M-BAD: A Multilabel Dataset for Detecting Aggressive Texts and Their Targets

Omar Sharif[Ψ], **Eftekhar Hossain**[$] and **Mohammed Moshiul Hoque**[Ψ]
[Ψ]Department of Computer Science and Engineering
[$]Department of Electronics and Telecommunication Engineering
[$Ψ]Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{omar.sharif, eftekhar.hossain, moshiul_240}@cuet.ac.bd

Abstract

Recently, detection and categorization of undesired (e. g., aggressive, abusive, offensive, hate) content from online platforms has grabbed the attention of researchers because of its detrimental impact on society. Several attempts have been made to mitigate the usage and propagation of such content. However, most past studies were conducted primarily for English, where low-resource languages like Bengali remained out of the focus. Therefore, to facilitate research in this arena, this paper introduces a novel multilabel Bengali dataset (named ***M-BAD***) containing 15650 texts to detect aggressive texts and their targets. Each text of M-BAD went through rigorous two-level annotations. At the primary level, each text is labelled as either *aggressive* or *non-aggressive*. In the secondary level, the aggressive texts have been further annotated into five fine-grained target classes: *religion*, *politics*, *verbal*, *gender* and *race*. Baseline experiments are carried out with different machine learning (ML), deep learning (DL) and transformer models, where Bangla-BERT acquired the highest weighted f_1-score in both detection (0.92) and target identification (0.83) tasks. Error analysis of the models exhibits the difficulty to identify context-dependent aggression, and this work argues that further research is required to address these issues.

1 Introduction

Social media platforms have become a powerful tool to spontaneously connect people and share information with effortless access to the internet. These platforms provide users with a cloak of anonymity that allows them to speak their opinions publicly. Unfortunately, this power of anonymity is misused to disseminate aggressive, abusive, hatred and illegal content. In the recent past, these mediums have been used to incite religious, political and communal violence (Hartung et al., 2017). A significant portion of such incidents has been communicated through textual content (Kumar et al., 2020a; Feldman et al., 2021). Therefore, it has become crucial to develop automated systems to restrain the proliferation of such undesired or aggressive texts. This issue has been taken seriously in English, German, and other high-resource languages (Caselli et al., 2021; Aksenov et al., 2021). However, minimal research effort has been made in low-resource languages, including Bengali. Systems developed in English or other languages can not detect detrimental texts written in Bengali due to the significant variations in language constructs and morphological features. Nevertheless, people use their regional language to communicate over social media. Therefore, developing benchmark datasets and regional language tools is monumental to tackle the undesired text detection challenges. This work develops M-BAD containing 15650 texts using a two-level hierarchical annotation schema. In level-1, texts are categorized into binary classes: aggressive or non-aggressive. In level-2, 8289 aggressive texts are further annotated with multilabel targets. These labels are used to identify aggression's target into five fine-grained classes, such as *religion*, *gendered*, *race*, *verbal* and *politics* (detailed taxonomy discussed in Section 3). Proper annotation guidelines and the detailed statistics of the dataset is described to ensure M-BAD's quality. Several experiments are performed using ML, DL and transformer models to assess the task. The experiments demonstrate that (i) transformer models are more effective in detecting aggressive texts and their targets than ML/DL counterparts, (ii) covert propagation of aggression using ambiguous, context-dependent and sarcastic words is difficult to identify. The significant contributions of this work can be summarized as follows,

- Study two new problems from the perspective of low-resource language (i.e. Bengali), *(i)* detecting aggressive texts and *(ii)* identifying the multilabel targets of aggression.

- Release a new benchmark aggressive dataset labelled with the target of aggression and detailed annotation steps.
- Perform baseline experimentation on the developed dataset (M-BAD) to benchmark the two problems, providing the first insight into this challenging task.

Reproducibility: The resources to reproduce the results are available at https://github.com/omar-sharif03/M-BAD. The appendix contains details about data sources, annotators and a few samples of M-BAD.

2 Related Work

This section briefly describes the past studies related to aggression and other undesired content detection concerning non-Bengali and Bengali languages.

Non-Bengali aggressive text classification: Kumar et al. (2018a) compiled a dataset of 15000 aggression annotated comments in English and Hindi with three classes: *overtly aggressive, covertly aggressive, non-aggressive*. In their subsequent work (Kumar et al., 2020b), Bengali aggressive comments were added in the corpus. Early works with neural network techniques such as LSTM (Nikhil et al., 2018), CNN (Kumari and Singh, 2020), combination of shallow and deep network (Golem et al., 2018) achieved good accuracy. However, with the arrival of BERT based models, it acquired superior performance and outperformed all the models on these datasets (Risch and Krestel, 2020; Gordeev and Lykova, 2020; Sharif et al., 2021). Bhardwaj et al. (2020) developed a multilabel dataset in Hindi with five hostile classes: *fake, defamation, offensive, hate, non-hostile*. Their baseline system was implemented with m-BERT embedding and SVM. Leite et al. (2020) introduced a multilabel toxic language dataset. The dataset contains 21k tweets manually annotated into seven categories: *insult*, *LGBTQ+phobia*, *obscene*, *misogyny*, *racism*, *non-toxic* and *xenophobia*. They also performed baseline evaluation with the variation of BERT models. In a similar work, Moon et al. (2020) developed a corpus to detect toxic speech in Korean online news comments.

