Proceedings of the ResT-UP2 Proceedings @LREC2022, pages 8–15
Marseille, 20-25 June 2022

BanglaHateBERT: BERT for Abusive Language Detection in Bengali

Md Saroar Jahan[⋆], **Mainul Haque** [⋄], **Nabil Arhab** [⋆], **Mourad Oussalah**[⋆]
[⋆]University of Oulu, CMVS, BP 4500, 90014, Finland
[⋄] City University, Dhaka, Bangladesh
{Md.Jahan,Nabil.Arhab, Mourad.Oussalah}@oulu.fi, mainul37@gmail.com

Abstract

This paper introduces BanglaHateBERT, a retrained BERT model for abusive language detection in Bengali. The model was trained with a large-scale Bengali offensive, abusive, and hateful corpus that we have collected from different sources and made available to the public. Furthermore, we have collected and manually annotated 15K Bengali hate speech balanced dataset and made it publicly available for the research community. We used existing pre-trained BanglaBERT model and retrained it with 1.5 million offensive posts. We presented the results of a detailed comparison between generic pre-trained language model and retrained with the abuse-inclined version. In all datasets, BanglaHateBERT outperformed the corresponding available BERT model.

Keywords: Bangla Hate BERT, Bangla Hate Dataset

1. Introduction

Bengali (pronunciation: [baŋla]) is the 6^{th} most spoken language worldwide, spoken by almost 260 million people, offering resources for potential hate speech detection. The Bengali language is Bangladesh's national language and the second most-spoken language in India (Thompson, 2012). The development of the internet in society promoted the freedom of speech at an unprecedented level. This has led to a continuous rise of hate speech and offensive language on social media. For instance, online abuse towards females is continuously rising in Bangladesh (Sambasivan et al., 2019). In addition, the development of machine learning models to tackle hate speech in real-time is challenging for low resource languages like Bengali because of a lack of datasets and tools for Bengali text classification (Hussain et al., 2018). Only a few works have been reported on Bengali hate speech detection in social media. For instance, we found the claim of six Bengali hate speech datasets and research work. However, only two datasets are publicly available. Among by (Karim et al., 2020), which is annotated into five different classes and follows the native Bengali dialect. Nevertheless, this dataset does not contain any non-hate classes that might fall short during model training for hate and non-hate detection. Another dataset by (Awal et al., 2018) of 2665 sentences translated from an English hate speech dataset that lacks the dialect of native Bengali. Furthermore, some datasets were code-mixed and written in English (Banik and Rahman, 2019). Besides, none of the datasets are balanced in terms of their classes, and only a tiny percentage contained hate samples (Romim et al., 2021). Table 1 shows a comparison of state-of-the-art datasets on Bengali hate speech.

We can distinguish three categories of automatic abusive language detection using natural language processing (NLP) pipeline: i) feature-based linear classifiers (Waseem and Hovy, 2016), (Ribeiro et al., 2018), ii) neural network architectures (e.g., CNN or Bi-LSTM) (Kshirsagar et al., 2018), (Mishra et al., 2018), (Mitrović et al., 2019), and, finally, iii) fine-tuning pre-trained language models, e.g., BERT, RoBERTa, (Liu et al., 2019), (Swamy et al., 2019). Results vary both across datasets and architectures, where linear classifiers showed good training performance but lower accuracy scores compared to neural architecture or BERT-like models. On the other hand, systems using pre-trained language models have gained momentum in the field. Although a common problem with pre-trained models is that the training language combination makes them well-fitted for general-purpose language understanding tasks, but their limits are well-acknowledged when facing domain-specific language tasks. To address this limitation, there is a growing interest in developing domain-specific BERT-like pre-trained language models, such as AlBERTo (Polignano et al., 2019) or TweetEval (Barbieri et al., 2020) for Twitter dataset, BioBERT for biomedical domain in English (Lee et al., 2020), FinBERT for the financial domain in English (Yang et al., 2020), IndicBERT (BERT for major Indian language (Kakwani et al., 2020)), LEGAL-BERT for the legal domain in English (Chalkidis et al., 2020) and HateBERT (BERT for English Hate speech) (Caselli et al., 2020). Similarly, for Bengali text classification, BnglaBERT (Sarker, 2021) has been promoted and shown to outperform other BERT models (i.e., indicBERT, m-BERT). However, this model was trained with general Bengali text and does not contain much hate text, which falls short in hate speech classification tasks. To enrich this model, we introduce BanglaHateBERT, a pre-trained BERT model for abusive language phenomena in social media in Bengali. Besides, since abusive language phenomena covers a wide spectrum, e.g., microaggression, stereotyping, offense, abuse, hate speech, threats, and doxing (Jurgens et al., 2019), our BenglaHateBERT contributes to identifying a wide range of Bengali abusive text.

