Identification and Classification of Textual Aggression in Social Media: Resource Creation and Evaluation

Omar Sharif and Mohammed Moshiul Hoque(✉)

Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Chittagong, Bangladesh
{omar.sharif,moshiul_240}@cuet.ac.bd

Abstract. Recently, social media has gained substantial attention as people can share opinions, expressions, emotions and carry out meaningful interactions through it spontaneously. Unfortunately, with this rapid advancement, social media misuse has also been proliferated, which leads to an increase in aggressive, offensive and abusive activities. Most of these unlawful activities performed through textual communication. Therefore, it is monumental to create intelligent systems that can identify and classify these texts. This paper presents an aggressive text classification system in Bengali. To serve our purpose a corpus (hereafter we called, 'ATxtC') is developed using hierarchical annotation schema that contains 7591 annotated texts (3888 for aggressive and 3703 for non-aggressive). Furthermore, the proposed system can classify aggressive Bengali text into religious, gendered, verbal and political aggression classes. Data annotation obtained a 0.74 kappa score in coarse-grained and 0.61 kappa score in fine-grained categories, which ensures the data's acceptable quality. Several classification algorithms such as LR, RF, SVM, CNN and BiLSTM are implemented on AtxtC. The experimental result shows that the combined CNN and BiLSTM model achieved the highest weighted f_1 score of 0.87 (identification task) and 0.80 (classification task).

Keywords: Natural language processing · Aggressive text classification · Bengali aggressive text corpus · Low resource languages · Deep learning

1 Introduction

With the phenomenal emergence of the internet, social media has become a powerful tool to spread and convey intentions, opinions, and feel to many people. However, it is very unpropitious that with this rise of social media, the incident of hate, abuse, cyberbullying and aggression has also increased significantly. Some people are misusing this power of social media to publicize aggressive and malicious contents, share fake news and spread illegal activities. Tech companies, academicians and policymakers are trying to develop NLP tools to identify

T. Chakraborty et al. (Eds.): CONSTRAINT 2021, CCIS 1402, pp. 9–20, 2021.
https://doi.org/10.1007/978-3-030-73696-5_2

these types of contents to mitigate unlawful activities. The aggressive/abusive text classification has much progressed for highly resource languages such as English [1,2], Arabic [3] etc. However, to the best of our knowledge, no significant resources have been developed to date for handling textual aggression in social media for low resource language like Bengali. Usually, people use their regional language to communicate over social media. For example, approximately 39 million people are using Facebook[1] through the Bengali language. Therefore, to improve the quality of conversation and reduce security threats over social media, we need to develop the necessary regional language tool. The key barriers to implement an aggressive text detection system in low resource language are the scarcity of benchmark corpora and related tools. Moreover, the overlapping characteristics of some correlated phenomena such as aggression, hate, abuse, profanity has made this task more complicated and challenging. Our goal is to compensate for this deficiency by developing an aggressive text classification framework for Bengali. The key contributions can be summarized as follows,

- Develop an Aggressive Text Corpus (ATxtC) which contains 3888 aggressive and 3703 non-aggressive Bengali texts. Hierarchical annotation schema uses to classify aggressive texts into religious, gendered, verbal and political aggression classes.
- Propose a benchmark system with experimental validation on ATxtC using machine learning and deep learning methods on each level of annotation.

2 Related Work

Detecting and classifying abusive contents (such as hate, aggression, and troll) has grabbed researchers' attention in recent years. Zampieri et al. [4] compiled a dataset of 14k offensive posts called 'OLID' to identify the type and target of an objectionable post. Their work proposed hierarchical annotation schema to detect abusive language. Ritesh et al. [5] developed aggression annotated corpus for Hindi-English code mixed data. They define various aggression dimension and corpus development process in detail. An ensemble approach was proposed by Arjun et al. [6] to identify the aggression in Hindi and English languages. XLMR and cross-lingual embeddings based model used by Ranasinghe et al. [7] on misogyny and aggression dataset [8]. For identifying and classifying abusive tweets, a corpus is created with three classes (offensive, hate and neither) [9]. This work used LR, DT, RF and SVM to classify tweets and concluded that it is challenging to identify covertly abusive texts. Although most of the works have carried out in English, a significant amount of related studies also focuses on Hindi, Greek, German and other languages too [10,11]. Due to the lack of benchmark corpora very few researches have been conducted in this area for Bengali. Ishmam et al. [12] develop a corpus of 5k Facebook post to categorize hateful Bengali language into six classes. Sharif et al. [13] proposed a system to detect suspicious Bengali texts. They trained their system on 7k suspicious and

