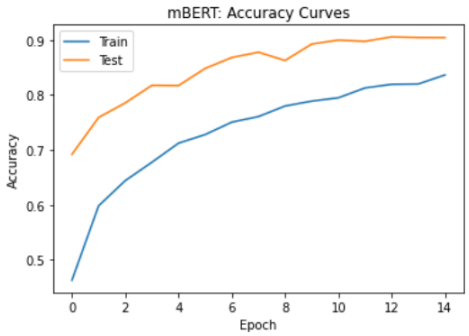

Figure 17: CyberBullDetector: mBERT - Train and Validation Accuracy per epoch for fold 2

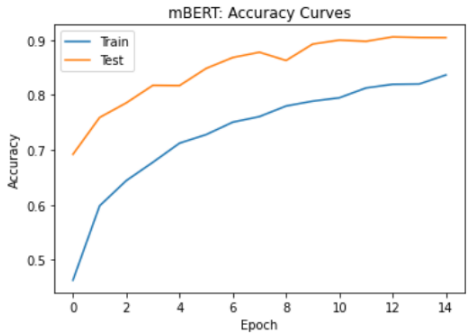

Figure 18: CyberBullDetector: mBERT - Train and Validation Loss per epoch for fold 2

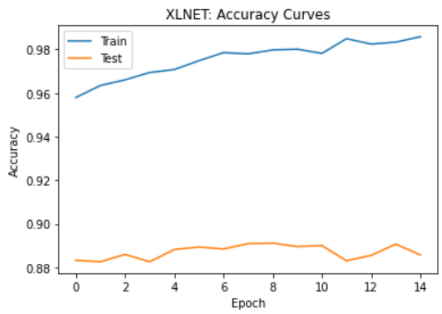

Figure 19: CyberBullDetector: XL-Net - Train and Validation Accuracy per epoch for fold 2

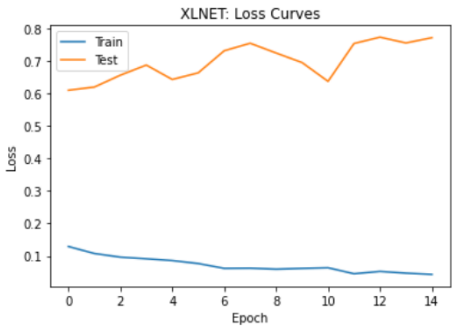

Figure 20: CyberBullDetector: XL-Net - Train and Validation Loss per epoch for fold 2

5.1.8 Results and Discussion

Results were calculated in the same way as we did for GenDisc. Here, after every fold, the accuracy of each of the model was calculated on test data. And then for each model, the accuracies from each fold were added and then averaged, which gave us the cross-validation score.

Observation

The results were similar to what we have got in the case of GenDisc. The multilingual BERT and DistilBERT models performed better than their vanilla models. But, contradictingly, although XLNet was supposed to outperform BERT, it didn't. Infact, the accuracy for the XLNet model was the lowest amongst all the five models.

Table 10: CyberBullDetector: Results

	XLNet	BERT	mBERT	DistilBERT	mDistilBERT	Ensemble
Fold 0	64.14	74.43	78.64	68.5	75.00	78.28
Fold 1	63.21	73.14	76.71	70.00	76.64	79.21
Average	63.67	73.85	77.67	69.24	75.82	**78.75**

Ensemble

After getting the predictions from each of the individual models. We followed an ensemble technique on the predictions. And it was seen that the accuracy for the ensemble model was surprisingly quite high, but it could not beat the multilingual-BERT model's accuracy.

5.1.9 Accuracy

There are several metrics for evaluating a model. There are multiple ways to generate an evaluation score. In our case, we calculated the Mcc, Precision, Recall and the F1 score.

5.1.10 Explainability

Explainability or interpretability is a concept that a ML model and its output can be explained in a way that makes sense to a human being at an acceptable level.