This aims to bridge the gap in availability of the Ben-

Table 1: A comparison of all state of the art datasets on Bengali hate speech

Paper	Total data	Number Of class	Language	Availability
Classification Benchmarks for Under-resourced Bengali Language based on Multichannel Convolutional-LSTM Network (Karim et al., 2020)	5,699	05	Native Bengali	Publicly available
Hateful speech detection in public Facebook pages for the Bengali language (Ishmam and Sharmin, 2019)	5,126	06	Native Bengali	Not available
Toxicity Detection on Bengali Social Media Comments using Supervised Models (Banik and Rahman, 2019)	10,219	05	Mixed Bengali English	No available
A Deep Learning Approach to Detect Abusive Bengali Text (Emon et al., 2019)	4,700	07	Native Bengali	Not available
Threat and Abusive Language Detection on Social Media in Bengali Language (Chakraborty and Seddiqui, 2019)	5,644	07	Native Bengali	Not available
Detecting Abusive Comments in Discussion Threads Using Naïve Bayes (Awal et al., 2018)	2,665	07	Translated English to Bengali	Publicly available
Hate Speech detection in the Bengali language: A dataset and its baseline evaluation (Romim et al., 2021)	30,000	02	Native Bengali	Not available

gali hate dataset and pre-trained BERT model for Bengali domain-specific abusive language detection. Overall, this paper claims threefold contributions as follows:

1. A new 1540k Bengali offensive corpus collected from Reddit-banned offensive comments is released.

2. A new 15k native Bengali offensive balanced corpus and manually labeled as offensive and non-offensive, collected from youtube, and Facebook users' comments, is made available.

3. We proposed a domain-specific pre-trained BERT model, referred BanglaHateBERT, for the purpose of Bengali offensive/hate speech detection.

Section 2 describes the dataset development process, including corpus statistics, hate categories identification, annotator and annotation guidelines, and disagreement handling. Section 3 illustrates the BanglaHateBERT construction, including a brief introduction of BERT. The results are provided in Section 4.2 and finally, Section 5 draws the main findings of this work.

2. Creation of Bengali Hate Dataset

We shall consider a new Bengali dataset for textual offensive speech annotated at the sentence level. To collect data, we used the beautiful-soup[1] python library to directly collect data and convert them into CSV file format. We collected data from Facebook and Youtube mainly from social media groups, celebrity pages, local Bengali news pages, political news posts, roasting videos, and funny content posts from 1 January 2021 to 05 April 2022. First, we collected 110k posts and then filtered 8.5k with Bengali profane word string matching to increase the chances of hitting offensive posts (examples of profane words shown in Table 3). At the same time, for the purpose of enforcing class balance, we also identified 8.5k posts from the original dataset that do not contain offensive/hate content. Finally, we manually label these total of 17k offensive and non-offensive posts, and after data preprocessing and manual scrutinizing, we kept 15k that held up to our standard by discarding noise comments or statements presenting only Bengali text. In other words, we mainly remove unidentified characters, symbols, numbers, mentioned tags, emojis, tab tokens, URLs, etc. We have not performed the removal of stop-words and stemming for preserving data quality. The statistics of the collected dataset are summarized in Table 2. Next, to identify hate-speech content from the collected dataset, we first highlight the categories of hate speech that are investigated in the subsequent analysis. This is detailed in the next subsection.

2.1. Hate categories identification

Hate speech often occurs with different linguistic connotations, even in subtle forms (Fortuna and Nunes, 2018). Due to the nature of its diversity, we identified eight hate speech targets, which we describe and

[1] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Table 2: Statistic of Dataset.