[1] www.statista.com/statistics/top-15-countries-based-on-number-of-Facebook-users.

non-suspicious Bengali texts using ML techniques. A system trained with SVM on 5.5k Bengali documents to detect offence and threat in social media [14]. To our knowledge, this work is the first attempt to create a benchmark corpus to identify and classify aggressive Bengali texts.

3 Task Definition

Hierarchical annotation schema [4] is used to divide ATxtC into two levels: (A) identify whether a text is aggressive or not (B) classify an aggressive text into fine-grained classes namely religious aggression, gendered aggression, verbal aggression and political aggression.

3.1 Level A: Aggressive Text Identification

It is challenging to decide whether a text is aggressive or not because of its subjective nature. One person may contemplate a piece of text as aggressive while it seems normal to others. Moreover, overlapping characteristics of aggression with hate speech, cyber-bullying, abusive language, profanity have made this task more challenging. It is monumental to define the aggressive text to implement the aggressive text classification system successfully. After exploring the literature [5,6,15] and pursuing the properties of aggression we discriminate aggressive and non-aggressive text as following,

- **Aggressive texts (AG):** attack, incite or seek to harm an individual, group or community based on some criteria such as political ideology, religious belief, sexual orientation, gender, race and nationality.
- **Non aggressive texts (NoAG):** do not contain any statement of aggression or express hidden wish/intent to harm others.

3.2 Level B: Classification of Aggressive Text

As interpretation of aggression varies considerably across individuals, it is very important to have a fine line between aggression categories. To minimize the bias during annotation by analyzing existing research on aggression [2,5,16], toxicity [17], hate speech [18], abuse [19] and other related terminologies guided us to present definition of the following aggression classes:

- **Religious Aggression (ReAG):** incite violence by attacking religion (Islam, Hindu, Catholic, etc.), religious organizations, or religious belief of a person or a community.
- **Gendered Aggression (GeAG):** promote aggression or attack the victim based on gender, contain aggressive reference to one's sexual orientation, body parts, sexuality, or other lewd contents.
- **Verbal Aggression (VeAG):** damage social identity and status of the target by using nasty words, curse words and other obscene languages.

- **Political Aggression (PoAG):** provoke followers of political parties, condemn political ideology, or excite people in opposition to the state, law or enforcing agencies.

As far as our exploration, no research has been conducted to date that classifies aggressive Bengali texts into these classes.

4 Aggressive Text Corpus

No corpus of aggressive Bengali texts is available to best of our knowledge, which has above discussed fine-grained class instances. Therefore, we develop an annotated aggressive text corpus in Bengali. We discuss corpus development steps and provide a brief analysis of ATxtC in following subsections.

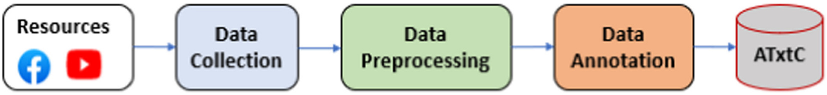


Fig. 1. ATxtC development steps

4.1 Corpora Development

To develop the corpus, we followed the directions given by Vidgen and Derczynski [20]. Figure 1 illustrates the ATxtC development steps, which has three major phases: data collection, data preprocessing and data annotation. After collecting raw data from different sources, we perform preprocessing to remove inconsistencies, and finally, human experts carry out annotation on these data.

