Tackling Cyber-Aggression: Identification and Fine-Grained Categorization of Aggressive Texts on Social Media using Weighted Ensemble of Transformers

Omar Sharif[a], Mohammed Moshiul Hoque[a,*]

[a] *Department of CSE, Chittagong University of Engineering and Technology, Chittagong, 4349, Bangladesh*

Abstract

The pervasiveness of aggressive content in social media has become a serious concern for government organizations and tech companies because of its pernicious societal effects. In recent years, social media has been repeatedly used as a tool to incite communal aggression, spread distorted propaganda, damage social harmony and demean the identity of individuals or a community in the public spaces. Therefore, restraining the proliferation of aggressive content and detecting them has become an urgent duty. Studies of the identification of aggressive content have mostly been done for English and other resource-high languages. Automatic systems developed for those languages can not accurately identify detrimental contents written in regional languages like Bengali. To compensate this insufficiency, this work presents a novel Bengali aggressive text dataset (called 'BAD') with two-level annotation. In level-A, 14158 texts are labeled as either aggressive or non-aggressive. While in level-B, 6807 aggressive texts are categorized into religious, political, verbal and gendered aggression classes each having 2217, 2085, 2043 and 462 texts respectively. This paper proposes a weighted ensemble technique including m-BERT, distil-BERT, Bangla-BERT and XLM-R as the base classifiers to identify and classify the aggressive texts in Bengali. The proposed model can readdress the softmax probabilities of the participating classifiers depending on their primary outcomes. This weighting technique has enabled the model to outdoes the simple average ensemble and all other machine learning (ML), deep learning (DL) baselines. It has acquired the highest weighted f_1-score of 93.43% in the identification task and 93.11% in the categorization task.

Keywords: Natural language processing, Aggressive text classification, Low resource language, Bengali aggressive text corpus, Deep learning, Transformers, Ensemble

1. Introduction

The phenomenal proliferation of social media platforms (i.e. Facebook, Twitter, YouTube) has dramatically transformed people's communication mode. These platforms have become the potential medium to express people's opinions on various topics such as politics, religion, finance, sports and other societal events. Information shared on social media platforms has the power to reach millions within a short period. This rapid growth of information has not only resulted in a positive exchange of information, but also allow a group of malign users to disseminate aggressive, offensive, hatred, and other illegal contents. Past surveys reported that social media platforms had been utilized to publicize aggression, incite political and religious violence that jeopardize communal harmony and social stability [1]. The viciousness of aggressive and offensive texts is strong enough to trigger massive violence, create mental health problems or even instigate suicide [2, 3]. Therefore, it is monumental to develop resources and methods to flag such contents for reducing unlawful activities and keep the information ecosystem clean from polluted contents. Statistics show that social media platforms such as Facebook and YouTube have

*Corresponding author.
Email addresses: omar.sharif@cuet.ac.bd (Omar Sharif), moshiul_240@cuet.ac.bd (Mohammed Moshiul Hoque)

more the 4.8 billion[1] users who generate millions of posts/comments every day. It is impractical to moderate and monitor this massive volume of contents manually. These phenomena pose the necessity to develop automated systems which can identify and classify online offence analyzing this bulk of information. Over the last few years, several studies have been carried out to develop an automatic and semi-automatic system to tackle the spread of undesired (aggressive, abusive, offensive) contents on online platforms [4, 5]. However, most of the resources developed for resource-rich language like English, Chinese, Arabic and other European languages [6, 7]. Nevertheless, people usually interact via their regional language to carry out day-to-day communication. System trained in resource-rich languages can not be directly replicated to detect aggressive/abusive texts written in the local language. Therefore, it is a prerequisite to develop resources, techniques, and regional language tools to reduce the effect of undesired texts.

Unfortunately, despite being the seventh most widely spoken language globally, Bengali is considered one of the notable resource-constrained languages [8]. Statistics reveal that more than 45 million users on Facebook and YouTube are using Bengali daily. Most of these users commonly interact on social media via the textual form. Many textual interactions contain hostile contents that cause the significant rise of hate, abuse, cyberbullying and aggression on social media. Thus, to ensure the quality of textual conversation and reduce unlawful activities on these platforms, developing an automated Bengali language system that can identify these aggressive activities is mandatory. Such a system will flag posts/comments that convey any aggressiveness that might threaten national security, try to break communal harmony, and publicize distorted propaganda. However, developing a system to detect aggressive textual conversation in a resource-constrained language like Bengali is challenging. The scarcity of benchmark dataset and deficiency of language processing tools are the key barriers to develop such a system in Bengali. Complicated morphological structure, presence of ambiguous words, diversities in different dialects and rich variations in the constituent parts of a sentence have made the task more complicated. Bengali has a rich vocabulary and unique writing script which has no overlap with other resource-high languages. Moreover, multilingual code-mixing in social media texts has added a new challenge to the existing task [9]. Therefore, the key research questions we are investigating in this paper are-"**RQ1:** How can we successfully develop an aggression annotated dataset in the Bengali language?". "**RQ2:** How can we effectively identify potential aggressive texts and categorize them into predefined aggression categories?"

This work develops a Bengali aggressive text dataset by analyzing aggressive and non-aggressive texts' properties to address the above research questions. Various aspects of the dataset are also explained to get better insights. Several machine learning (ML), deep learning (DL) and transformer-based techniques are investigated to build the aggressive text identification and classification system. Exploring the models' outcomes, this work proposes a weighted ensemble technique that exploits the best performing models' strength. Finally, we investigate the proposed model's results and errors and compare it with other existing techniques. Major contributions of this work can be illustrated in the following:

- **Dataset:** present a new Bengali aggressive text dataset which contains 6807 aggressive and 7351 non-aggressive texts. Furthermore, by employing a hierarchical annotation schema, aggressive texts are annotated into religious, political, verbal and gendered aggression classes.
- **Insights:** provide useful insights and detailed statistics of the data that ensure the quality of the dataset.
- **Model:** develop a weighted ensemble model using m-BERT, distil-BERT, Bangla-BERT, XLM-R to identify and categorize aggressive Bengali texts. The proposed model emphasizes the participating classifiers' softmax probabilities based on their previous performance on the dataset. This weighting technique outperforms the simple average ensemble approach and enhances the classifier performance in the developed dataset.
- **Benchmarking:** investigate and compare the performance of the proposed model with other ML, DL baselines and existing techniques, thus setting up a benchmark work to compare in the future.

[1] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

- **Error analysis:** deeply analyze the results and errors of the proposed model. Presents qualitative and quantitative analysis that shed light on the reasons behind some of the errors and provide a few directions that might help to mitigate the system's deficiency.

This research is one of the pioneering works that aims to identify and classify aggressive texts in Bengali as per our exploration. We expect that the resources developed in this work will pave the way for aggressive text classification researchers in Bengali. The remaining of the paper is organized as follows. Section 2 discusses the studies related to unwanted text detection and classification on online platforms. The detailed definition of aggressive text and its categories included in Section 3. Dataset development steps are described in Section 4. Section 5 presents the analysis and statistics of the developed dataset. Section 6 illustrates the techniques adopted to develop the proposed system. Experimental findings with quantitative and qualitative error analysis are reported in Section 7. Section 8 points out the future scopes with concluding remarks.

2. Related work

Over the last few years, a significant amount of work has been carried out to identify and categorize unwanted texts on various online platforms such as twitter, facebook, reddit and so on. Works included aggression classification [10, 11, 12], hate speech detection [13, 14, 15], abuse detection [16, 17, 18], toxicity classification [19, 20], misogyny [21, 22], trolling identification [23, 24], cyberbullying detection [25, 26], and offensive text classification [27, 28, 29]. Although most of the researches focused in English however a considerable body of work has been conducted for other languages too. This part briefly describes the researches related to aggression, hate, offense detection/classification with other co-related phenomena concerning non-Bengali and Bengali languages.