Bengali aggressive text classification: No significant research has been conducted yet to detect multilabel aggression in Bengali. The scarcity of benchmark corpora is the primary reason behind this. Few works have been conducted to develop datasets and models in other correlated domains such as hate, abuse, fake and offence. Karim et al. (2021) developed a hate speech dataset of 3000 samples with four categories: *political*, *personal*, *religious*, *geopolitical*. Emon et al. (2019) presented a dataset comprised of 4.7k abusive Bengali texts collected from online platforms. They proposed LSTM based classifier to categorize texts into seven classes. However, they did not investigate other DL models' performance, which might get similar accuracy with less computational cost. To detect the threat and abusive language, a dataset of 5.6k Bengali comments is created by Chakraborty and Seddiqui (2019). In recent work, Sharif and Hoque (2021a) introduced a benchmark Bengali aggressive text dataset. They employed a hierarchical annotation schema to divide the dataset into two coarse-grained (aggressive, non-aggressive) and four fine-grained (political, religious, verbal, gendered) aggression classes. In their later work (Sharif and Hoque, 2021b), they extended the dataset from 7.5k texts to 14k texts.

Differences with existing studies: As far as we are concerned, very few works have been accomplished to detect aggressive texts and identify the target of aggression (e.g. religion, gender, race). Existing works (Sharif and Hoque, 2021b; Zampieri et al., 2019; Kumar et al., 2018b) have framed it as a multi-class classification problem and ignored the overlapping phenomena of classes. However, a text can express aggression towards multiple targets simultaneously. Suppose a text has an aggressive write up against political women, expressing political and gendered aggressions. The proposed work addresses the issues that are previously overlooked and differs from the existing research in the following ways, (i) develop a novel Bengali aggressive text dataset annotated with the multiple targets of an aggressive text. As our knowledge goes, this is the first attempt to develop such a dataset in Bengali, (ii) illustrate a detailed annotation guideline which can be followed to develop resources for the similar domains in Bengali and other low-resource languages, (iii) perform experimentation with multilabel classes with various ML, DL and transformer-based models.

3 Dataset Development Taxonomy

This work presents a two-level hierarchical annotation schema to develop a novel multilabel ag-

gression dataset in Bengali (M-BAD). Level-1 has two coarse-grained categories: aggressive and non-aggressive. In contrast, level-2 has five fine-grained multilabel target classes (religion, politics, verbal, gender, race). This work differs from previous work done by Sharif and Hoque (2021b) in two ways; (i) overlapping phenomena between aggression targets are considered, (ii) a new target class (i.e., racial aggression) is added into the M-BAD. Figure 1 illustrates the taxonomic structure of M-BAD.

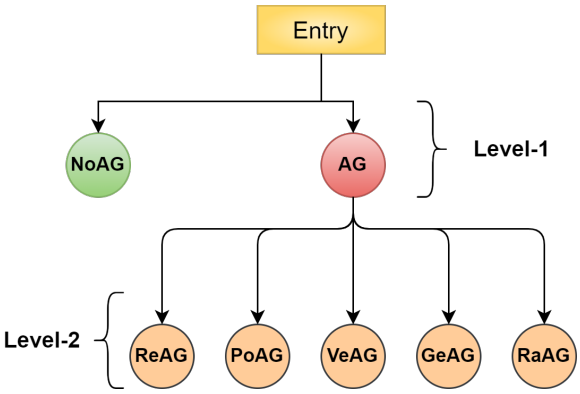


Figure 1: Taxonomic structure

Because of the subjective nature of the dataset, it is crucial to have a clear understanding of the categories. It helps develop a quality dataset by mitigating annotation biases and reducing ambiguities. After analyzing past studies (Sharif and Hoque, 2021b; Bhardwaj et al., 2020; Zampieri et al., 2019; Vidgen et al., 2021) on textual aggression and other related phenomena, we differentiate between the coarse-grained and fine-grained categories.

Coarse-grained Aggression Classes : The system initially identifies an input text as aggressive (AG) or non-aggressive (NoAG) classes.

- **(AG)**: excite, attack or seek harm to the individual, group or community based on a few criteria such as gender identity, political ideology, sexual orientation, religious belief, race, ethnicity and nationality.
- **(NoAG)**: do not contain any aggressive statements or express any evil intention to harm others.

Fine-grained Target Classes: An AG text is further classified into five fine-grained categories: religious aggression (ReAG), political aggression (PoAG), verbal aggression (VeAG), gendered aggression (GeAG) and racial aggression (RaAG). Each of the classes is defined in the following:

- **ReAG**: excite violence by attacking religion, religious organization or religious belief (Catholic, Hindu, Jew, or Islam, etc.) of a community
- **PoAG**: demean political ideology, provoke followers of political parties, or incite people in against law enforcement agencies and state.
- **VeAG**: seek to do evil or harm others, denounce the social status by using curse words, obscene words, outrageous and other threatening languages.
- **GeAG**: attack an individual or group by making aggressive reference to sexual orientation, sexuality, body parts, or other lewd contents.
- **RaAG**: insult or attack some and promote aggression based on race.