Data Collection: We accumulated aggressive and non-aggressive texts manually from Facebook and YouTube as most of the Bengali social media users are active on these platforms. Most of the religious aggression data collected from comment threads of various Facebook pages and YouTube channels spread hatred and misinformation about religion. Most of the aggression's expressed in social media is against women which contain obscene and toxic comments. Texts related to gendered aggression is accumulated from several sources, including fashion pages, fitness videos, and news coverage on women/celebrities. Verbally aggressive texts include nasty words and obscene language. Political aggression texts procured from different pages. These pages stated about political parties and influential political figures and peoples' reaction to the government's different policies. Non-aggressive data culled from newspapers, Facebook and YouTube contents and these texts do not have any properties of aggression.

Data Preprocessing: To remove inconsistencies and reduce annotation efforts, we preprocessed the accumulated texts. All the flawed characters (!@#$%&) are dispelled. As concise texts do not contain any meaningful information, text having length fewer than two words discarded. Duplicate texts and texts written in languages other than Bengali are removed. After performing these steps processed texts passed to the human experts for manual annotation.

Data Annotation: As we noticed, annotation of aggression is entirely subjective, thus to reduce annotation bias, we choose annotators from the different racial, religious and residential background. A total of 5 annotators perform manual annotation. Some key characteristics of annotators are: a) age between 20–28 years, b) field of research NLP and experience varies from 10–30 months, c) all are native Bangla language speakers, d) active in social media and view aggression in these platforms. Prior to annotation, we provided examples of each category to the annotators and explained why a sample should be labelled as a specific class. Each of the instances was labelled by two annotators. In case of disagreement, we called an academician experienced in this domain to resolve the issue through discussion. During annotation, we observe that some of the texts have overlap among aggression dimensions. As these numbers are deficient, we do not include such instances in the current corpus for simplicity. We plan to address this issue in future when we get a large number of such cases. Some annotated samples of our ATxtC presented in Table 1.

Table 1. Some annotated instances in ATxtC. Here level A and level B indicates hierarchical annotation schema. English translation given for understanding

Text	**Level A**	**Level B**
“ মুসলিমদের মসজিদে নামায পড়া আর পাগলের রাস্তায় লাফালাফি করা এক জিনিস।" (Muslims praying in the mosque and jumping off an insane on the street both are same)	AG	ReAG
“বাংলাদেশর উন্নতির জন্য দরকার এই সরকারের পতন।"(The fall of this government is necessary for the betterment of Bangladesh)	AG	PoAG
“মেয়েরা হিংস্র জানোয়ার"(Girls are ferocious beast)	AG	GeAG
“মেসি বার্সেলোনার হয়ে রিয়াল মাদ্রিদের বিপক্ষে দুটি গোল করেছেন" (Messi has scored two goals for Barcelona against Real Madrid)	NoAG	-

4.2 Corpora Analysis

In order to check the quality and validity of the annotation, we measure the inter-annotator agreement. To examine the inter-rater agreement, we used Cohen's kappa [21] coefficient, which can be measured by Eq. 1.

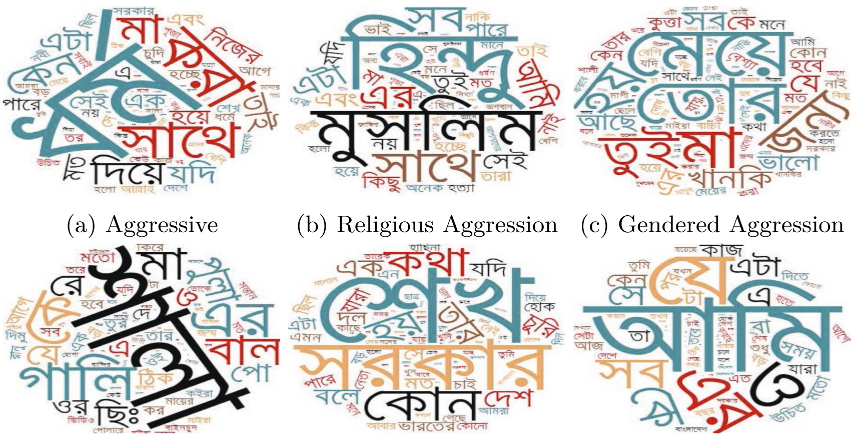
$$k = \frac{P(o) * P(e)}{1 - P(e)} \tag{1}$$

here P(o), P(e) are observed and the probability of chance agreement among annotators. The inter annotation agreement for coarse-grained classes is slightly lower than 74% while for fine-grained classes agreement is approximately 61%. The scores indicate that there exist substantial agreement between annotators.