English: Kumar et al. [10] present an aggressive language identification dataset that has three category: *overt, covert, non-aggressive.* The Dataset contains 15k aggression annotated comments/posts written in English and Hindi. Aroyehun et al. [30] develop deep neural network-based models on English with data augmentation and pseudo labelling strategy. Their system achieved macro f_1-score of 0.64 and 0.59 using LSTM and CNN-LSTM methods, respectively. Risch et al. [31] employ bootstrap aggregating based ensemble with multiple fine-tuned BERT on TRAC-2 [32] dataset to identify aggression and misogyny. They obtain an 80.3% weighted f_1-score on the test set of English social media posts. Zampieri et al. [33] compile an offensive language identification dataset *(OLID)* of 14k English tweets. They used a three-layer hierarchical annotation schema to detect, categorize and identify the target of texts whether it attack individuals or a group of people. Baseline evaluation is performed using SVM, BiLSTM and CNN. In all three levels, CNN outperforms others by achieving macro-f_1 of 0.80, 0.69 and 0.47. Fortuna et al. [34] offer a dataset of 80000 tweets labelled with four categories: *hateful, abusive, spam and normal.* They performed a holistic approach to identify confusion among various categories. Davidson et al. [35] develop a hate speech dataset of 25k tweets with three categories: *hate, offence and neither.* Logistic regression with tf-idf and n-gram features obtains the best macro f_1-score of 0.90.

Hindi: Mathur et al. [36] introduce a Hindi-English code switched dataset of 3.6k tweets split into three categories: *abuse, hate speech, and non-offensive.* They proposed a system based on CNN and transfer learning which achieves f_1-score of 71.4%. Aggression annotated Hindi-English code mixed Dataset of 21k Facebook comments and 18k tweets are developed by Kumar et al. [37]. Instances were labelled with three top-level tags and ten discursive classes. Annotation performed by four annotators where the inter-annotator agreements were 72% and 57% on the top-level and 10 class annotations. Bhardwaj et al. [38] presented a multi-label hostility detection dataset of 8.2k online posts in Hindi. Dataset divided into five dimensions: *fake, hate, offensive, defamation, and non-hostile* where SVM achieved the highest weighted f_1-score of 84.11% with m-BERT embedding.

Arabic: Mulki et al. [39] built a dataset of 5.8k tweets in the Arabic Levantine dialect and manually annotated tweets into *abusive, hate or normal* classes. Their system achieved f_1-score of 89.6% in binary *(abusive or normal)* and 74.4% in ternary *(abusive, hate or normal)* classification scenario using naive Bayes (NB). Mubarak et al. [40] present a dataset to classify Arabic tweets into *offensive, obscene or clean* categories. Their system obtains a maximum of f_1-score of 0.60 with the combination of seed words and unigram features. Hassan et al.

[41] employ ensemble technique over SVM, CNN-BiLSTM and m-BERT to identify offensive Arabic texts in OffensEval-2020 dataset. They used character n-grams, word n-grams, character and pre-trained word embedding features. The system acquired of 90.16% f_1-score on the Arabic test set.

Spanish & German: Carmona et al. [42] manually annotated 11k Mexican Spanish tweets into *aggressive* and *non-aggressive* classes. They organized a task over this dataset and lexicon-based approach [43] obtained the best performamnce with a macro f_1-score of 0.62. Few tasks organized at GermEval [44, 45] aimed to classify German tweets into *offensive* and *non-offensive* classes. The top performance achieved f_1-score of 76.77% using feature ensemble method on a dataset of 8.5k German tweets.

Portuguese: Leite et al. [46] presented a toxic language dataset (ToLD-Br) composed of 21k tweets. They manually annotated tweets into seven classes: *LGBTQ+phobia, racism, insult, xenophobia, obscene, misogyny, and non-toxic.* Their system obtained macro-f_1 of 76% using BERT models. An offensive dataset consisting of 1250 comments is developed by Pelle et al. [47]. They split offensive texts into six fine-grained labels: *racism, sexism, xenophobia, LGBTQ+phobia, cursing, and religious intolerance.* N-gram features with SVM achieved the best f_1-score ranging from 77% to 82%. Fortuna et al. [48] presented a Portuguese hate speech dataset consisting of 5668 tweets. The Dataset was labelled into binary *(hate or non-hate)* and 81 hierarchical categories. LSTM with pre-trained word embedding acquires 78% f_1-score on the binary labels.

Multilingual: In recent years, a series of the shared task and academic events have organized, focusing on multilingual identification and classification of aggressive, abusive, offensive and hatred contents in social media. Shared task on trolling, aggression and cyberbullying (TRAC-1 [10]) aims to classify English and Hindi texts into overtly, covertly and non-aggressive classes. In the second iteration (TRAC-2 [32]), they added Bengali texts with an additional task of identifying gendered aggression. The best outcome was achieved with variants of transformer models [49, 50]. OffensEval-2020 [51] provided manually annotated offensive texts in five different languages *(English, Arabic, Turkish, Greek and Danish)* that follow the hierarchical annotation schema of 'OLID'. The top system of all the languages have employed ensemble technique with fine-tuned transformers [52, 53]. HASOC-2020 [15] offered hate and offensive language dataset in *Tamil, Malayalam, Hindi, English and German* to perform two tasks. At first, identify hate or offensive posts and further categorize them into *hate, offense* and *profane* classes. The best system for Hindi, German and English achieved 0.53, 0.52 and 0.51 macro f_1-score, respectively. Other notable works included *Automatic Misogyny Identification* [54], *Workshop on Abusive Language* [55] and *HatEval* [56] that investigated hate speech against women and immigrants in Spanish and English.

Bengali: Identification and categorization of aggressive texts in Bengali is an open avenue for future research. Due to the scarcity of benchmark dataset, linguistic tools and other resources, no significant works have been carried out to date in this arena. However, with multi-lingual and cross-lingual models' arrival, few works have been conducted recently related to the detection/classification of hate, aggression, offence, and abuse. Ranasinghe et al. [57] developed a model to classify aggressive Bengali texts into overtly, covertly and non-aggressive classes. They used 4k texts from the Bengali dataset presented in the TRAC-2 shared task [32]. Their system achieved the highest weighted f_1-score of 84.23% by leveraging inter-language transfer strategy with XLM-R. Karim et al. [58] collected 3k Bengali text samples and categorized them into four hatred classes: *political, personal, geopolitical, religious.* They used an ensemble of BERT variants to develop their system and obtained a 0.88 f_1 score. The class definitions provided by the authors may result in contradiction. An instance may be expressed political and religious hate simultaneously. No insight on the countermeasure is provided during such situations. Romim et al. [59] presented a hate speech dataset which contains 30k with *hate* or *non-hate* comments crawled form Facebook and YouTube. The baseline system obtained 87.5% f_1-score using SVM. A recent work [60] presented a logistic regression-based model to classify *suspicious* and *non-suspicious* Bengali texts. Five different ML algorithms are applied on the extended Dataset of 7k texts [61]. SGD classifier with tf-idf features obtains the best accuracy of 84.57%. Emon et al. [62] develop an abusive Bengali dataset of 4.7k texts consisting of seven classes (*slang, religious-hatred, political-hatred, personal attack, anti-feminism, neutral, positive*). The model gained 82.2% accuracy by utilizing LSTM.

An SVM based system is developed to identify the threat and abuse from Bengali texts [63], which achieved an accuracy of 78% on a dataset of 5644 texts. In our previous work, we develop an aggressive text identification and classification dataset of 7591 texts where combined CNN, BiLSTM methods obtained the highest weighted f_1-score [64]. Here, we perform experimentation with a wide range of methods on the extension of the existing dataset.

Availability of a standard dataset is the prerequisite to develop any classification system. Previous research in Bengali mainly focuses on classifying hatred and abusive contents using ML and other feature-based methods. None of the research has been conducted to identify aggression and categorize aggressive texts into fine-grained classes in Bengali. Therefore, to perform the aggressive text identification and classification in Bengali, we develop a dataset (named 'BAD') using a hierarchical annotation schema. Computational systems developed over other languages and datasets can not be replicated directly on a new dataset. The main reason is that models available in one language would not be able to capture the features in another language without proper modifications in the model architecture (i.e., no. of layers, no. of neurons, no. of filters) and fine-tuning of hyperparameters (i.e., learning rate, batch size, dropout rate, epochs, optimizer). Therefore, the proposed weighted ensemble method optimizes the various hyperparameters to perform the aggressive text identification and classification tasks in Bengali more efficiently and providing more insights than existing techniques.

3. Definition of the Task

This work aims to develop an aggressive text identification and classification system that can detect whether a potential text $t_i \epsilon T$ is aggressive or not from a set of m texts, $T = \{t_1, t_2, ..., t_m\}$ in the first phase. In the next phase, the system categorizes the aggressive texts into one of n predefined aggression classes, $AC = \{ac_1, ac_2, ..., ac_n\}$. The task of the system is to assign at_i automatically to ac_j where at_i and ac_j represents the aggressive text and aggressive class, respectively. In order to accomplish the task, dataset is split into two levels using hierarchical annotation schema [33]: (A) coarse-grained identification of aggressive texts (B) fine-grained categorization of aggressive texts. This section defines the aggressive texts and their fine-grained classes to perform the tasks mentioned above.