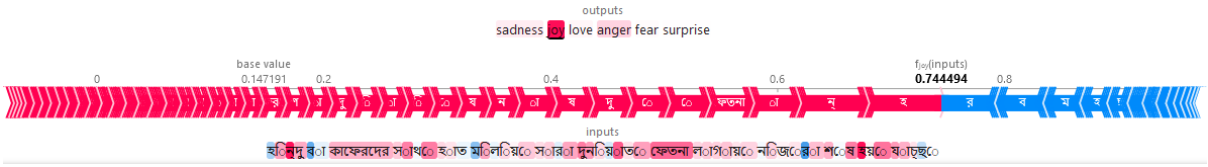

Figure 21: CyberBullDetector: Sentiment Analysis Explainer example '1' using SHAP on our test data

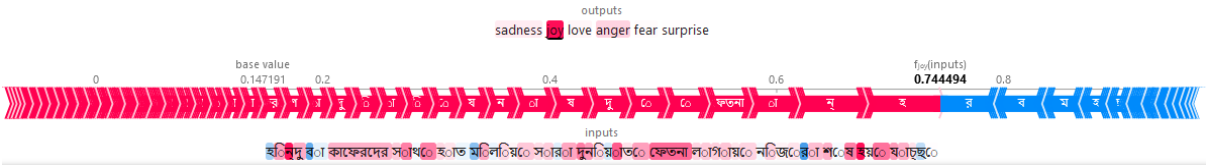

Figure 22: CyberBullDetector: Sentiment Analysis Explainer example '2' using SHAP on our test data

The way we tried to implement explainability in our thesis was not possible due to a variety of reasons. We faced a lot of problems while implementing it. The main issue was while tokenizing, the explainability was done letter by letter instead of word by word which didn't really make much sense. So using our model was not really a viable option here. So what we did was we used a different sentiment analysis model with our data text there to get the desired output. We did that to get a better understanding of how the explainability actually worked using SHAP. It is basically an example of sorts to show how explainbility could work on our models. So implementing the explainability on our models is something we hope to do in future.

Significance of Explainability in Cyberbullying

Accusing someone of cyberbullying has repercussions. General Data Protection Regulation (Council 2016) says, we need to able to explain how a machine decided to label something as cyberbullying.

SHAP (SHapley Additive exPlanations) : Model agnostic approach

For making our models explainable, we used the SHAP algorithm. It is a model agnostic approach to explain the output of any machine learning model.

It uses the Shapley Values from game theory to connect the local explanations with the

optimal credit allocation.

6 Conclusion and Future Scopes

6.0.1 Conclusion

From all the results that we have got we can infer that different models perform differently.Expectedly the multi lingual models (mBERT,mDistilBERT) perform better than their normal counterparts. Very much to our surprise XLnet performs relatively poorly. The ensembling technique does yield better results overall in the end. As far as explainability goes SHAP sometimes gives us a heatmap of letter by letter which makes less of a sense than word by word.

6.0.2 Future scopes

We intend to expand the dataset by introducing newer categories of classes in the Bangla cyberbullying-based dataset. We intend to conduct experiments with newer models that make use of cutting-edge NLP techniques. We also intend to expand the GenDisc dataset and use it for a different type of NLP task.

References

[1] R. M. Kowalski and G. W. Giumetti, "Bullying in the digital age," in *Cybercrime and its victims*, pp. 167–186, Routledge, 2017.

[2] V. Kumar and P. Nanda, "Social media in higher education: A framework for continuous engagement," *International Journal of Information and Communication Technology Education (IJICTE)*, vol. 15, no. 1, pp. 97–108, 2019.

[3] 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, 2018.

[4] R. Kshirsagar, T. Cukuvac, K. McKeown, and S. McGregor, "Predictive embeddings for hate speech detection on twitter," *arXiv preprint arXiv:1809.10644*, 2018.

[5] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *arXiv preprint arXiv:2106.00742*, 2021.

[6] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *European semantic web conference*, pp. 745–760, Springer, 2018.