Statistics	Count
Number of Tokens	190,823
Vocabulary Size	26430
Number of Posts	15000
Average number of Tokens per post	12.7
Non-hate class	7500
Hate class	7500

provide examples from the corpus as follows:

Xenophobia: is a term that primarily represents the form of discrimination manifested through biased actions and hate against foreigners (DE OLIVEIRA, 2020). An example: 'রোহিঙ্গারা আসার পর ইয়াবা ব্যাবসা অনেক বেড়ে গেছে। ' - 'After the arrival of Rohingyas, there was increased Yaba drug business'.

Racism: racism or racial segregation consists of a tendency of racial domination (Wolfe, 1999). (Clair and Denis, 2015) pointed out that racism is a biological or a cultural dominance of one or more racial groups related to, e.g., skin color or physical look differences. For example, from the corpus: 'রোহিঙ্গারা সব হারামি, ওদের দেশে না রাখাই ভালো ' -Rohingyas are all bastards, it is better not to keep them in the country'.

Sexual: This includes expressions with a sexual meaning or intention. Examples from the corpus is: 'আর আমি, বাড়া দিলে তুমি স-যত্নে গ্রহণ করিতে মনজিলে মাকসুদের পানি শুকিয়ে প্লাস্টিকের অর্গানিক বাঁড়া চাইতে!'-'And me, give you my d**k, you would take care of it, dry the water in the floor and ask for organic d**k!'. However, innocent sexual talk and sex educational conversions are considered differently (e.g., 'হস্তমৈথুন ভাল বা খারাপ ?'- 'Masturbation good or bad?').

Religious fundamentalism/Religious Intolerance: This is consistently associated with high levels of intolerance and prejudices toward targeting specific religious groups (Altemeyer and Altemeyer, 1996). This is exemplified in the following post: 'হিন্দুরা শিশ্ন পূজা করে'-'Hindu people worship dick'.

Homophobia: This corresponds to negative attitudes and feelings toward homosexuality. This includes people who are identified or perceived as being lesbian, gay, and bisexual. An example of this case from our corpus is: 'সালা সমকামী, পিছন দিয়ে করে '-'He is gay, he get f**k in his back'.

Besides the above-mentioned categories, we have also considered hate toward a person, geopolitical or political organization. For example, 'বিএনপির ঘরে ধুকে একটা একটা করে মারবো'-'We will target and kill each BNP by entering their house', this is a severe threat towards a political party which does not fall into the above categories; however, it fulfills the definition of hate speech. In the next section, we describe the process of manual annotation, indicating how a given post lies within a specific category of hate speech.

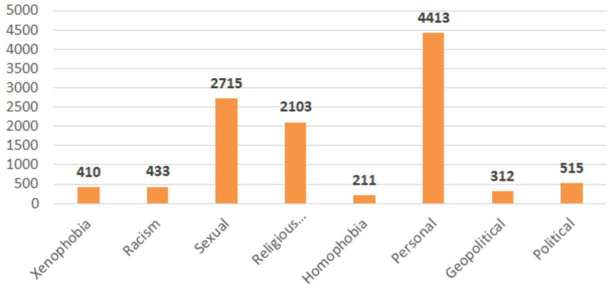

Figure 1: The number of hate samples in each category (same sample can exist in multiple categories).

2.2. Annotation Guidelines

The annotation involves identifying whether each sentence contains a hate speech or not by following the previously described hate categorization. In this study, all the annotators created and discussed the guidelines to ensure all participants had the same understanding of hate speech. A total of 27 independent native Bengali labelers have been employed separately to avoid bias. All annotators hold a minimum of a Bachelor's degree or are final-year Bachelor's students with a full ability to understand annotation guidelines. Besides, a research fellow has resolved disagreements between more than two annotators, who is a Ph.D. candidate in this field, and was called whenever a disagreement arises (total disagreement 339). If a sentence includes a hate, regardless of its hate category, it is given the label '1'; otherwise, it is assigned '0'. See examples shown in Table 4.