3.1. Level A: Aggressive Text Identification

Determining whether a text is aggressive or not aggressive is very ticklish, even for language and psychology experts due to its subjective nature. People may define aggression in different ways, which leads to the heterogeneous interpretation of aggression. One person may contemplate a piece of text as aggressive, while another may consider it as usual. Moreover, overlapping characteristics of aggression with hate speech, cyber-bullying, abusive, offence and profanity have made this task more complicated and challenging. Understanding the phenomena of aggression in a better way requires a large amount of literature study in aggression and impoliteness from psychological and linguistic perspectives. However, this task's aim is much simpler, and this work performs a surface level classification of aggressive text on social media. Thus, it is monumental to define aggressive text first to implement the aggressive text classification system successfully. To do this, several pieces of literature have been explored to interpret the aggression, incitement, violence, suspicion, and hatred contents from different sources. Table 1 presents a summary of the definitions culled from various trending social networking sites, human rights organizations, psychological and scientific studies.

Baron et al. [73] defined aggression as a behaviour that expresses the desire to harm another individual verbally, physically, and psychologically. The distinction between physical, verbal, and relational aggression exhibited by Buss et al. [74]. Kumar et al. [37] discriminated overtly and covertly aggressive texts. In overtly aggressive texts, aggression expressed directly with the strong verbal attack. While covertly aggressive texts attack the victim in rhetorical queries, satire, metaphorical reference, and sarcasm. The majority of these statements provide the broader prospect of aggression from images, videos, texts and illustrations. However, this work focuses on detecting and classifying aggression from textual contents only. Analyzing the interpretation of aggression and exploration of literature lead us to distinguish between aggressive and non-aggressive texts as follows:

- Aggressive texts **(AG)**: Text contents that incite, attack, or wish to harm an individual, group or community based on some criteria

Table 1: Definitions of aggression, incitement, violent and hatred contents according to different scientific studies, human rights organizations and various social networking sites.

Source	Definition
Anderson et al. [65]	"Language that used toward other individuals with the intent to cause harm".
Facebook [66]	"Contents that attack or pose credible threats to personal or public safety, facilitate high severity violence, misinformation and unverifiable rumours that contribute to risk of imminent violence".
Torres et al. [67]	"Aggressive language intents to hurt or harm an individual or a group by referring to or exciting violence".
Nobata et al. [68]	"Language which attacks or demeans a group based on race, ethnic origin, religion, gender, age, disability, or sexual orientation/gender identity".
YouTube [69]	"Contents that promote violence or hatred against individual or groups, based on age, sexual orientation, religion, disability, nationality etc".
Council of Europe (COE) [70]	"Expression which spread, incite, justify or promote violence, hatred and discrimination against a person or group of persons for variety of reasons".
Paula et al. [71]	"Language that glorify violence and hate, incite people against groups based on religion, ethnic or national origin, physical appearance, gender identity or other".
Roy et al. [72]	"Aggressive language directly attack group or person using abusive words, comparing in a derogatory manner or support false attack toward others".

such as religious belief, gender, sexual orientation, political ideology, race, nationality and ethnicity.

- Non-aggressive texts **(NoAG)**: Text contents that do not contain any statement of aggression or express hidden intention to harm an individual, group or society.

3.2. Level B: Fine-grained Categorization

In recent years, the thriving interest in aggression/abuse from various perspectives have created a conglomeration of typologies and terminologies. Few works attempted to provide a uniform understanding of this complex phenomenon. Waseem et al. [75] proposed two-level categorization of abusive online language: nature of the abuse (implicit or explicit) and the target of the abuse (group or individuals). However, Kumar et al. [10] pointed out that in the majority of the abusive instances, individuals and groups are targeted simultaneously. Therefore, it would not be wise to distinguish between these classes while annotating many instances. The authors suggested that the distinction between various abuse/aggression dimensions can be made considering the attack's locus such as gender, religion, specific ideology, politics, race, and ethnicity [10, 37]. Most previous works in Bengali [58, 62] illustrated that political, gendered, verbal and religious abuse/offence classes are occurred more frequently in Bengali texts than others (such as racial, geographic). Furthermore, our exploration revealed that a higher amount of Bengali texts are available in four coarse categories: political, religious, verbal, and gendered aggression. Therefore, this work also concentrated on these four aggression dimensions due to their much textual contents availability. As these classes interpretation varies considerably across individuals, it is essential to draw a fine line among these aggression categories. In order to minimize the bias as well as overlap during annotation after analyzing existing research on aggression detection [76, 37, 77, 78], toxicity classification [46, 79, 80], hate speech identification [81, 82, 83], abuse detection [84, 85, 86], cyber-bullying categorization [87, 25] and other related terminologies guided us to make a distinction between aggression classes as the following:

- Religious aggression **(ReAG)**: incite violence by attacking religion (Islam, Hindu, Catholic, and Jew), religious organizations, or religious belief of a person or a community.

- Political aggression **(PoAG)**: provoke followers of political parties, condemn political ideol-

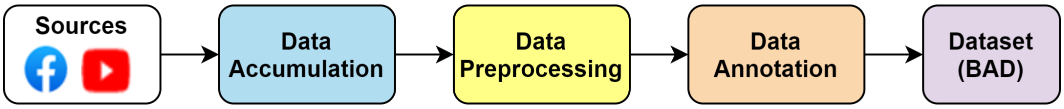


Figure 1: Dataset development steps. BAD stands for "Bengali Aggressive Text Dataset".

ogy, or excite people in opposition to the state, law or enforcing agencies.

- Verbal aggression **(VeAG)**: damage social identity and status, describe a wish to harm or do evil of the target by using nasty words, curse words and other outrageous languages.
- Gendered aggression **(GeAG)**: promote aggression or attack the victim based on gender, contain an aggressive reference to one's sexual orientation, body parts or sexuality, or other lewd contents.

To the best of our knowledge, no research has been conducted yet classifying aggressive Bengali texts into these fine-grained classes.

4. Dataset Development

As per our exploration, none of the datasets on aggressive Bengali text is available that deals with the defined fine-grained class instances. Therefore, we develop a Bengali aggressive text dataset (called 'BAD') to serve our purpose. To develop 'BAD', we have followed the directions given by Vidgen and Derczynski [88]. Figure 1 illustrates the data collection and annotation pipeline. The detailed discussion on the dataset development process described in the following subsections.

4.1. Data Accumulation

A total of **14443** aggressive and non-aggressive texts are accumulated manually from various social media platforms. Most of the dataset instances are collected from Facebook and YouTube since majority of the Bengali social media users are active on these platforms. According to social media stats[2], 94.88% and 2.68% social media users in Bangladesh use Facebook and YouTube. Although the recent statistics exhibited a rise in Twitter users in Bangladesh, the people mostly use English for social communication. Due to the scarcity of Bengali texts related to the aggressive contents, the current work did not consider Twitter data. Dataset utilized in this work was acquired from July 1, 2020, to February 25, 2021. Within this duration, we have considered only those texts that were composed after June 30, 2019. Strategies that followed to collect aggressive and non-aggressive texts illustrated in the following:

- For ***aggressive texts***, the general approach is to collect the posts and comments that incite violence or express aggression. Additionally, we analyze the replays of aggressive posts/comments. In a significant number of cases, we found that to counter an aggressive comment; people use another aggressive comment. Furthermore, to get additional aggressive texts, the user's timeline is scanned who like, share or comment in support of aggression-related posts.

  Most religious aggressive data is collected from the comment threads of YouTube channels and Facebook pages concerning religion. The majority of the gendered aggression expressed in social media is against women compared to the male counterpart. Texts related to this category are accumulated from various domains such as fitness videos, fashion pages, and media coverage on celebrities/women. Texts that use curse/outrageous words and wish to do evil to others added into the verbal aggression category. Politically aggressive texts procured from Facebook pages of political parties, pages of their supporters and opposition parties, influential political figures, and people's reaction to the government's different policies.

- ***Non-aggressive texts*** are cumulated from the news/posts related to science & technology, entertainment, sports and education. The primary sources of these data are Facebook pages and YouTube channels of popular Bangladeshi newspapers (such as Somoy-news, Prothom-Alo, Jamuna-tv). Table 2 illustrates the popularity and activity status of these sources. The data was collected only

[2] https://gs.statcounter.com/social-media-stats/all/bangladesh

Table 2: Statistics of few sources from where data were gathered. Here FP, YC indicates Facebook page and YouTube channel respectively.