[7] "Youths call for continued guidance to tackle online bullying amid increased internet use,"

[8] C. L. Nixon, "Current perspectives: the impact of cyberbullying on adolescent health," *Adolescent health, medicine and therapeutics*, vol. 5, p. 143, 2014.

[9] "https://www.findlaw.com/criminal/criminal-charges/cyber-bullying.html,"

[10] 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)*, pp. 1–6, IEEE, 2019.

[11] 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)*, pp. 53–60, IEEE, 2018.

[12] 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)*, pp. 1–10, IEEE, 2021.

[13] "What is cyberbullying by unicef,"

[14] "What falls under cyberbullying and the criteria for cyberbullying,"

[15] "What is cyber bullying?,"

[16] "psychological analysis of cyberbullying,"

[17] S. Paul and S. Saha, "Cyberbert: Bert for cyberbullying identification," *Multimedia Systems*, pp. 1–8, 2020.

[18] 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*, pp. 1900–1904, 2021.

[19] L. Bacco, A. Cimino, F. Dell'Orletta, and M. Merone, "Explainable sentiment analysis: A hierarchical transformer-based extractive summarization approach," *Electronics*, vol. 10, no. 18, p. 2195, 2021.

[20] I. Guellil, A. Adeel, F. Azouaou, F. Benali, A.-E. Hachani, K. Dashtipour, M. Gogate, C. Ieracitano, R. Kashani, and A. Hussain, "A semi-supervised approach for sentiment analysis of arab (ic+ izi) messages: Application to the algerian dialect," *SN Computer Science*, vol. 2, no. 2, pp. 1–18, 2021.

[21] S. Alsafari, S. Sadaoui, and M. Mouhoub, "Effect of word embedding models on hate and offensive speech detection," *arXiv preprint arXiv:2012.07534*, 2020.

[22] H. Faris, I. Aljarah, M. Habib, and P. A. Castillo, "Hate speech detection using word embedding and deep learning in the arabic language context.," in *ICPRAM*, pp. 453–460, 2020.

[23] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the indonesian language: A dataset and preliminary study," in *2017 International Confer-*

ence on Advanced Computer Science and Information Systems (ICACSIS), pp. 233–238, IEEE, 2017.

[24] N. I. Pratiwi, I. Budi, and I. Alfina, "Hate speech detection on indonesian instagram comments using fasttext approach," in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 447–450, IEEE, 2018.

[25] M. A. Fauzi, "Random forest approach fo sentiment analysis in indonesian," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 1, pp. 46–50, 2018.

[26] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of hindi-english code-mixed social media text for hate speech detection," in *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pp. 36–41, 2018.

[27] T. Santosh and K. Aravind, "Hate speech detection in hindi-english code-mixed social media text," in *Proceedings of the ACM India joint international conference on data science and management of data*, pp. 310–313, 2019.

[28] S. Tarwani, M. Jethanandani, and V. Kant, "Cyberbullying detection in hindi-english code-mixed language using sentiment classification," in *International conference on advances in computing and data sciences*, pp. 543–551, Springer, 2019.

[29] 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)*, pp. 385–388, IEEE, 2018.

[30] A. M. Ishmam and S. Sharmin, "Hateful speech detection in public facebook pages for the bengali language," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 555–560, IEEE, 2019.

[31] R. Kumar, B. Lahiri, and A. K. Ojha, "Aggressive and offensive language identification in hindi, bangla, and english: A comparative study," *SN Computer Science*, vol. 2, no. 1, pp. 1–20, 2021.

[32] N. Romim, M. Ahmed, H. Talukder, S. Islam, *et al.*, "Hate speech detection in the bengali language: A dataset and its baseline evaluation," in *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pp. 457–468, Springer, 2021.

[33] A. K. Das, A. Al Asif, A. Paul, and M. N. Hossain, "Bangla hate speech detection on social media using attention-based recurrent neural network," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 578–591, 2021.

[34] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification for low-resource languages," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1–13, 2021.

[35] S. Sazzed, "Abusive content detection in transliterated bengali-english social media corpus," in *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pp. 125–130, 2021.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[37] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[38] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.