In our annotation, a sentence is considered *hate* if it satisfies the following criteria drawn from the hate definition by (Brown, 2017; Anis and Maret, 2017; Chetty and Alathur, 2018): *deliberate attack directed towards a specific group of people or organization employing sexual attack, curse, defamation, threat, gender, ethnicity, and identity.* Similar guidelines are also followed by Facebook and the youtube community for considering hate speech, which states *'Hate speech is a sentence that dehumanizes one or multiple persons or a community'.* Dehumanizing can be done by comparing the person or community to an insect, object, or criminal. It can also be done by targeting a person based on their race, gender, physical and mental disability [2]. A sentence might contain slang or inappropriate language. But unless that slang dehumanizes a person or community, we did not consider it to be associated with a hate speech [3]. Indeed, the presence or absence of offensive/abusive/profane words in a sentence cannot systematically be considered an acceptable proof to establish the existence of hate or not-hate. For example, Sentence 3 ('Gf why your two things are big') from Table-1 does not contain any offensive word, though, by definition, it is very offensive to someone. Another example in Sentence 4 ('some

[2] https://web.facebook.com/communitystandards/

[3] https://www.youtube.com/howyoutubeworks/policies/community-guidelines/

Table 3: Example of profane words.

Type	**Words**	**English Trsnlation**
Offensive	চুদি	F**k
Offensive	মাগী	Bi*ch
Offensive	সমকামী	Gay
Offensive	কাইলা	Black(skin color)
Offensive	খানকির পোলা	Bastard
Swear words	জাহান্নামে যাবি	Go to Hell
Swear words	মাইরা ফেলমু	kill you

people are just bastards, just ignore them'), includes the profane word 'Bastards'; however, it does not target any specific group; rather, it might have supported the victim, which makes it a non-hate sentence. Therefore, with regards to hate speech (HS), we decided to consider two characteristics for its identification:

1. There must be a target (i.e., an individual, race/group/community, or an organization), and
2. The action, or intention of the statement (Searle and Searle, 1969): this means that we must deal with a message that incites, spreads, promotes, or support violence or hatred towards the given target or a statement that aims at dehumanizing, delegitimizing, hurting or intimidating the target.

To understand the action or intention of the speaker, the use of profane words plays an important role. This is defined as socially improper use of language that includes offensive, cursing, swearing, or expletive wording. Table 3 highlights examples of frequent profane words extracted from the corpus.

Once labeled, 50% (7.5k) of the dataset was identified as hate, while the rest 50% (7.5k) were non-hate sentences. The final version of the dataset is saved in a CSV file that contains three columns (Posts: refer to collected sentences; Label: the judgment of the annotator in terms of hate or non-hate; Category: the type of hate speech). The details of the dataset collection are made available at this GitHub page[4].

2.3. Inter Annotator Agreement

We used Krippendorff's alpha (α) (Krippendorff, 1970) to measure the inter-annotator agreement because of the nature of our annotation setup. This robust statistical measure accounts for possible incomplete data and, therefore, does not require every annotator to annotate every sentence systematically.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

Here α is calculated by Equation (1), where (D_o) is the observed number of disagreements and (D_e) stands for the estimated likelihood of a disagreement occurring.

We used nominal metrics to calculate annotator agreement. The range of α is between 0 and 1, 1 α 0. When $\alpha = 1$, there is perfect agreement between annotators, and when α=0, the agreement is entirely due to a chance. Our annotation produced an agreement reliability score of 0.919 using nominal metric .

Disagreement Cases:
Our inter-annotator agreement score was satisfactory ($\alpha = 0.913$); however, some minor disagreements occurred. Below we summarize some problematic annotating examples that raise conflict among annotators.