Name	Type	Affiliation	Popularity (No. of followers/ subscribers)	Reactions per post (in avg.)	Activity (frequency of posting)
Prothom Alo	FP/YC	Newsgroup	15M	5k	200 post/day
Rafiath Mithila	FP	Artist	3M	25k	3 post/week
Mizanur Azhari	YC	Religious speaker	1.67M	30k	1 post/week
Jamuna tv	YC	Media	7.69M	4k	50 post/day
Asif Mohiuddin	FP	Public figure	118k	1.5k	1 post/day
Awami League	FP	Political org.	799k	3.5k	13 post/day
Salman BrownFish	YC/FP	Musician	2.6M	20k	2 post/week
Pinaki Bhattacharya	FP	Author	342k	9k	6 post/day
Somoynews tv	YC/FP	Media	7.8M	1K	150 post/day
Basher kella	FP	Political	42k	300	20 post/day

from the Facebook and YouTube pages of the newsgroups. None of the data is accumulated from news portals. Moreover, while procuring aggressive texts, we found plenty of non-aggressive examples and added them into this category.

The potential texts are manually accumulated from more than 100 Bengali Facebook pages and YouTube channels affiliated with media, political organizations, authors, artists, and newsgroups to develop the dataset. Table 2 illustrates detailed statistics to understand the quality of the data gathered from Facebook and YouTube platforms[3]. Data were culled from only those threads that received at least 200 reactions (like, comment or share) in total. We did not use any list of keywords or phrases to collect data.

4.2. Data Preprocessing

To reduce the annotation effort and remove inconsistencies, few preprocessing filters are applied to the accumulated texts. Steps have followed in processing the texts are,

- All the flawed characters (#@!&%) dispelled from the texts.
- As concise texts do not contain any meaningful information, the text having a length of fewer than three words are discarded.
- Texts written in languages other than Bengali and duplicate texts are removed.

We eliminated **94** texts in this step, and the remaining **14349** texts are passed to the human annotators for manual annotation.

4.3. Data annotation

"How to achieve the correct annotation" is one of the most crucial questions to answer when labelling a training dataset [89]. Therefore, to clarify the queries regarding annotation in this part, we recapitulate the annotators' identity, annotation guidelines, and data labelling process that we pursued to develop 'BAD'.

4.3.1. Identity of the annotators

Bedner and Friedman [90] emphasize knowing about the identity of the annotators since their perception and experience might influence the annotations. Binns et al. [91] pointed out that in the context of online abuse, the gender of the annotators has an impact on the annotations. Moreover, a homogeneous group of annotators might not capture all the examples of aggression and abuse [92]. To mitigate these issues, we choose annotators from different racial, residential and religious backgrounds. Five annotators carry out manual annotation: two undergraduate, two graduate and one academic expert. Experience, expertise and other relevant demographic information about the annotators are presented in Table 3.

All of the annotators are native Bengali speakers. Some key characteristics of undergraduate

[3]To develop the dataset, we have considered only the public posts/comments from these sources. The source pages or channels might contain personal information; thus, we avoided disclosing the source link.

Table 3: Summary of the demographic information, field of research, research experience and personal experience of aggression in social media of the the annotators. Here AN, OA denotes annotator and online aggression respectively.

	AN-1	AN-2	AN-3	AN-4	Expert
Research-status	Undergrad	Undergrad	RA	RA	Professor
Research-field	NLP	NLP	NLP	NLP	NLP, HCI, Robotics
Experience	1 year	1 year	2 years	3 years	20 years
Age	23	23	24	26	46
Religion	Islam	Islam	Hindu	Islam	Islam
Gender	Male	Female	Male	Male	Male
Viewed OA	yes	yes	yes	yes	yes
Targeted by OA	no	yes	no	yes	yes

and graduate annotators are: a) age between 22-26 years, b) field of research NLP and experience varies from 1-3 years, c) do not have extreme perspective about religion, d) not a member of any political organization e) active in social media and view aggression in these platforms. Although while selecting the annotators, we tried to keep demographic aspects balanced; however, the annotators pool is still biased in religion (Islam) and gender (Male).

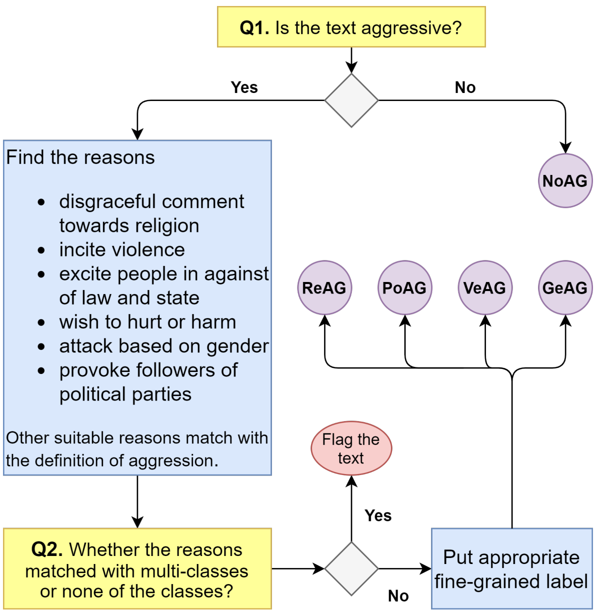


Figure 2: Guidelines for data annotation. Reasons denotes a subset of possible reasons.

Algorithm 1: Final label assigning process

Input: Set of texts with initial labels
Output: Aggressive text dataset with final labels

$T \leftarrow \{t_1, t_2, ..., t_m\}$ (set of accumulated texts);
$BAD \leftarrow []$ (Bengali aggressive text dataset);
$FL \leftarrow []$ (final class labels);
$IL[m][2] \leftarrow \{a_1, a_2, .., a_m\}$ (initial labels);
$D \leftarrow []$;

for $t_i \epsilon T$ **do**
    $l_1 = IL[i][1]$ (first label);
    $l_2 = IL[i][2]$ (second label);
    if ($l_1 == flag$ & $l_2 == flag$) **then**
        //text is discarded;
    else if ($l_1 == l_2$) **then**
        $BAD.append(t_i)$;
        $FL.append(l_1)$;
    else
        $D.append(t_i)$ (disagreement: put this text in separate list);
    end
    $i = i + 1$;
end
for $d_j \epsilon D$ **do**
    1. expert discuss with annotators;
    2. based on discussion either add d_j to '$BAD = []$' with final label or discard it;
    $j = j + 1$;
end

4.3.2. Annotation Guidelines

To ensure the quality of annotation and better understand the dataset, it is crucial to provide detailed guidelines for annotation [93]. In few cases, dataset creators had given the liberty to the annotators to apply their perspective [48]. However, it is risky since individual interpretation and perceptions vary considerably. We ask the annotators to follow the process depicted in Figure 2 during annotation to avoid such issues.

To determine the initial label at first, we have to identify whether a text is aggressive or not. If it is non-aggressive, then put the label NoAG. However, if it is aggressive, we need to ascertain the reasons. In case the reasons match with multiple or none of the defined aggression dimensions *flag* the text for further discussion. Otherwise, assign an appropriate fine-grained (ReAG, PoAG, VeAG, GeAG) label. Prior annotation, we provide few samples of each category to the annotators and explain why an example should be labelled with a specific class. Each processed texts labelled by two annotators, and in case of disagreement, the expert resolved the issue through discussion.

After receiving the initial label, we follow the algorithm 1 to set the final labels. For each text t_i, we check the two initial labels l_1 and l_2. A text is discarded when both of the initial labels contain *flag*. If l_1 and l_2 match, then the text and associated label added into the final lists. When disagreement is raised expert discusses with the annotators whether to keep or remove the text. The final label of such text also decided on the discussion. For **105** texts, we observe overlap among aggression dimensions and **86** texts do not fall into any defined aggression categories. Table 4 shows few examples with the reasoning that have been discarded due to overlap among aggression dimensions and other disagreements. Since these numbers are deficient, such instances are not included in the current corpus. We plan to address this issue in future when we attain a significant number of such instances. Finally, we get the aggressive text ('BAD') containing **14158** processed and annotated texts.

5. BAD: Bengali Aggressive Text Dataset

Further analysis is performed to understand the properties of the dataset. This section presents the various statistical analysis of 'BAD'[4].