A summary of the ATxtC exhibited in Table 2. Out of 7591 texts, 3888 texts are labelled as aggressive while remaining 3703 texts are non-aggressive. Aggressive texts further classified into fine-grained classes where religious, gendered, verbal and political aggression classes have 1538, 381, 1224 and 715 texts respectively. From this distribution, we can see that our corpus is highly imbalanced. This problem happened because of the scarcity of resources, and we could not cull a sufficient amount of data for some classes. We plan to tackle this issue by collecting more texts for rare classes. The average number of words in a non-aggressive text is higher than an aggressive text. Moreover, frequent words of various aggressive and non-aggressive categories depicted in word clouds are shown in Fig. 2(a) to Fig. 2(f). More highlighted words are most frequent than other words in a class.

Table 2. ATxtC statistics

	AG	ReAG	GeAG	VeAG	PoAG	NoAG
No. of texts	3888	1568	381	1224	715	3703
Total words	53850	27670	4200	11287	10693	75027
Unique words	12653	7553	1837	3794	3706	17501
Max. text length (words)	132	98	57	58	132	225
Avg. no. of words in texts	13.85	17.64	11.02	9.22	14.95	20.25



(a) Aggressive (b) Religious Aggression (c) Gendered Aggression

(d) Verbal Aggression (e) Political Aggression (f) Non Aggressive

Fig. 2. Word clouds representation of frequent words for each class in ATxtC

5 Methodology

In this section, we briefly describe the methods used to develop our models. Initially, features are extracted from texts with different feature extraction technique. We use logistic regression (LR), random forest (RF) and support vector machine (SVM) for preliminary model building. After that, we apply deep learning models, i.e. convolution neural network (CNN) and bidirectional long short term memory (BiLSTM) network to capture semantic features of the texts. Finally, we combine these deep models to check out its performance in ATxtC. Architecture and parameters of different methods illustrated in the following paragraphs.

Feature Extraction: Machine learning and deep learning methods could not possibly learn from the raw texts. So we have to extract features to train these models. For ML methods, we extract unigram, bigram and trigram features using tf-idf technique [22]. For DL, we use Word2Vec [23] embedding technique for feature extraction. ATxtC corpus is utilized to create the embedding matrix by using the embedding layer of Keras. Embedding maps textual data into a dense vector which holds the semantic meaning of words. The embedding dimension determines this dense vector's size, and we choose 200 as our optimal dimension. Similar features are used for both coarse-grained and fine-grained classifications.

Machine Learning Methods: ML methods popular for solving different text classification problem are used to build the baseline models. We use 'lbfgs' optimizer and 'l2' regularization technique with $\mathrm{c} = 0.9$ to implement LR. In RF, we use 100 estimators and take 'entropy' as a criterion. If there exist at least two samples in a decision branch, it is split. We construct SVM with 'linear' kernel and set value $\mathrm{c} = 0.5$ and $\gamma = 1$. In each case, parameters were chosen with the trial and error approach.

CNN: CNN has already been used to successfully identify aggression from texts [24]. CNN has a convolution layer which can adopt inherent features and syntactic information of the texts. We use one convolution layer with 64 filters and kernel size 3. We apply max pooling with poll size 3 to downsample the features. To add non-linearity 'relu' activation function used with CNN. Increasing the number of filters or layers harms the system results.

BiLSTM: LSTM is well-known for its ability to capture contextual information and long-term dependencies. We employed bidirectional LSTM to make use of the information from both past and future states. One layer of BiLSTM used with 64 cells and dropout rate of 0.2. Dropout technique applied to reduce overfitting. Dense layer takes the output from BiLSTM for prediction.

CNN+BiLSTM: Combined deep learning models have already been proven fruitful for aggressive text classification [25]. In this approach, we combine the CNN and BiLSTM networks sequentially. We make a slight modification in the parameter values of the network. Previously, we use 64 BiLSTM units with a dropout rate of 0.2. In the combined model, we use 32 LSTM units and reduce the dropout rate to 0.1.