1. 'দিঘী এখন সাগর হয় গেছে' - Dighi has now become the sea': Not sure whether the speaker used word 'sea' in a vulgar way in Bengali targeting a Bangladeshi actress 'Dighi'.
2. 'ব্যশ্যাস্যালয়ের মাটি ছাড়া দূর্গা মূর্তি গড়া অসম্পূর্ণ!' - 'It is incomplete to build a Durga idol without the soil of the brothel! ': Not sure whether the speaker intends to provide information or to devalue targeting the Hindu religion.
3. 'তাহসানের বউ এখন আরেকজনের বৌ' - 'Tahsan's wife is now someone else's wife': This post doesn't consist of any hate/swear words; however, mentioning someone's 'wife' might have the intention of defamation or insult or no intention at all. Therefore, it was complex to comprehend the intention of the speaker.
4. 'তুমি তো বরিশাইল্যা'-' You are Barisallya': The word 'Barishallya' is an ethnic slur typically used refers to a particular people of a region. Sometimes this is used as an insult and sometimes as a fun connotation.
5. 'ওরাল সেক্স কি করা যাবে'-'Can we do oral sex?': Despite the fact that this sample contains offensive terms, the speaker's goal may be harmless, and the question may be asked for educational purposes.

3. Creation of BanglaHateBERT

The Bidirectional Encoder Representations from Transformers (BERT) is a seminal transformer-based language model that involves an attention mechanism that enables contextual learning relations between words in a text sequence (Devlin et al., 2018). Two training strategies were used in our BERT model:

1. Masked-Language Modeling (MLM): where 15 % of the tokens in a sequence are masked, and then the model learns to predict those tokens.
2. Next sentence prediction (NSP): here, the model accepts two sentences as input and learns whether the second sentence is a successor of the first sentence in their original document context.

[4] `https://github.com/saroarjahan/BanglaHateBert`

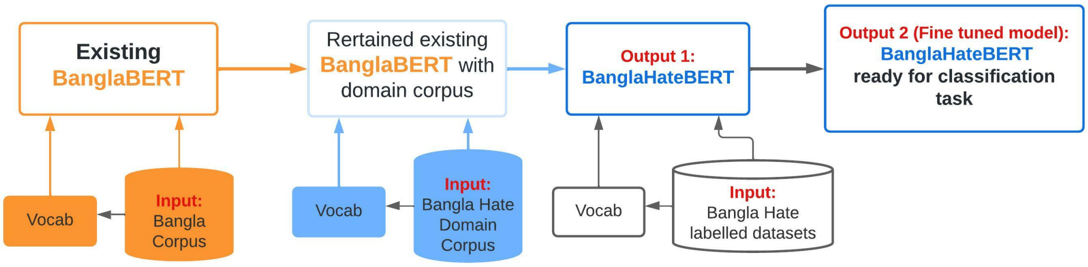

Figure 2: The architecture of BanglaHateBERT (Output1), can be used for further training with labeled corpus and ready for text classification tasks.

Sentence	Translation	Label	Categories
1. ওমি একটা পাগলা হালায় আছিলো।মাথায় গু আছিলো হালার	Omi is a crazzy, head full off shit	1	Personal
2. যেগুলো মানুষের নিত্যপ্রয়োজনীয় সেই গুলো দাম বাড়ছে। এই বাজেট কোনভাবে জনহিতকর বাজেট হতে পারে না	The prices of the necessities of life are rising. This budget can in no way be a public interest budget.	0	-
3. বান্ধবী তোমার ও দুটো এতো বড় কেনো	Gf why your two things is big	1	Personal, Sexual
4. কিছু মানুষ এমনিতেই হারামি এদের এত পাত্তা দিয়েন না	some people are just bastards just ignore them	0	-
5. তোদের মত নাস্তিকের বাচ্চার জন্য মুসলমানদের আজকের বদনাম	Today's notoriety of Muslims for the child of an atheist like you	1	Religious
6. ১৯৭১ এ ভারতের সাহায্য না করলে আজ পাকিস্তানের পা চাটতে	If India had not helped in 1971, Pakistan would have been licked today	1	Geoplitical
7. তুইতো একটা রেইনবো	Your are a rainbow (meaing gay)	1	Personal, Homophobia

Table 4: Annotation examples from original dataset with English translation. Label 1 refers to hate, and 0 refers to non-hate, categories column refers to type of hate speech.

The creation of BanglaHateBERT follows a two-step process is highlighted in Figure 3. First, we collected the large-scale Bangla offensive corpus, and then we retrained the existing BnaglaBERT with this Bengali offensive corpus.