[4] **Disclaimer:** Authors would like to state that the comments/examples referred to in this section presents as they were accumulated from the original source. Authors do not use these examples to hurt individuals or a community. Moreover, authors do not promote aggressive language usage, and this research work aims to mitigate the practice of such language.

Table 4: Few examples of excluded texts. Label 'flag' indicates that the expressed aggression does not match with any predefined aggression classes and remarks provided by the expert reveals the reasons for discarding the samples.

Text	Label	Remarks
আমাদের এলাকা হলে আমরা নিজেরাই ওই হিন্দুদের মার্ডার করে দিতাম। (If it was our area, we would have killed those Hindus by ourselves)	VeAG, ReAG	describe a wish to harm Hindus
এই সরকার কয়েক বছর ক্ষমতায় থাকলে বাংলাদেশে কেও আর ধর্ম পালন করতে পারবে না (If this government stays in power for a few years, no one will be able to practice religion in Bangladesh)	ReAG, PoAG	incite people in opposition to state misusing the religion
আমরা কোন সংসদে নারি মন্ত্রী দেখতে চাইনা, হোক আওয়ামী লীগ বা বিএনপি (We do not want to see any women ministers in the parliament, be it from Awami League or BNP)	PoAG, GeAG	discrimination towards women from political perspective
ধর্ষণ কারির মৃত্যুদণ্ড চাই (I want the death penalty for the rapist)	flag	aggression against a person who commit hateful crime
বাংলাদেশ থেকে টিকটক নামে বিষধর অ্যাপটিকে চিরতরে বন্ধ করা হোক (Let ban the poisonous TikTok app forever from Bangladesh)	flag	disgust against app or media
নারী মানে কলঙ্ক। সব নষ্টের মূলে নারী। (Women mean stigma. Women are the root of all evil)	GeAG, VeAG	verbal attack toward women

Table 5: Few examples of annotation divergence. Label-1 and label-2 denotes the first and second annotations for each text.

Text	Label-1	Label-2
মুসলিম উম্মাহ কে ধ্বংস করার জন্য একদল নারীবাদী উঠে পড়ে লেগেছে (A group of feminists has risen up to destroy the Muslim Ummah)	GeAG	ReAG
আওয়ামীলীগের লোকেরা জাহান্নামী, কারণ কুরআন হাদিস আওয়ামী লীগের ভদ্রলোকেরা মানেনা (Ihe people of Awami League are hellish because the gentlemen of Awami League do not accept Quran and Hadith)	PoAG	ReAG
এই যুগের মেয়েরা এক একটা ডাইনি, এমন মেয়েদের পুড়িয়ে মারা দরকার (The girls of this age are witches, they need to be burnt to death)	VeAG	GeAG
চাল চুরি করা এই সরকারের ঐতিহ্য। শূধু চাল নয় ভোট ও চুরি করে তারা (Stealing rice is the tradition of this government. Not just rice, they steal vote as well)	VeAG	PoAG

5.1. Annotation Quality

Two annotators labelled each instance of the dataset, and an expert resolved the issue through deliberations and discussions when disagreement raised between them. To check the validity and quality of the annotations, we measured the inter-rater agreement. Cohen's kappa coefficient [94] is used to calculate the agreement between annotators (equation 1).

$$k = \frac{O(a) - H(ca)}{1 - H(ca)} \tag{1}$$

Here, $O(a)$ and $H(ca)$ denoted the observed and hypothetical chance of agreement between annotators. Table 6 presents the kappa score on each annotation level.

Table 6: Kappa score on each level of annotation.

	Class	K-score	Mean
Level-A	NoAG	0.87	0.80
	AG	0.73	
Level-B	ReAG	0.54	0.65
	PoAG	0.63	
	VeAG	0.72	
	GeAG	0.69	

The highest agreement of 0.87 is achieved for NoAG class, which exhibits that this class has a more distinctive lexicon compare to other classes. Among the fine-grained classes, the maximum and minimum k-score of 0.72, 0.54 are obtained for VeAG and ReAG classes. Investigation reveals that in many cases, the aggression was expressed covertly, which is difficult to classify. This covert form of expression may be a reason behind the low agreement in fine-grained classes. The mean k-score in coarse-grained classes is 80%, while fine-grained classes obtained the mean k-score of 65%. These scores indicate substantial agreement between the annotators. Table 5 shows few instances for which disagreement occurred during annotation.

5.2. Dataset Statistics

This work's main objective is to detect aggressive texts and categorize them into one of the fine-grained classes. The developed (BAD) uses to build the computational models. For training and evaluation, the dataset split into three sets: train (80%), validation (10%) and test (10%). Instances of the dataset are shuffled randomly before partitioning to eliminate bias and ensure randomness. Table 7 illustrates a summary of the dataset. Out of 14158 texts, 7351 texts are labelled as NoAG, while the remaining 6807 texts belong to the AG class. Aggressive texts are further categorized into fine-grained classes where religious, political, verbal and gendered aggression classes have 2217, 2085, 2043 and

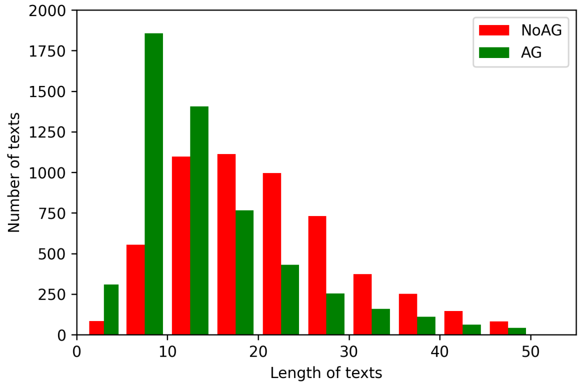
Table 7: Summary of the train, validation and test set

Class	Train	Valid	Test	Total
NoAG	5845	769	737	7351
AG	5481	647	679	6807
ReAG	1794	210	213	2217
PoAG	1655	229	201	2085
VeAG	1629	194	220	2043
GeAG	368	48	46	462

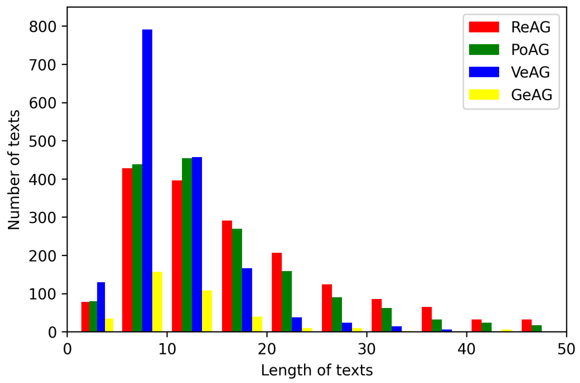
Table 8: Statistics of the training set. Here MTL, ANW, ANUW stands for maximum text length, average number of words and average number of unique words respectively.

	Level-A		Level-B			
	NoAG	**AG**	**ReAG**	**PoAG**	**VeAG**	**GeAG**
Total words	160745	78714	32282	26099	16378	3955
Unique words	26804	16155	8294	6819	5214	1738
MTL (words)	635	132	98	132	60	44
ANW (per text)	27.50	14.36	17.99	15.76	10.05	10.74
ANUW (per text)	4.59	2.95	4.62	4.12	3.20	4.72

462 text samples.



(a) Coarse-grained classes



(b) Fine-grained classes

Figure 3: Number of text fall into various length range for different classes in training set

Since the classifier models learn from the training set examples to acquire more valuable insights, we further investigated this set. Detailed statistics of the training set presented in Table 8. From the distribution, it notices that the training set is highly imbalanced for coarse-grained as well as fine-grained classes. There is a significant difference between aggressive and non-aggressive class in level-A in terms of the number of total words and total unique words. The NoAG class has a total of 160k words, while the AG class contained only 78k words. On average, NoAG class contained 4.6, and the AG class hold 2.9 unique words per text. In level-B, ReAG class has two and eight times as many as total words compare to VeAG and GeAG classes. ReAG consisting the maximum (18), and VeAG contained the minimum (10) number of words per text. On average, all the fine-grained classes have four unique words in each text.

In-depth investigation of the training set texts length reveals some interesting facts. Figure 3 depicts the number of texts vs the length of texts distribution of the training set for coarse-grained and fine-grained classes. It observed that the aggressive texts tend to be shorter than the non-aggressive ones. Approximately 4000 aggressive texts have less than 20 words among 6807 aggressive texts. On the other hand, $\approx$ 4500 non-aggressive texts have a length of higher than 20 words among 7351 non-aggressive texts. Only a fraction of texts has more than 40 words. In level-B, most of the fine-grained class texts have a length of 8 to 15 words. Several texts in PoAG and ReAG classes are approximately similar in every length range.