To choose the optimal hyperparameters, we played with different parameter combination. Parameters values adjusted based on its effect on the validation set result. For coarse-grained and fine-grained classification, we use ‘binary_crossentropy’ and ‘categorical_crossentropy’ loss, respectively. Models are trained with the ’adam’ optimizer up to 30 epochs. In each batch, there are 64 instances and learning rate set to 0.001. Keras callbacks used to save the best intermediate model. We employ the same architecture for both coarse-grained and fine-grained classification with marginal modification on the parameter values. Finally, the trained model evaluated on the unknown test set instances.

6 Experiments and Result Analysis

In this work, our goal is to identify whether a text is aggressive or not and classify potential aggressive texts into fine-grained classes, namely ReAG, GeAG, VeAG and PoAG. We used weighted f_1 score to determine the models’ superiority and present models precision, recall, and f_1 score for each class. We employ three machine learning techniques (LR, RF, SVM), two deep learning techniques (CNN, BiLSTM) and one combined (CNN+BiLSTM) model to serve our purpose. We conduct experiments on open-source google colaboratory platform. Keras==2.4.0 framework used with tensorflow==2.3.0 in the backend to create DL models. We use scikit-learn==0.22.2 to implement ML models and pandas==1.1.4 to process and prepare data.

Before developing the models, ATxtC partitioned into three mutually exclusive sets: train, validation and test. Train data used to build the models while we tweak model parameters based on the validation set results. Finally, models are evaluated on the blind test set. To eliminate any bias, we perform random shuffling before data partitioning. A detail statistics of the dataset presented in the Table 3.

Table 3. Class-wise distribution of train, validation and test set in ATxtC. Level A indicates coarse-grained classes (aggressive and non-aggressive). Level B indicates fine-grained classes (religious, gendered, verbal and political aggression).

	Level A		Level B			
	AG	NoAG	ReAG	GeAG	VeAG	PoAG
Train	2721	2601	1078	266	858	501
Validation	386	364	165	38	125	79
Test	781	738	325	77	241	135
Total	3888	3703	1568	381	1224	715

Models performance of coarse-grained classification reported in Table 4. Here we aim to identify whether a text is aggressive (AG) or non-aggressive (NoAG). All the models get mentionable accuracy on this task. All three DL models achieve a weighted f_1 score of 0.87. Among the ML models, LR gets maximum, and RF achieves a minimum f_1 score of 0.86 and 0.81, respectively. LR, CNN and BiLSTM get highest f_1 score of 0.87 on AG class. However, in NoAG class, combined model along with CNN achieve maximum 0.88 f_1 score.

Table 4. Evaluation results for coarse-grained identification aggressive texts. Here P, R, F1 denotes precision, recall and f_1 score respectively and (C+B) indicates combined CNN & BiLSTM model.

	Measures	LR	RF	SVM	CNN	BiLSTM	C+B
AG	P	0.84	0.77	0.79	0.86	0.85	0.91
	R	0.90	0.86	0.93	0.88	0.90	0.84
	F1	**0.87**	0.81	0.85	**0.87**	**0.87**	0.86
NoAG	P	0.89	0.85	0.92	0.88	0.90	0.86
	R	0.83	0.75	0.77	0.87	0.85	0.90
	F1	0.86	0.80	0.84	**0.88**	0.87	**0.88**
Weighted	P	0.87	0.81	0.86	0.87	0.87	0.87
	R	0.86	0.81	0.84	0.87	0.87	0.87
	F1	0.86	0.81	0.84	**0.87**	**0.87**	**0.87**

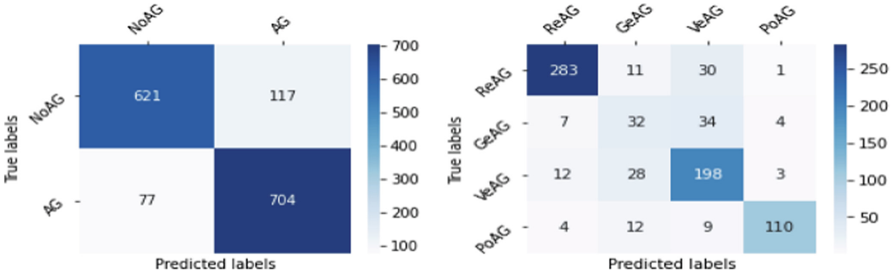
Table 5. Evaluation results for fine-grained classification of aggressive texts. Here P, R, F1 denotes precision, recall and f_1 score respectively and (C+B) indicates combined CNN & BiLSTM model.