Large-scale Bengali offensive corpus: Because of the lack of large-scale Bengali hate corpus for BERT training, we initially translated 16 offensive English hate datasets, with a total of 157k offensive sentences[5] to Bengali using Google API[6]. Furthermore, we have collected and translated (English to Bengali) 1478k Reddit-banned sentences that were considered offensive posts by the Reddit community. Finally, we have used these offensive sentences to retrain the BERT model.

Large scale Bengali pre-trained BERT model: To retrain the BERT model, we used an existing BanglaBERT model, a Bangla language model trained on 18.6 GB of Internet-crawled data from Wikipedia Bangla pages. In other words, the BanglaBERT model is trained on 1 million training steps over 3 billion tokens (24B characters) of Bengali text drawn from news, online discussion, and internet crawl (Sarker, 2021).

From the offensive corpus, we used 1,635,348 messages (a total of 40,309,341 tokens) to retrain the BanglaBERT base-uncased model by applying the Masked Language Model (MLM). We retrained for 15 epochs (almost 2 million steps) in batches of 64 samples, including up to 512 sentence tokens. We used Adam with a learning rate of 5e-5, which is an optimization solver for the Neural Network algorithm that is computationally efficient and requires little memory (Kingma and Ba, 2014). We trained using the huggingface code on one NvidiaRTX 3070 GPU. The result is a shifted BanglaBERT model to BanglaHateBERT.

4. Evaluation of BanglaHateBERT

To verify the validity and suitability of BanglaHateBERT, we compared it with other popular BERT models related to Bengali (i.e., BanglaBERT, multilingual-BERT, indicBERT). In addition, we also compared BERT performance with other deep-learning models CNN. We used one Bengali benchmarked dataset (Karim et al., 2020) for testing model performance. In contrast, we used our collected Bengali hate speech dataset as well.

[5] https://hatespeechdata.com/
[6] https://pypi.org/project/googletrans/

Table 5: Performance comparison of BanglaHateBERT vs. other models in terms of classifier Accuracy (%) and F1 scores (%) for Bengali hate speech detection using(Karim et al., 2020) dataset. Best scores are in bold.

Classifier	Accuracy	F1
CNN + fastText	92.1	91.3
BERT-multilingual	80	79.4
IndicBERT	89.4	88.1
BanglaBERT	92.4	92
BanglaHateBERT	**93.1**	**92.8**

4.1. Classifier Architecture

We performed a binary hate speech classification. For consistency, we used the same training, validation, and test samples for all models. We randomly shuffled and divided the entire collected dataset into three parts: training, validation, and testing set. For both datasets, we have used 70% for training, 10% for validation, and 20% for testing the model.

CNN-fastText Model Structure: We adopted (Kim, 2014) CNN architecture, where the input layer is represented by a concatenation of the words forming the post (up to 70 words), except that each word is now represented by its fastText embedding representation with a 300 embedding vector. Word embedding maps each token to a vector of real numbers aiming to quantify and categorize the semantic similarities between linguistic terms based on their distributional properties in a large corpus using machine learning or related dimensional reduction techniques. We used the pre-trained word embeddings; namely, Bengali fastText [7]. A convolution 1D operation with a kernel size of 3 was used with a max-over-time pooling operation over the feature map with a layer dense 50. Dropout on the penultimate layer with a constraint on l2-norm of the weight vector was used for regularization.

BERT Model Structure: We used Huggingface Transformers (Wolf et al., 2019) library for implementing the classifiers. We fine-tuned different transformer training data using 70% training data. The following models were tested: BanglaBERT, IndicBERT(covering 12 major Indian languages, multilingual-BERT (mBERT uncased), and BanglaHateBERT. Each model was fine-tuned for 6 epochs with a learning rate of 5e-6, maximum input sequence length of 128, and batch size 4. After each epoch, the model was evaluated on the test set. Fig. 3 illustrates our BERT architecture