Table 9: *Jaccard* similarity between pair of coarse-grained and fine-grained classes. (c1) NoAG; (c2) AG; (f1) ReAG; (f2) PoAG; (f3) VeAG; (f4) GeAG.

Level-A			Level-B				
	c1	**c2**		**f1**	**f2**	**f3**	**f4**
c1	-	0.39	f1	-	0.40	0.24	0.35
c2	-	-	f2	-	-	0.23	0.30
			f3	-	-	-	0.33

For quantitative analysis, the *Jaccard* similarity

Table 10: Some examples of BAD. Level-A and level-B indicates coarse-grained and fine-grained class labels.

Text	Level-A	Level-B
ধর্ম পালন করা মানে শয়তানের উপাসনা করে। আমাদেরকে ধর্ম থেকে দূরে থাকতে হবে (Practicing religion means worshiping Satan. We have to stay away from religion)	AG	ReAG
দেশকে এই সরকারের হাত থেকে মুক্ত করতে হলে যুদ্ধ ছাড়া কোনো উপায় নেই নেই (There is no way to free the country from this government without war)	AG	PoAG
তুই দেশের বাইরে আছিস বলে এখনও বেছে আছিস।তোর সাহস থাকলে বাংলাদেশ আয় তোকে সবার সামনে হত্যা করব (You are still alive because you are out of the country. If you have the courage, come to Bangladesh. I will kill you in front of everyone)	AG	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
হাজারো সালাম জানাই শিক্ষকদের, যাদের অবদানে এগিয়ে যাচ্ছে বাংলাদেশ (Thousands of salutations to the teachers, who are helping Bangladesh to move forward)	NoAG	-

is calculated between 200 most frequent words of each class. The similarity values between each pair exhibited in Table 9. ReAG-PoAG pair obtain the highest similarity score of 0.40. VeAG has maximum similarity with GeAG, while GeAG has more words in common with ReAG. Table 10 shows a few annotated samples of the BAD.

6. System Overview

This work's primary concern is to identify the aggressive texts (task-A) and categorize them into four fine-grained aggression classes (task-B): ReAG, PoAG, VeAG and GeAG. To accomplish these tasks, we develop computational models using various machine learning, deep learning and transformer based methods. This section briefly describes the methods and techniques employed to address the tasks. Figure 4 shows the schematic diagram of the system. Parameters and architectures of different approaches have discussed in the subsequent subsections.

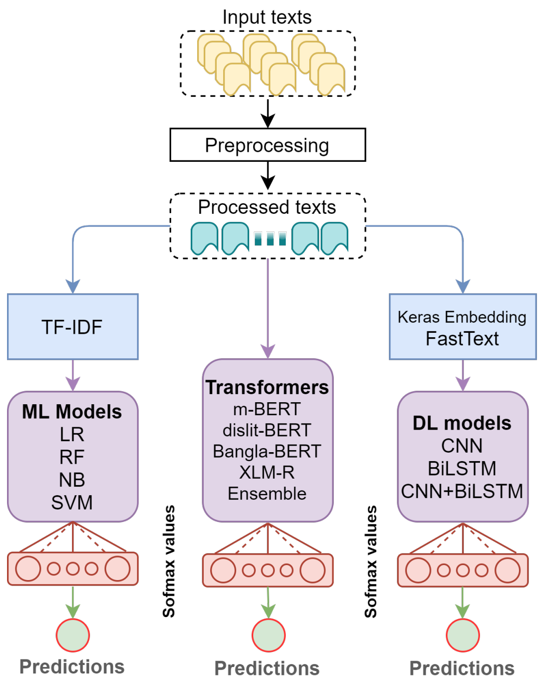


Figure 4: Abstract process diagram of Bengali aggressive text identification and categorization system.

6.1. Preprocessing and Feature Extraction

Raw input texts contain noises such as punctuation, digits, unwanted symbols and characters written in other languages than Bengali. All of these were removed during the preprocessing step. Various techniques such as TF-IDF and FastText word embedding are applied to extract the texts' relevant features [95].

- **TF-IDF**: To train the ML-based methods, we extract the n-gram features of the texts using the term frequency-inverse document frequency technique [96] (TF-IDF). A combination of unigram and bigram features are utilized for both tasks. To reduce the computation, 20k and 10k most frequent features are considered for task-1 and task-2, respectively.

Inverse document reweighting technique is enabled while maximum and minimum document frequency value settled to 1.

- **Word Embedding**: Although TF-IDF is an effective feature extraction (FE) technique, it could not hold the words' semantic information. Therefore, the word embedding technique is employed to capture the semantics of the words regarding the context [97]. Default Keras embedding layer used to obtain the embedding features. Texts are needed to be converted into fixed-length numeric sequences to acquire the features. Therefore, a vocabulary of x unique words is created where the value of x is set to 35000 and 16000 for task-1 and task-2, respectively. To achieve numeric mapping, words in a text are replaced by the word's index in the vocabulary. Since each text has a different number of words, we get a variable-length sequence which is not suitable for feature extraction. Using the Keras pad-sequences method, each sequence is converted into a fixed-length vector of size l. The value of l is set to 70 for task-1 and 50 for task-2. Extra values are removed from the long sequences, and short sequences are padded with the value 0. The Embedding layer converts a text of length l into a matrix of size $l * e$. Here, e indicates the embedding dimension that determines the word's length of the embedding vector. For both task embedding dimension value is set to 100.

  FastText: The Keras embedding layer could not handle the out of vocabulary words. It set the vectors of those words to 0. The FastText embedding technique [98] is used to alleviate this problem. This technique holds the subword information since words are represented as the sum of character n-grams. This work uses the pre-trained word vectors of Bengali where the embedding dimension settled to 300 [99].

6.2. Methodology

Four ML methods (such as LR, RF, NB, SVM), three deep learning techniques (such as CNN, BiLSTM, CNN+BiLSTM) and four transformer-based models (m-BERT, distil-BERT, Bangla-BERT, XLM-R, ensemble) are implemented to investigate the Bengali aggressive text classification task performance.

6.2.1. Machine Learning Models

The various parameters are tuned to prepare the LR, RF, NB and SVM models before performing the classification task. A summary of the parameters adopted for each model presented in Table 11.

Table 11: Parameter summary of ML models.

Model	Parameters
LR	optimizer='lbfgs', regualizer='l2', C=1.0, max_iter=400
RF	criterion='gini', n_estimators=100, max_features=no_of_features
NB	α=1.0, class_prior=none, fit_prior=true
SVM	kernel='rbf', γ='scale', tol='0.001', random_state=0

All four ML models utilize the combination of unigram and bigram features extracted by the TF-IDF technique. In the LR model, the 'lbfgs' optimizer is used with 'l2' regularizer. The inverse regularization strength is fixed to 1.0, and a maximum of 400 iterations are taken for solvers to converge. RF is implemented with 100 trees, and the 'gini' criterion is utilized to measure the quality of split in the tree. An internal node is partitioned if there exist at least two samples. All the system features are considered during node partitioning. The additive smoothing parameter of the NB model is set to 1. Prior class probabilities are settled based on the number of instances in the class. For SVM, the 'rbf' kernel is used with 'l2' penalizer and kernel coefficient value decided using the number of features. Tolerance of stopping criterion and random state set to 0.001 and 0, respectively.

6.2.2. Deep Learning Models

Keras and FastText embedding are used to develop deep learning models that have been applied successfully to offensive text classification [100], hostility detection [101] and aggressive texts categorization [102]. Hyperparameters and their corresponding values significantly effect DL models performance [103, 104]. Due to linguistic diversity, one model developed for a particular language can not perform similarly in another language. Thus, DL models should be prepared with their optimized hyperparameters depending on the task and language types. The preparation of DL models for Bengali aggressive text classification illustrates in the following:

Table 12: Hyperparameter summary of DL models. C+B denotes combined CNN, BiLSTM method.