	Measures	LR	RF	SVM	CNN	BiLSTM	C+B
ReAG	P	0.74	0.79	0.73	0.91	0.90	0.92
	R	0.95	0.82	0.95	0.90	0.85	0.87
	F1	0.83	0.80	0.83	**0.90**	0.87	**0.90**
GeAG	P	0.91	0.70	0.90	0.33	0.28	0.39
	R	0.13	0.25	0.12	0.06	0.29	0.42
	F1	0.23	0.37	0.21	0.11	0.28	**0.40**
VeAG	P	0.80	0.67	0.78	0.71	0.78	0.73
	R	0.90	0.87	0.89	0.93	0.81	0.82
	F1	**0.84**	0.76	0.83	0.80	0.79	0.77
PoAG	P	0.95	0.83	0.96	0.86	0.79	0.93
	R	0.65	0.64	0.62	0.81	0.84	0.81
	F1	0.77	0.72	0.75	0.84	0.82	**0.87**
Weighted	P	0.82	0.75	0.81	0.78	0.78	0.81
	R	0.79	0.74	0.78	0.81	0.78	0.81
	F1	0.76	0.73	0.75	0.78	0.78	**0.80**

Table 5 exhibits model results on fine-grained evaluation. Models classify aggressive texts into four pre-defined aggression classes. These classes are religious aggression (ReAG), gendered aggression (GeAG), verbal aggression (VeAG) and political aggression (PoAG). We can see that DL models perform better compare to ML models. The combined method outdoes all others by achieving a maximum of 0.80 weighted f_1 score. In ReAG, GeAG and PoAG classes combined model get highest f_1 score of 0.90, 0.40 and 0.87 respectively. LR get highest 0.84 f_1 score for VeAG class. The performance of all models for GeAG class is lower compare to other classes. The fewer number of training examples in this class might be the reason behind this unusual performance. Among the models, RF performs poorly in all classes, and the combined model achieve the superior result in most of the classes.

6.1 Error Analysis

Combination of CNN and BiLSTM is our best performing model for both identification and classification task. In this section, we discuss the detail error analysis of this model. To analyze the errors, we perform a quantitative analysis from the confusion matrix. Figure 3 shows the confusion matrix of the combined model in the test set.



(a) Coarse-grained classification

(b) Fine-grained classification

Fig. 3. Confusion matrix for CNN+BiLSTM model

Figure 3(a) shows that the false positive rate is higher than the false-negative rate and classifier wrongly classifies 117 non-aggressive texts as aggressive. This occurs because some aggressive words may present in the texts in a sarcastic way which does not mean any aggression or harm. In 77 cases model fail to identify aggressive texts because some texts might hold covert aggression which is very difficult to locate. Figure 3(b) observes that texts from ReAG class commonly confused with VeAG class and PoAG class texts confuse the GeAG class. It is to be noted that among 77 VeAG texts model inappropriately classified 34 texts as GeAG, which is higher than the number of correct predictions.

This misclassification happened because most of the GeAG texts contain a large amount of vulgar and nasty words. Increasing the number of training examples for this class might help the model to generalize better. Few misclassification examples on the test set listed in Table 6.

Table 6. Few misclassified examples by CNN+BiLSTM model. A and P denotes actual and predicted class respectively.