4 M-BAD: Multilabel Aggression Dataset

As far as we are concerned, no dataset is available to date for detecting or classifying multilabel aggressive texts and their targets in Bengali. However, the availability of a benchmark dataset is the prerequisite to developing any deep learning-based intelligent text classification system. This drawback motivates us to construct **M-BAD**: a novel multilabel Bengali aggressive text dataset. This work follows the guidelines and directions given by (Sharif and Hoque, 2021b; Vidgen and Derczynski, 2021) to ensure the quality of the dataset. This section briefly describes the data collection and annotation steps with detailed statistics of M-BAD.

4.1 Data Collection

We have manually accumulated **16000** aggressive and non-aggressive texts from different social platforms within the duration from 16 June to 27 December 2021. During this period, we only collected those texts that were posted, composed or shared after 1 January 2020. Potential texts were accumulated from YouTube channels and Facebook pages affiliated with political organizations, religion, newsgroups, artists, authors, celebrities, etc. Appendix A presents detailed statistics of the data collection sources.

Aggressive texts were cumulated from comments and posts that express aggression or excite violence. User profiles were also scanned who promoted, shared, or glorified aggression information to acquire additional texts. On the other hand, non-aggressive posts have been collected from news/comments/posts related to sports, education, entertainment, science and technology. Furthermore, while collecting aggressive texts, many data samples were found that did not express any aggression. Such texts were added to the corpus. We did not store any personal information (name, phone number, birth date, location) of the users during data accumulation. Each sample text is anonymized in the dataset. Thus, we do not know who has posted or created the collected texts. Finally, a few preprocessing filters are applied to remove inappropriate texts. 255 samples are discarded based on the following filtering criteria, (i) contains non-Bengali texts, (ii) has length fewer than three words, (iii) duplication. Remaining **15745** texts passed to the annotators for manual labelling.

4.2 Annotation Process

Section 3 describes the annotation schema and class definitions used to annotate the texts. Six annotators carried the annotation: four undergraduate and two graduate students. An expert verified the label in case of disagreement. Appendix B illustrates the detailed demographics of annotators. Annotators were split into three groups (two in each), and each group labelled a different subset of processed texts. To achieve quality annotations, we trained the annotators to define classes and associated examples. We tried to ensure that annotators understood what an aggressive text is and how to determine the target of aggression. Moreover, annotators are carefully guided in the weekly lab meetings.

Two annotators annotated each text, and the final label was assigned based on the agreement between the annotators. In case of disagreement, an expert resolve the issue through deliberations with the annotators. During the final label assignment, we found 95 texts that did not fall into any defined aggression categories and subsequently discarded them. Finally, we get M-BAD, an aggression dataset annotated with their targets containing **15650** texts. Appendix C shows few samples of M-BAD.

We measure the inter-annotator agreement using kappa score (Cohen, 1960) to check the validity of annotations. Table 1 presents the κ-*score* on both coarse-grained and fine-grained classes. The table shows that agreement is higher (0.77) in coarse-grained classes. The agreement is consistently 'moderate' (≈ 0.62) among the fine-grained classes but a bit lower in ReAG. Scores indicate difficulty in detecting targets of aggression by the annotators. Analysis reveals that sarcastic, implicit and ambiguous words made this difficult.

		κ-score	Average
Level-1	AG	0.85	0.77
	NoAG	0.69	
Level-2	ReAG	0.55	0.62
	PoAG	0.61	
	VeAG	0.62	
	GeAG	0.67	
	RaAG	0.65	

Table 1: Kappa (κ) score on each annotation level

4.3 Dataset Statistics

For training and evaluation purposes, the developed M-BAD is divided into the train (80%), test (10%), and validation (10%) split using a stratified strategy. The identical split ratio is used for both coarse-grained and multilabel fine-grained experiments. Table 2 presents the class-wise distribution of the texts for both Level-1 and Level-2. It is noticed that the distributions are slightly imbalanced with Level-2, which will be very challenging to handle in a multilabel setup.