Hyperparameter	**Hyperparameter space**	**CNN**	**BiLSTM**	**C+B**
Input length	-	70(task-A), 50 (task-B)		
Embedding dimension	[32, 64, 100, 128, 200, 256, 300, 400]	100(task-A), 300 (task-B)		
Filters (layer-1)	[8, 16, 32, 64 128]	128	-	128
Kernel size	[3, 5, 7]	3	-	3
Filters (layer-2)	[16, 32, 64 128, 256]	64	-	-
Pooling type	'max', 'average'	'max'	-	'max'
LSTM cell (layer-1)	[8, 16, 32, 64 128]	-	128	64
Dropout rate	[0.1, 0.15, 0.20, 0.25, 0.30, 0.35]	-	0.2	0.2
LSTM cell (layer-2)	[16, 32, 64 128, 256]	-	64	32
Learning rate	[0.3, 0.2, 0.1, 0.001, 0.0001, 0.00001]	0.001		
Optimizer	'adam', 'Nadam', 'RMSprop'	'adam'		
Batch size	[8, 16, 32, 64, 128]	16		
Epochs	-	30		

CNN: Embedding features are propagated into a two-layer CNN architecture. The first and second layers contain 128 and 64 filters, respectively. Each layer consisting of kernels size (3×3) and features are downsampled by max-pooling technique with a $(1{\times}3)$ size window. Softmax layer take features from CNN to make the prediction. To add non-linearity *'relu'* activation function is used.

BiLSTM: a BiLSTM architecture is used to capture long-range dependencies and hold information from both past and future. Like CNN, it also has two layers where the first layer has 128 and the second layer has 64 bidirectional LSTM cells. The dropout value settled to 0.2, and the features passed to the softmax layer for prediction.

CNN + BiLSTM: in the combined method, CNN and BiLSTM added sequentially with slight modifications in their previous architecture. One layer of CNN with 128 filters and a kernel size of $(3{\times}3)$ is used. Features from CNN are downsampled using a pooling layer and propagated through two layers of BiLSTM. The first layer has 64, and the second layer has 32 LSTM units. The dropout rate is unaltered, and hidden representation is passed to the softmax layer.

The input sequence length is set to 50 and 70 for task-A and task-B. These values are fixed based on the insights from length analysis shown in Figure 3. The dimension for Keras and FastText embedding settled to 100 (task-A) and 300 (task-B), respectively. The rest of the architecture is similar for both types of tasks. All the models use the 'adam' optimizer with a learning rate of 0.001. Models are trained with 16 samples per batch for 30 epochs. The model with the highest validation accuracy is stored using callbacks. Table 12 summarizes hyperparameter values used by the DL models. Experimentation was performed using the values from the hyperparameter space. Optimum hyperparameter values have been settled in a trial and error fashion depending on the validation set outcomes.

6.2.3. Transformer Models

Past studies reveal that the transformer models trained in monolingual, multi-lingual or cross-lingual settings are achieving the state of the art performance in categorising unwanted texts [51, 32, 15]. Thus, this work employed four pre-trained transformer models: Multilingual Bidirectional Encoder Representations from Transformers (m-BERT) [105], distilled version of BERT (distilBERT) [106], Bangla-BERT [107] and cross-lingual version of Robustly Optimized BERT (XLM-R) [108]. By varying hyperparameters, these models are fine-tuned over the (BAD). Models are fetched from the HuggingFace[5] library and built with ktrain packages [109].

Multilingual-BERT is a large model which has been trained with 104 monolingual datasets. We use the 'bert-base-multilingual-uncased' model with 12 layers, 12 heads, and 110M parameters. The model is fine-tuned by altering the learning rate, batch size and epochs. A distilled version of m-BERT is utilised, having six layers, 768 dimensions and 12 heads. This model reduced the computational time and preserved the overall system

[5]https://huggingface.co/models

performance up to 95%. The system also trained with the base version ('distilbert-base-multilingual-cased') fetched from the HuggingFace library. Another pre-trained model, 'bangla-bert-base', is also implemented. This model is trained with monolingual Bengali CommonCrawl corpus and utilises the BERT base model's architecture. XLM-R is a cross-lingual model which outdoes m-BERT in various benchmarks. This model is built over the of 100 languages and has 12 layers, eight heads and approximately 125M parameters. We implement the 'xlm-roberta-base' model for our purpose. Table 13 shows a list of parameters for BERT variants.

Table 13: Fine-tuned parameter values of transformers.

Hyperparameter	**Value**
Fit method	'auto_fit'
Learning rate	$2e^{-5}$
Epochs	20
Batch size	12
Max sequence length	50, 70

All the models are fine-tuned on BAD using the ktrain 'auto_fit' method. Models are trained for 20 epochs with a batch size of 12. A triangular learning rate policy is adopted with a maximum learning rate of $2e^{-5}$. Max sequence length for the texts is settled to 50 for task-1 and 70 for task-2. Model weights are stored using checkpoint, and the best model is chosen based on its efficiency in the validation set.

6.2.4. Proposed Ensemble Model

Recent works exhibited that the ensemble of transformers can significantly improve the efficiency of a classification task [110, 111]. Ensemble methods exploit the strength of the individual models and increase the system's predictive accuracy. Four transformer models are used (m-BERT, distil-BERT, Bangla-BERT, XLM-R) that is fine-tuned on the developed dataset. Figure 5 shows the architecture of the proposed weighted ensemble technique. This work employs two types of ensemble techniques: average (A-ensemble) and weighted (W-ensemble). The average (A) ensemble computes the average of the softmax probabilities of the participating models. This averaging technique considers a class with the maximum probability as the output class. In this method, prior results of the base classifiers is not considered [112, 113]. On the other hand, this work proposes a weighted ensemble technique which strengthen the classifiers' performance to identify and categorize Bengali aggressive texts.

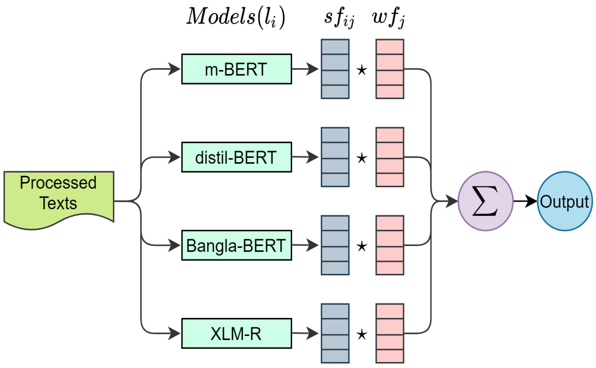


Figure 5: Architecture of the proposed model.

Algorithm 2: Process of W-ensemble

Input: Softmax probabilities and WF score
Output: Predictions of the W-ensemble

$sp \leftarrow []$ (softmax probabilities);
$wf \leftarrow []$ (weighted f_1 scores);

$sum = []$ (weighted sum);
for $i\epsilon(1, m)$ **do**
    for $j\epsilon(1, l)$ **do**
        $sum[i] = sum[i] + (sp_{ij}[] * wf_j)$;
        $j = j + 1$;
    end
    $i = i + 1$;
end

$n\_sum = 0$;
for $j\epsilon(1, l)$ **do**
    $n\_sum = n\_sum + wf_j$;
    $j = j + 1$;
end

$P = (sum/n\_sum)$ //normalized probabilities;
$O = \arg\max(P)$ // set of predictions;

Rather than simple or traditional averaging, the proposed method offers an additional weight to the softmax probabilities of a model based on its prior results. Lets consider, we have 'l' existing models and 'm' validation/test set instances. A model classifies each instances m_i into one of n predefined classes. For each m_i, a model l_j provides a softmax probability vector of size 'n', $sp_{ij}[n]$. Thus, models output becomes: $\langle sp_{11}[],\ sp_{21}[],\ ...,\ sp_{m1}[]\rangle$, $\langle sp_{12}[],\ sp_{22}[], ..., sp_{m2}[]\rangle$,.., $\langle sp_{1l}[],\ sp_{2l}[], ..., sp_{ml}[]\rangle$. Prior weighted f_1-scores of 'l' models measured on the validation set are $wf_1, wf_2, ..., wf_l$. Utilizing these values, the proposed technique computes the output as described in Eq. (2).

$$O = \arg\max\left(\frac{\forall_{i\epsilon(1,m)} \sum_{j=1}^{l} sp_{ij}[n] * wf_j}{\sum_{j=1}^{l} wf_j}\right) \quad (2)$$

Here, O denotes the vector of m, which contains the ensemble method's predictions.

Algorithm 2 describes the process of calculating ensemble weights. Softmax probabilities of the models are aggregated after multiplying with the WF scores. Probabilities are normalized by dividing with the sum of WF scores. Finally, output predictions are computed by taking the maximum from the probabilities.

7. Experiments and Results

This section presents a comprehensive performance analysis of the approaches that we employed for Bengali aggressive text classification. Various evaluation measures and the outcomes of the different models will be described here subsequently. Moreover, this section explains the proposed model's error analysis and compares its performance with other existing techniques.