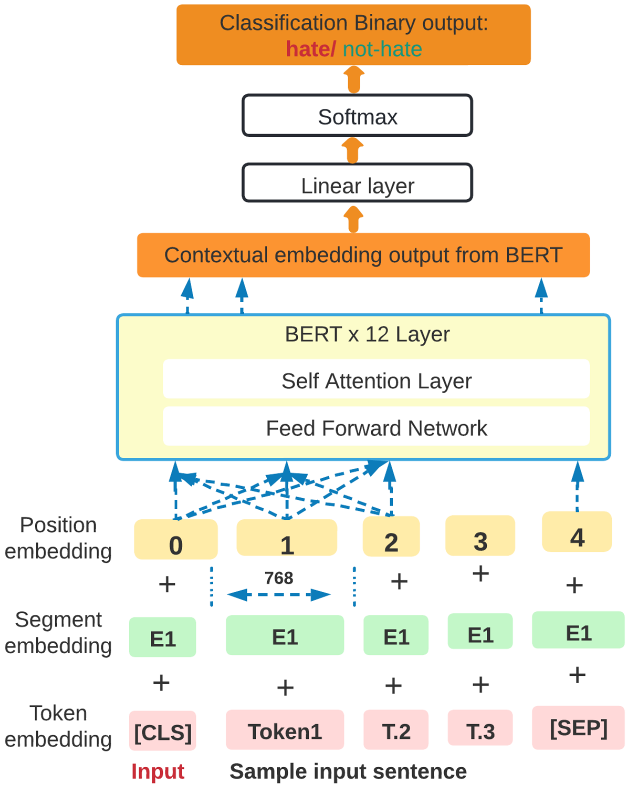

Figure 3: The general BERT architecture for text classification.

Table 6: Performance comparison of BanglaHateBERT Vs. other models in terms of hate speech classification Accuracy (%) and F1 scores (%) using our 15k balanced dataset. Best scores are in bold.

Classifier	Accuracy	F1
CNN + fastText	92.6	92.1
BERT-multilingual	82.1	81.3
IndicBERT	89.8	89.3
BanglaBERT	93.1	93
BanglaHateBERT	**94.3**	**94.1**

4.2. Results

The results of the binary classification of Bengali hate dataset by (Karim et al., 2020) and our collected dataset are summarized in Table 5, which shows classifier accuracy and F1 score for all four types of classifiers.

Among all five classifiers, BnaglaHateBERT outperformed all other models, indicating that the suggested BanglaHateBERT contextual model works better than the general one. These results have been observed for both balanced and unbalanced datasets, which followed an identical model performance rank: mBERT, IndicBERT, FastText, BanglaBERT, and BanglaHateBERT). For example, in both datasets, mBERT performed the lowest in terms of accuracy and F1 score compared to IndicBERT and BanglaBERT. This low

[7] https://fasttext.cc/docs/en/crawl-vectors.html (accessed 30.12.2021)

performance of mBERT can be explained by the fact that mBERThas was trained in over 102 languages. However, since it has only a small percentage of Bengali tokens, it falls short for domain-specific tasks. On the other hand, InbdicBERT performed overall better than mBERT, although it is also a multilingual BERT model. However, IndicBERT was trained over large-scale corpora covering 12 major Indian languages, containing a large portion of Bengali tokens (850 million). In both experiments, BnaglaBERT performed much better than mBERT and indicBERT since it has 3 billion Bengali tokens, which is much higher than mBERT and indicBERT. However, BanglaHateBERT performed even better than BanglaBERT since it has an additional 4 million tokens, which are primarily derived from the offensive corpus. The in-domain results confirm the validity of the re-training approach to generate better models for the detection of abusive language phenomena. On every dataset, BanglaHateBERT outperforms the corresponding general BERT model. These results can be further explained by observing the fastText model performance. For example, fastText did not perform better than BanglaBERT and BanglaHatebERT, which suggests that NLP contextual model is preferable compared to non-contextual word embeddings like fastText. However, interestingly it has outperformed indicBERT and mBERT as well, which indicates that the number of tokens highly influences model performance.

Strictly speaking, as far as we know, the (Karim et al., 2020) dataset has not been tested with BERT model previously. However, it has been tested with the deep learning model with word2vec embeddings and yielded 92.1% accuracy, which is 2% lower than our best performing model.