Class	Train	Test	Valid	Total
ReAG	2391	327	305	3023
PoAG	2408	310	275	2993
VeAG	3939	498	472	4909
GeAG	1306	148	167	1621
RaAG	175	21	28	224
NoAG	5893	710	758	7361
AG	6642	840	807	8289

Table 2: Number of instances in train, test and validation sets for each category

	Class	#Words	#Unique words	Avg. #words/text
Level-1	AG	80553	17413	12.12
	NoAG	106573	24617	18.08
Level-2	ReAG	30748	9093	12.85
	PoAG	28410	8496	11.79
	VeAG	42342	11587	10.74
	GeAG	13817	4796	10.57
	RaAG	1711	1206	9.77

Table 3: Training set statistics in each level and class

To obtain in-depth insights, training set is further analyzed which is reported in Table 3. The statistics illustrated that in Level-1, NoAG class has the highest number of words (≈106k) and unique words (≈24k) compared to the AG class. Meanwhile, in Level-2, VeAG has the maximum number of words (≈42k) and unique words (≈11k) while RaAG class has the lowest (≈1.7k, ≈1.2k). However, the average number of words per text ranges from 10 to 12 among the aggression categories. Figure 2 shows the histogram of the texts length of each category. It is observed that ≈5000 texts of NoAG class have a length between ≈15-40. On the other hand, most of the length of the texts falls between 5-30 in VeAG class while ≈ 1000 texts of RaAG class has a length < 20. It is also noticed that only a small number of texts have length > 50.

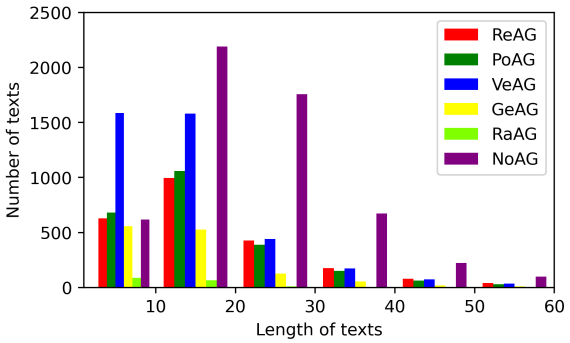


Figure 2: Histogram of the text length for each categories

	PoAG	VeAG	GeAG	RaAG
ReAG	0.38	0.47	0.36	0.18
PoAG		0.42	0.29	0.16
VeAG			0.50	0.25
GeAG				0.23
	NoAG			
AG	0.22			

Table 4: Jaccard similarity of 400 most frequent words between each pair of classes

We calculated the Jaccard similarity scores between the most 400 frequent words for quantitative analysis. Table 4 presents the similarity values among each pair of categories from Level-1 and Level-2. The VeAG-GeAG pair obtained the highest similarity score (0.50), while the PoAG-RaAG pair got the lowest score (0.16). It is observed that VeAG class has maximum similarity with almost all the classes except RaAG.

5 Methodology

Several computational models are investigated to develop the target aware aggression identification system. At first, the investigation is carried out for classifying the aggressive texts, and then we develop models for categorizing the target of the aggression (ReAG, PoAG, VeAG, GeAG, RaAG) considering the multilabel scenario. Machine learning and deep learning-based methods are employed to build the system. This section briefly discussed the techniques and methods used to develop the system.

5.1 ML-based methods

Two ML-based methods, Logistic Regression (LR) (Sharif and Hoque, 2019) and Naive Bayes with Support Vector Machine (NBSVM) (Wang and Manning, 2012) have been investigated for the classification task. Bag of words (BoW) features are used to train these models. The LR model is built with the 'lbfgs' optimizer and 'l2' regularization technique. Apart from this, the inverse regularization parameter C settled to 1.0. On the other hand, for NBSVM, the additive smoothing (α) and regularization parameters (C) are settled at 1.0 whereas the interpolation value is selected to $\beta = 0.25$.

5.2 DL-based Methods

Several popular DL methods are also investigated including BiGRU (Marpaung et al., 2021) and pretrained transformers (Vaswani et al., 2017) to identify the multi-label textual aggression.

BiGRU+FastText: The FastText (Joulin et al., 2016) embeddings are used as the input of the BiGRU model. Before that, a 1D spatial dropout technique is applied over the embedding features and then fed to a BiGRU layer with 80 hidden units. The last time step hidden output from the BiGRU is passed to a 1D global average pooling and a 1D global max-pooling layer. Subsequently, the two pooling layers outputs are concatenated and propagated to the classification layer.

Pretrained Transformers: In recent years, transformer (Vaswani et al., 2017) models trained on multilingual and monolingual settings achieved outstanding result in solving undesired text classification related tasks (Sharif and Hoque, 2021b; Hossain et al., 2021). As our task deals with a dataset of low-resource language, we employed three transformer-based models: (i) Multilingual

Bidirectional Encoder Representations for transformers (m-BERT) (Devlin et al., 2018) (ii) BERT for Bangla language (Bangla-BERT) (Bhattacharjee et al., 2021), and (iii) BERT for Indian languages (Indic-BERT) (Kakwani et al., 2020). The models have culled from the hugging face[1] transformers library and fine-tuned them with default arguments on the developed dataset.

Both ML and DL-based models are trained for two classification tasks: coarse-grained and multilabel fine-grained. To allow the reproducibility of the models and mitigate the training complexity, we use identical hyperparameters values for both classification tasks. We employed the Ktrain (Maiya, 2020) wrapper that provides easy training and implementation of the models. For multilabel classification, we enabled the Ktrain default multilabel settings. The BiGRU+FastText model is trained with a learning rate of $7e^{-3}$ while the transformer models with $8e^{-5}$. The models are trained using the triangular policy method (Smith, 2017) for 20 epochs with a batch size of 32. To save the best intermediate models, we utilized the early stopping criterion.