Text	**A**	**P**
“ কুত্তার মতো পিটবো বেয়াদব মেয়েটাকে" (Beat the rude girl like a dog)	GeAG	VeAG
“জিকির না গান না উচ্চাঙ্গসংগীত বুঝতে পারলামনা, হয়তো বা শয়তান বুঝবে" (I don’t understand whether it is Zikir or classical music, maybe the Shaitan will understand)	ReAG	VeAG
“সৃজিতকে দেখলেই মনে হয় খারাপ মানুষ" (Seeing Srijit, it seems that he is a bad person)	VeAG	GeAG

7 Conclusion

This paper describes the development process of a benchmark aggressive text corpus using hierarchical annotation schema. This corpus manually annotated 7591 texts with four fine-grained classes (religious, gendered, verbal and political aggression). As the baseline, several supervised machine learning (LR, RF, SVM) and deep learning (CNN, BiLSTM, CNN+BiLSTM) models are investigated. The proposed system evaluated into two tasks: aggressive text identification and classifying aggressive texts to fine-grained classes. In both cases, the combined model (CNN+BiLSTM) outperforms others by achieving a maximum of 0.87 and 0.80 weighted f_1 score. Attention mechanism with BERT, ELMo, and other word embedding techniques may be applied to observe their classification performance effects. As deep learning algorithms do very well, it will be interesting to see how they perform when we pursue ensemble techniques. Finally, adding more diverse data in the corpus will undoubtedly help the models to generalize better.

Acknowledgement. This work supported by ICT Division.

References

1. Prabhakaran, V., Waseem, Z., Akiwowo, S., Vidgen, B.: Online abuse and human rights: WOAH satellite session at RightsCon 2020. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 1–6 (2020)
2. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Evaluating aggression identification in social media. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pp. 1–5 (2020)
3. Mubarak, H., Rashed, A., Darwish, K., Samih, Y., Abdelali, A.: Arabic offensive language on Twitter: analysis and experiments. arXiv preprint arXiv:2004.02192 (2020)
4. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666 (2019)
5. Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T.: Aggression-annotated corpus of Hindi-English code-mixed data. arXiv preprint arXiv:1803.09402 (2018)
6. Roy, A., Kapil, P., Basak, K., Ekbal, A.: An ensemble approach for aggression identification in English and Hindi text. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 66–73 (2018)

7. Ranasinghe, T., Zampieri, M.: Multilingual offensive language identification with cross-lingual embeddings. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5838–5844 (2020)
8. Bhattacharya, S., et al.: Developing a multilingual annotated corpus of misogyny and aggression. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pp. 158–168 (2020)
9. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv:1703.04009 (2017)
10. Bhardwaj, M., Akhtar, M.S., Ekbal, A., Das, A., Chakraborty, T.: Hostility detection dataset in Hindi. arXiv preprint arXiv:2011.03588 (2020)
11. Pitenis, Z., Zampieri, M., Ranasinghe, T.: Offensive language identification in Greek. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 5113–5119 (2020)
12. Ishmam, A.M., Sharmin, S.: 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)
13. Sharif, O., Hoque, M.M., Kayes, A., Nowrozy, R., Sarker, I.H.: Detecting suspicious texts using machine learning techniques. Appl. Sci. **10**(18), 6527 (2020)
14. Chakraborty, P., Seddiqui, M.H.: 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)
15. Baron, R.A., Richardson, D.R.: Human Aggression, 2nd edn. Plenum Press, New York (1994)
16. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Benchmarking aggression identification in social media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 1–11 (2018)
17. van Aken, B., Risch, J., Krestel, R., Löser, A.: Challenges for toxic comment classification: an in-depth error analysis. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 33–42 (2018)
18. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (CSUR) **51**(4), 1–30 (2018)
19. Ibrohim, M.O., Budi, I.: Multi-label hate speech and abusive language detection in Indonesian Twitter. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 46–57 (2019)
20. Vidgen, B., Derczynski, L.: Directions in abusive language training data: garbage in, garbage out. arXiv preprint arXiv:2004.01670 (2020)
21. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)
22. Tokunaga, T., Makoto, I.: Text categorization based on weighted inverse document frequency. In: Special Interest Groups and Information Process Society of Japan (SIG-IPSJ). Citeseer (1994)
23. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
24. Kumari, K., Singh, J.P.: 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, pp. 113–119 (2020)
25. Aroyehun, S.T., Gelbukh, A.: Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 90–97 (2018)