5. Conclusion

This paper introduced a new Bengali hate speech annotated dataset and BERT model for Bengali hate speech detection and experimented with mBERT, IndicBERT, BanglaBERT, and CNN models. To the best of our knowledge, this work is the first application of the BERT hate model trained with a domain-specific 1.5 million hate domain posts for Bengali. In addition, we published a balanced dataset (50% hate and 50% non-hate), which contains 15k posts collected from youtube and Facebook, which were then manually labeled and covered large categories of hate speech. In all cases, BanglaHateBERT has performed outstandingly in detecting hate speech compared to the mBERT, indicBERT, BanglaBERT, and CNN models, suggesting the effectiveness of domain-based contextual model performance over the non-domain-based contextual model. The developed BanglaHateBERT yields 94.3% accuracy and 94.1% F1 scores, which outperformed alternative models by a non-negligible margin.

6. Acknowledgements

This work is partially supported by EU YoungRes (ID. 823701) and Cost Action Nexus Linguarum, which are gratefully acknowledged.

7. Bibliographical References

Altemeyer, R. A. and Altemeyer, B. (1996). *The authoritarian specter*. Harvard University Press.

Anis, M. and Maret, U. (2017). Hatespeech in arabic language. In *International Conference on Media Studies*.

Awal, M. A., Rahman, M. S., and Rabbi, J. (2018). Detecting abusive comments in discussion threads using naïve bayes. In *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 163–167. IEEE.

Banik, N. and Rahman, M. H. H. (2019). Toxicity detection on bengali social media comments using supervised models. In *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, pages 1–5. IEEE.

Barbieri, F., Camacho-Collados, J., Neves, L., and Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Brown, A. (2017). What is hate speech? part 1: The myth of hate. *Law and Philosophy*, 36(4):419–468.

Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. (2020). Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Chakraborty, P. and Seddiqui, M. H. (2019). 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)*, pages 1–6. IEEE.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chetty, N. and Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118.

Clair, M. and Denis, J. S. (2015). Sociology of racism. international encyclopedia of the social and behavioral sciences 2nd.

DE OLIVEIRA, L. M. (2020). Imigrantes, xenofobia e racismo: uma análise de conflitos em escolas municipais de são paulo.

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

Emon, E. A., Rahman, S., Banarjee, J., Das, A. K., and Mittra, T. (2019). A deep learning approach to detect abusive bengali text. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5. IEEE.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Hussain, M. G., Al Mahmud, T., and Akthar, W. (2018). An approach to detect abusive bangla text. In *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, pages 1–5. IEEE.

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

Jurgens, D., Chandrasekharan, E., and Hemphill, L. (2019). A just and comprehensive strategy for using nlp to address online abuse. *arXiv preprint arXiv:1906.01738*.

Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (2020). IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Karim, M. R., Chakravarti, B. R., P. McCrae, J., and Cochez, M. (2020). Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In *7th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA,2020)*. IEEE.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

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

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Liu, P., Li, W., and Zou, L. (2019). Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *SemEval@ NAACL-HLT*, pages 87–91.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2018). Neural character-based composition models for abuse detection. *arXiv preprint arXiv:1809.00378*.

Mitrović, J., Birkeneder, B., and Granitzer, M. (2019). nlpup at semeval-2019 task 6: A deep neural language model for offensive language detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 722–726.

Polignano, M., Basile, P., De Gemmis, M., and Semeraro, G. (2019). Hate speech detection through alberto italian language understanding model. In *NL4AI@ AI* IA*, pages 1–13.

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., and Meira Jr, W. (2018). Characterizing and detecting hateful users on twitter. In *Twelfth international AAAI conference on web and social media*.

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

Sambasivan, N., Batool, A., Ahmed, N., Matthews, T., Thomas, K., Gaytán-Lugo, L. S., Nemer, D., Bursztein, E., Churchill, E., and Consolvo, S. (2019). " they don't leave us alone anywhere we go" gender and digital abuse in south asia. In *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Sarker, S. (2021). Banglabert: Bengali mask language model for bengali language understanding.

Searle, J. R. and Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Swamy, S. D., Jamatia, A., and Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 940–950.

Thompson, H.-R. (2012). *Bengali*, volume 18. John Benjamins Publishing.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wolfe, A. (1999). The bridge over the racial divide: Rising inequality and coalition politics. *The Journal of Blacks in Higher Education*, (26):127.

Yang, Y., Uy, M. C. S., and Huang, A. (2020). Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.