6 Experiments

The experiments were carried out in a google collaboratory platform with a GPU environment. The evaluation of the dataset is performed based on the weighted f_1-score. Due to the highly skewed distribution of the classes, we considered macro f_1-score (MF1) as our primary metric in multilabel evaluation. Besides, the individual class performance is measured through precision (P), recall (R), and f_1-score (F1) matrices.

6.1 Results

Table 5 presents the outcome of the different models on the test set concerning the coarse-grained classification. In terms of weighted f_1-score (WF1), both LR and NBSVM obtained an identical score of 0.91 while BiGRU + FastText and m-BERT model got a slightly low score (0.90). However, the Bangla-BERT model achieved the highest F1 across the two coarse-grained classes (AG/NoAG = 0.92) and thus outperformed all the models by achieving the highest WF1 score of 0.92.

Table 6 reports the evaluation results of the multilabel fine-grained classification. The outcome illustrates that the NBSVM obtained the lowest MF1 (0.61) and WF1 score (0.77). Both Indic-BERT and BiGRU+FastText models acquired identical WF1 of 0.79. Meanwhile, macro and weighted f_1-score is slightly (MF1 $\approx$4%, WF1 $\approx$ 1%) improved with the m-BERT model. However, the Bangla-BERT model exceeds all the models by achieving the highest MF1 (0.72) and WF1 (0.83). In terms of class-wise performance, Bangla-BERT obtained the highest f_1-score in four fine-grained aggression classes: ReAG (0.94), PoAG (0.92), VeAG (0.81), and GeAG (0.68). One interesting finding is that in RaAG class, some models (LR, NBSVM, Indic-BERT) did not identify a single instance correctly. Moreover, the models' performance degrades with the classes (GeAG, RaAG) having fewer training samples than other classes. Thus, a large dataset with balanced data distribution needs to be developed for classifying the problematic multilabel samples.

	AG			NoAG			
Method	P	R	F1	P	R	F1	WF1
LR	0.93	0.90	0.91	0.89	0.92	0.91	0.91
NBSVM	0.94	0.89	0.91	0.89	0.94	0.91	0.91
BG+FT	0.88	0.93	0.90	0.92	0.87	0.89	0.90
m-BERT	0.90	0.89	0.90	0.89	0.90	0.89	0.90
Indic-BERT	0.88	0.90	0.89	0.89	0.87	0.88	0.89
Bangla-BERT	0.93	0.91	0.92	0.91	0.93	0.92	**0.92**

Table 5: Performance of the Coarse-grained classification on the test set. Here, BG+FT represents BiGRU+FastText model

6.1.1 Error Analysis

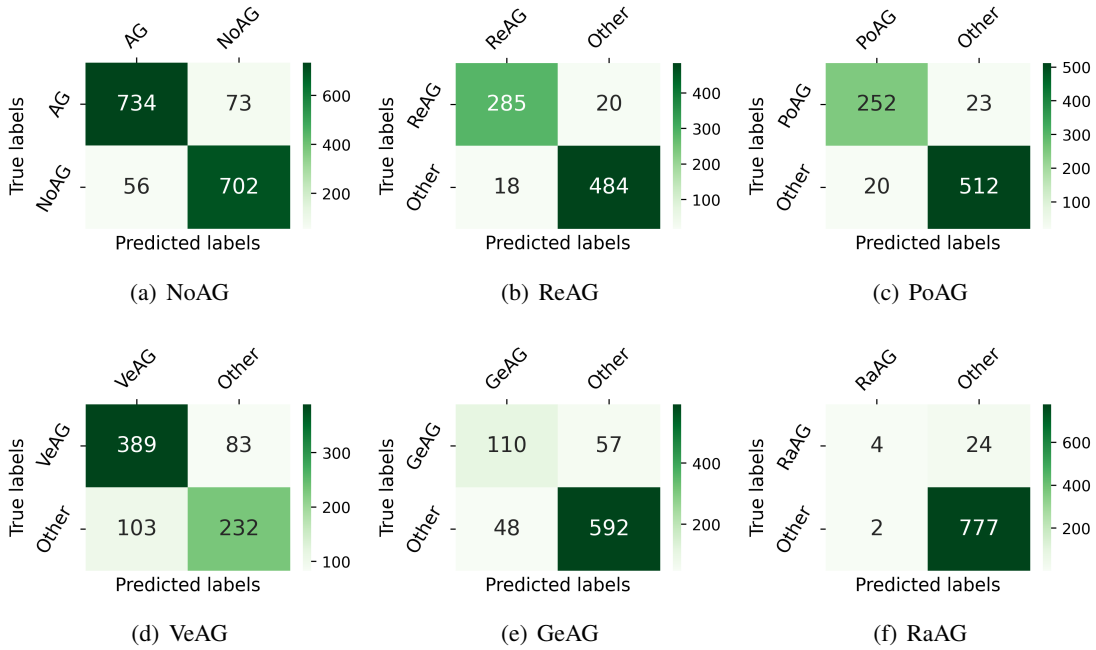
The results confirmed that Bangla-BERT is the best performing model in both coarse-grained and fine-grained classification tasks (Table 5, 6). We perform a thorough error analysis to know the model mistakes across different classes.

Quantitative analysis: Figure 3 shows the confusion matrices for the Bangla-BERT model. Figure 3 (a) depicts that with coarse-grained classification, the model incorrectly identified 73 (out of 807) and 56 (out 758) instances as NoAG and AG texts, respectively. The confusion matrices for fine-grained classes are shown in Figure 3 (b)-(f). It is noticed that in ReAG and PoAG classes model misclassified 20 (out of 305) and 23 instances (out of 275), respectively. The model yields the most incorrect predictions (24 out of 28) with RaAG class. The reason might be that the model did not get enough samples for learning and thus failed to discern the correct class in the testing phase. Meanwhile, in the case of VeAG, the model gets confused and mis-

[1] https://huggingface.co/

	ReAG			PoAG			VeAG			GeAG			Racism				
Method	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	MF1	WF1
LR	0.93	0.84	0.88	0.93	0.81	0.86	0.74	0.75	0.75	0.75	0.51	0.61	0.00	0.00	0.00	0.66	0.77
NBSVM	0.93	0.85	0.89	0.95	0.82	0.88	0.74	0.73	0.74	0.72	0.47	0.57	0.00	0.00	0.00	0.61	0.77
BG+FT	0.89	0.89	0.89	0.90	0.85	0.87	0.75	0.74	0.75	0.67	0.64	0.66	0.50	0.11	0.18	0.67	0.79
m-BERT	0.92	0.89	0.90	0.90	0.93	0.92	0.81	0.71	0.76	0.71	0.60	0.65	0.50	0.25	0.33	0.71	0.80
Indic-BERT	0.89	0.90	0.89	0.94	0.87	0.90	0.75	0.75	0.75	0.68	0.64	0.66	0.00	0.00	0.00	0.64	0.79
Bangla-BERT	0.94	0.93	0.94	0.93	0.92	0.92	0.79	0.82	0.81	0.70	0.66	0.68	0.67	0.14	0.24	**0.72**	**0.83**

Table 6: Fine-grained classification performance on the test set. Here, MF1 indicates the macro f_1-score



(a) NoAG (b) ReAG (c) PoAG

(d) VeAG (e) GeAG (f) RaAG

Figure 3: Confusion matrices of each category for Bangla-BERT model

classifies other classes instance (103 out of 232) as VeAG. The appearance of outrageous words in other fine-grained aggressive classes may be the reason for this confusion. Table 7 presents the false-negative rate (FNR) of the fine-grained categories. We noticed that the FNR is very high with GeAG (0.34) class while ReAG (0.065) and PoAG (0.08) classes FNR is deficient.

	False negative Rate
ReAG	20/305 (0.065)
PoAG	23/275 (0.08)
VeAG	83/472 (0.17)
GeAG	57/167 (0.34)
RaAG	4/28 (0.14)

Table 7: Error analysis for each fine-grained category

Qualitative Analysis: Figure 4 shows some correctly and misclassified sample texts from fine-grained classification tasks. The output predictions are obtained from the Bangla-BERT model. It is observed that the first two samples are correctly classified into different fine-grained aggression classes. However, in the third example, the model was only able to identify the text as **ReAG** and incorrectly predicted it as VeAG. Similarly, in the case of the last example model, it was not even able to classify it as RaAG. These examples illustrate the underlying difficulties of the multilabel classification problem. From the analysis, we found that the texts implicitly express aggression, which makes it arduous for the model to determine the multiple classes simultaneously. Moreover, some words have extensively appeared in the fine-grained classes. Perhaps, these words confuse the model to distinguish the classes and thus makes the task more difficult. Adding more training samples across all the classes might eradicate the problem to some extent.

7 Conclusion

This paper presented a multilabel aggression identification system for Bengali. To accomplish the purpose, this work introduced *M-BAD*, a multilabel

Text	Actual	Predicted
'ভারত আর বাংলাদেশের বিশ্ব বিদ্যালয়ে ধ**ণ কে পারফেক্ট করানোর কোর্স চালু হবে।' (Courses to make r**e perfect will be introduced in universities in India and Baladesh.)	PoAG, VeAG	PoAG, VeAG
'হিন্দুরা ভারতকে সাপোর্ট করে বাংলাদেশের চেয়ে বেশী। এই কু**র বাচ্চা হিন্দুদের দেশ হতে বের করে দেয়া দরকার' (Hindus support India more than Bangladesh. Get this Hindu bi**h out of the country)	ReAG, PoAG, VeAG	ReAG, PoAG, VeAG
'এই ধর্মান্ধ শালারা পরে সেক্স কন্ট্রোল করতে না পেরে ছাগল লাগাই' (These fanatical bastards can't control sex and then rape goats)	ReAG, RaAG	ReAG, VeAG
'পররাষ্ট্রমন্ত্রী এ সরিলে ভারতের হিন্দুর রক্ত আছে' (Foreign Minister has the blood of Hindus of India)	ReAG, PoAG, RaAG	ReAG, PoAG

Figure 4: Some correctly and incorrectly classified samples by the Bangla-BERT model

benchmark dataset consisting of 15650 texts. A two-level hierarchical annotation schema has been followed to develop the corpus. Among the levels, Level-1 is concerned with either aggressive or not aggressive, whereas Level-2 is concerned with the targets (religious, political, verbal, gender, racial) of the aggressive texts in a multilabel scenario. Several traditional and state of the art computational models have been investigated for benchmark evaluation. The results exhibit that the Bangla-BERT model obtained the highest weighted f_1-score of 0.83 for the multilabel classification. The error analysis revealed that it is challenging to identify the multiple targets of aggressive text as words are frequently overlapped across different classes. In future, we aim to mitigate this issue by exploring multitask learning and domain adaption approaches. Moreover, future work considers including more data samples with a significant period to minimize the bias towards a limited set of events.

Acknowledgements

This work supported by the ICT Innovation Fund, ICT Division, Ministry of Posts, Telecommunications and Information Technology, Bangladesh.

References

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. 2021. Fine-grained classification of political bias in German news: A data set and initial experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131, Online. Association for Computational Linguistics.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, Md Saiful Islam, M. Sohel Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier in multilingual models for low-resource language understanding. *CoRR*, abs/2101.00204.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Puja Chakraborty and Md. Hanif Seddiqui. 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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Estiak Ahmed Emon, Shihab Rahman, Joti Banarjee, Amit Kumar Das, and Tanni Mittra. 2019. A deep learning approach to detect abusive bengali text. In *2019 7th International Conference on Smart Computing Communications (ICSCC)*, pages 1–5.

Anna Feldman, Giovanni Da San Martino, Chris Leberknight, and Preslav Nakov, editors. 2021. *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. Association for Computational Linguistics, Online.

Viktor Golem, Mladen Karan, and Jan Šnajder. 2018. Combining shallow and deep learning for aggressive text detection. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 188–198, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Denis Gordeev and Olga Lykova. 2020. BERT of all trades, master of some. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 93–98, Marseille, France. European Language Resources Association (ELRA).

Matthias Hartung, Roman Klinger, Franziska Schmidtke, and Lars Vogel. 2017. Ranking right-wing extremist social media profiles by similarity to democratic and extremist groups. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–33, Copenhagen, Denmark. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2021. NLP-CUET@DravidianLangTech-EACL2021: Investigating visual and textual features to identify trolls from multimodal social media memes. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 300–306, Kyiv. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Bharathi Raja Chakravarthi, Md. Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language.

Ritesh Kumar, Atul Kr. Ojha, Bornini Lahiri, Marcos Zampieri, Shervin Malmasi, Vanessa Murdock, and Daniel Kadar, editors. 2020a. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA), Marseille, France.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020b. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018b. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kirti Kumari and Jyoti Prakash Singh. 2020. AI_ML_NIT_Patna @ TRAC - 2: Deep learning approach for multi-lingual aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 113–119, Marseille, France. European Language Resources Association (ELRA).

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*.

Angela Marpaung, Rita Rismala, and Hani Nurrahmi. 2021. Hate speech detection in indonesian twitter texts using bidirectional gated recurrent unit. In *2021 13th International Conference on Knowledge and Smart Technology (KST)*, pages 186–190. IEEE.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Nishant Nikhil, Ramit Pahwa, Mehul Kumar Nirala, and Rohan Khilnani. 2018. LSTMs with attention for aggression detection. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 52–57, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Julian Risch and Ralf Krestel. 2020. Bagging BERT models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 55–61, Marseille, France. European Language Resources Association (ELRA).

Omar Sharif and Mohammed Moshiul Hoque. 2019. Automatic detection of suspicious bangla text using logistic regression. In *International Conference on Intelligent Computing & Optimization*, pages 581–590. Springer.

Omar Sharif and Mohammed Moshiul Hoque. 2021a. Identification and classification of textual aggression in social media: Resource creation and evaluation. In *Combating Online Hostile Posts in Regional*

Languages during Emergency Situation, pages 1–12. Springer Nature Switzerland AG.

Omar Sharif and Mohammed Moshiul Hoque. 2021b. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. NLP-CUET@DravidianLangTech-EACL2021: Offensive language detection from multilingual code-mixed text using transformers. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 255–261, Kyiv. Association for Computational Linguistics.

Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE.

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar, Samuel R. Bowman, and Yoav Artzi. 2021. Crowdsourcing beyond annotation: Case studies in benchmark data collection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 1–6, Punta Cana, Dominican Republic & Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Sida I Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Appendix

A Data Sources

Data samples were collected from public post/comment threads of Facebook and YouTube. We did not store the profile information of any users. The data collection procedure is consistent with the copyright and terms of service of these organizations [2]. Potential texts were culled from more than 200 Bengali YouTube channels and Facebook pages. The popularity and activity status of a few data sources are presented in table A.1.

B Annotator Demographics

Past studies (Suhr et al., 2021; Zhou et al., 2021) on benchmark dataset creation have emphasized knowing about the demographic, geographic, research and other related information of the annotators. Since aggression is a very subjective phenomenon, annotators perspective and experience play a crucial role in developing the dataset. Six students and an expert were involved in our dataset construction process. Annotators demographic information, research experience, the field of research, and personal experience of viewing online aggression are summarized in table B.1.

Some key characteristics of the annotators' pool are, (i) native Bengali speakers, (ii) have prior experience of annotation, (iii) not an active member of any political parties, (iv) not hold extreme view against religion, (v) viewed online aggression. Before requiting, the annotators' necessary ethical approval was taken, and they are substantially paid according to university regulations.

C Data Samples

The authors would like to state that the examples referred to in the figure C.1 presented as they were accumulated from the source. Authors do not use these examples to hurt individuals or promote aggressive language usage. The goal of this work is to mitigate the propagation of such language.

[2] https://www.facebook.com/help/1020633957973118, https://www.youtube.com/static?template=terms

Page/channel name	Type	Affiliation	No. of followers/ subscribers	Reactions per post (in avg.)	Frequency of posting)
Bidyanondo	FP	Non political org.	5M	10k	10 post/day
Prothom Alo	FP/YC	Newsgroup	14M	4.5k	180 post/day
Rafiath Mithila	FP	Artist	3.8M	15k	4 post/week
Mizanur Azhari	YC	Religious speaker	1.9M	50k	1 post/month
Jamuna tv	FP/YC	Media	12.9M	3.7k	80 post/day
Awami League	FP/YC	Political org.	890k	4.6k	15 post/day
Abu Toha Adnan	FP	Religious speaker	2M	18k	10 post/week
Salman BrownFish	YC/FP	Musician	3M	15k	7 post/month
Arif Azad	FP	Author	742k	87k	8 post/month
Somoynews tv	FP/YC	Media	8.1M	2K	120 post/day
Basher kella	FP	Political	45k	400	15 post/day
Roar Bangla	FP/YC	Media	50K	300	3 post/day
Shakib Al Hasan	FP	Public figure	15.3M	50k	15 post/month

Table A.1: Activity and popularity statistics of a few sources from where data were gathered. FP indicates a Facebook page, and YC denotes a YouTube channel. Reactions are counted in terms of likes, comments and shares.

	AN-1	AN-2	AN-3	AN-4	AN-5	AN-6	Expert
Research-status	Undergrad	RA	Undergrad	Graduate	RA	Graduate	Professor
Research area	NLP	NLP	NLP	NLP	NLP	NLP	NLP, Social computig, HCI
Experience (years)	1	1	0.5	2.5	1.5	3	21
Prior annotation experience	yes	yes	no	yes	yes	yes	yes
Gender	Male	Male	Female	Female	Male	Male	Male
Age	22	23	22	25	23	26	47
Religion	Islam	Hindu	Hindu	Islam	Islam	Islam	Islam
Viewed online aggression	yes	yes	yes	yes	yes	yes	yes
Targeted by online aggression	yes	no	no	yes	no	yes	yes

Table B.1: Summary of annotators information.

Text	Level-1	Level-2
আরে ভাই কিট পতংগের সাইজো বড়ো আছে হিন্দুধর্মের তুলনায় (Hey brother, the size of few insects are bigger than Hinduism)	AG	ReAG
মোল্লাদের ঘরের মেয়েদের এরকম ধরে যতদিন ধ**ন না করা হবে, তত দিন এরা শায়েস্তা হবেনা। (As long as the girls of the mullah's house are not r**d like this, they will not be punished)	AG	ReAG, VeAG, GeAG
ভোটার বিহীন অবৈধ সরকার ছাত্রলীগ দিয়ে দেশটাকে ধ**নের স্বর্গরাজ্যে পরিনত করেছে। (The illegitimate government without voters has turned the country into a paradise of r**e with Chhatra League.)	AG	PoAG, VeAG
মেয়েদের এত পড়ালেখা করে আর কি লাভ হুদাই টাকা নষ্ট (What is the benefit of educating girls so much. It is just a waste of money)	AG	GeAG
মহিলা রাজনীতিবিদদের সংসদ থেকে বের করা দেয়া উচিত। সবগুলো ছাগল দেশের টাকা নষ্ট করতেছে (Women politicians should be expelled from Parliament. All the goats are wasting the country's money)	AG	GeAG, PoAG
চাকমাদেরকে দেশ থেকে বের করে দেয়া হক। (The Chakmas should be expelled from the country)	AG	RaAG
হাজারো সালাম জানাই শিক্ষকদের, যাদের অবদানে এগিয়ে যাচ্ছে বাংলাদেশ (Thousands of salutations to the teachers, who are helping Bangladesh to move forward)	NoAG	-
সাকিব আল হাসান, বাংলাদেশের জান বাংলাদেশের প্রান। এগিয়ে যাও (Shakib Al Hasan, the soul of Bangladesh. Go ahead)	NoAG	-

Figure C.1: Few samples of M-BAD