7.1. Experimental Setup and Evaluation Measures

Experiments carried out on Google colaboratory platform with python 3 Google cloud engine backend (GPU). A 12.5GB RAM and 64GB disk space have been utilized to implement the models. To process and prepare the data, we used pandas (1.1.4) and numpy (1.18.5). The machine learning models are built with scikit-learn (0.22.2) packages, while the training of DL models is performed using Keras (2.4.0) and TensorFlow (2.3.0). Transformer models are developed with ktrain (0.25) packages [109].

Since the models explored in this work is computationally intensive, therefore a brief analysis of their complexities presented for better understanding. Table 14 provides the number of trainable parameters of the deep neural networks and transformer models as well as reports their execution time on this experimental setup. As the training set of coarse-grained classification is much bigger, its complexity is also higher than the fine-grained classification task. Although the pre-trained models performed better, their execution time is 4-5 times higher than the custom deep neural networks. Among the models, XLM-R has the highest number of parameters and also requires the highest execution time.

The train, validation and test instances are utilized to develop the models. It ensured that all the instances of these sets are mutually exclusive. Models learn from the training set instances while the hyperparameter values are settled based on the validation set. Finally, the trained models are eval-

Table 14: Computational complexity of deep neural networks and transformer models. Execution time reported here is for completing 30 epochs (deep neural networks) and 20 epochs (transformers) in the GPU facilitated Google colab platform. Here task-A, task-B indicates coarse-grained and fine-grained classification task respectively.

	Task-A		**Task-B**	
Method	**Trainable Parameters**	**Execution Time**	**Trainable Parameters**	**Execution Time**
CNN (C)	3564450	13min 6s	1664196	7min 6s
BiLSTM (B)	3899106	19min 45s	1999364	5min 6s
C+B	3678690	16min 18s	1778820	4min 12s
CNN (FastText)	10641250	35min 6s	4940996	10min 6s
BiLSTM (FastText)	11103906	42min 12s	5404164	10min 1s
C+B (FastText)	10755490	40min 18s	5055620	9min 9s
m-BERT	167357954	1h 16min 29s	167359492	35min 20s
distil-BERT	135326210	46min 35s	135327748	20min 2s
Bangla-BERT	164398082	1h 15min 40s	164398082	36min 40s
XLM-R	278045186	1h 33min 1s	278046724	42min 13s

uated using the unseen instances of the test set. Various statistical measures are used to calculate and compare the performance of the systems. Few measures utilized for evaluation illustrated in Eqs. (3)-(6).

- Precision: calculate the number of samples (s_i) actually belong to class (c) among the samples (s_i) labeled as class (c).

$$P = \frac{True\ positive}{True\ positive + False\ positive} \tag{3}$$

- Recall: calculate how many samples (s_i) are correctly labeled as class (c) among the total number of samples (s_i) of class (c).

$$R = \frac{True\ positive}{True\ positive + False\ negative} \tag{4}$$

- Error: gives the value that how many samples are wrongly classified.

$$E = \frac{False\ positive + False\ negative}{Number\ of\ samples} \tag{5}$$

- F_1-score: calculated by simply averaging precision and recall ($F = \frac{2PR}{P+R}$). Since the dataset

Table 15: Evaluation results of different models on the test set for coarse-grained identification of aggressive texts. FT denotes FastText embedding and the superiority of the models determined based on WF scores.

		NoAG			AG			
	Method	**P**	**R**	**F**	**P**	**R**	**F**	**WF**
	LR	0.89	0.91	0.90	0.90	0.88	0.89	0.8968
	RF	0.80	0.91	0.85	0.89	0.76	0.82	0.8370
	NB	0.90	0.89	0.90	0.88	0.89	0.89	0.8913
	SVM	0.88	0.93	0.90	0.92	0.86	0.89	0.8953
	CNN (C)	0.92	0.90	0.91	0.90	0.92	0.91	0.9110
	BiLSTM (B)	0.92	0.90	0.91	0.89	0.92	0.90	0.9061
	C+B	0.90	0.92	0.91	0.91	0.89	0.90	0.9067
	CNN (FT)	0.91	0.91	0.91	0.90	0.90	0.90	0.9053
	BiLSTM (FT)	0.89	0.93	0.91	0.91	0.87	0.89	0.8995
	C+B (FT)	0.93	0.87	0.90	0.87	0.93	0.90	0.8997
	m-BERT (MB)	0.95	0.90	0.92	0.90	0.95	0.92	0.9223
	distil-BERT (DB)	0.91	0.93	0.92	0.92	0.90	0.90	0.9145
	Bangla-BERT (BB)	0.92	0.91	0.92	0.91	0.91	0.91	0.9124
	XLM-R (XR)	0.93	0.93	0.93	0.92	0.93	0.92	0.9272
A-ensemble models	MB+DB	0.92	0.92	0.92	0.91	0.91	0.91	0.9166
	MB+BB	0.94	0.92	0.93	0.91	0.94	0.92	0.9251
	MB+XR	0.94	0.93	0.94	0.92	0.94	0.93	0.9329
	DB+BB	0.92	0.92	0.92	0.91	0.92	0.92	0.9187
	DB+XR	0.93	0.93	0.93	0.92	0.92	0.92	0.9265
	BB+XR	0.94	0.93	0.94	0.93	0.93	0.93	0.9321
	MB+DB+BB	0.94	0.92	0.93	0.91	0.94	0.93	0.9286
	MB+DB+XR	0.94	0.93	0.94	0.92	0.94	0.93	0.9323
	DB+BB+XR	0.94	0.92	0.93	0.92	0.94	0.93	0.9308
	MB+DB+BB+XR	0.94	0.93	0.94	0.92	0.94	0.93	0.9336
W-ensemble models	MB+DB	0.92	0.91	0.92	0.92	0.91	0.91	0.9173
	MB+BB	0.94	0.93	0.93	0.92	0.94	0.91	0.9258
	MB+XR	0.94	0.93	0.94	0.92	0.94	0.93	0.9329
	DB+BB	0.93	0.92	0.92	0.92	0.92	0.92	0.9209
	DB+XR	0.93	0.93	0.93	0.93	0.92	0.92	0.9279
	BB+XR	0.93	0.94	0.94	0.93	0.94	0.93	0.9336
	MB+DB+BB	0.94	0.92	0.93	0.92	0.94	0.93	0.9287
	MB+DB+XR	0.94	0.93	0.94	0.93	0.94	0.93	0.9332
	DB+BB+XR	0.94	0.92	0.93	0.94	0.92	0.93	0.9315
	MB+DB+BB+XR	0.95	0.93	0.94	0.92	0.94	0.93	**0.9343**

is imbalance we calculate the weighted f_1-score which is defined as,

$$WF = \frac{1}{N}\sum_{i=1}^{c} F_i n_i, \quad N = \sum_{i=1}^{c} n_i \qquad (6)$$

Here N, F_i and n_i denotes total samples in test set, f_1-score and number of samples in class (i).

The weighted f_1-score (WF) is considered to determine the superiority of the models. Other scores such as precision, recall, error rate are also reported to get an understanding of the model's performance on different classes.

7.2. Results

The current work investigated all possible combinations of the base classifiers (i.e., transformers) for both tasks (i.e., fine-grained and coarse-grained). Table 15 exhibits the outcomes of the

Table 16: Evaluation results of various models on the test set for fine-grained classification. AE, WE, F represents A-ensemble, W-ensemble and FastText embedding respectively.

	ReAG			**PoAG**			**VeAG**			**GeAG**			
Method	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**WF**
LR	0.87	0.90	0.89	0.91	0.93	0.92	0.86	0.90	0.88	0.75	0.39	0.51	0.8689
RF	0.89	0.74	0.81	0.82	0.88	0.85	0.76	0.94	0.84	0.83	0.33	0.47	0.8088
NB	0.79	0.90	0.84	0.88	0.91	0.89	0.84	0.87	0.86	0.00	0.00	0.00	0.8049
SVM	0.84	0.89	0.87	0.90	0.92	0.91	0.82	0.91	0.86	1.00	0.13	0.23	0.8342
CNN (C)	0.89	0.87	0.88	0.93	0.89	0.91	0.81	0.89	0.85	0.54	0.41	0.47	0.8504
BiLSTM (B)	0.88	0.88	0.88	0.90	0.91	0.90	0.84	0.87	0.85	0.65	0.52	0.58	0.8569
C+B	0.84	0.89	0.86	0.91	0.93	0.92	0.86	0.87	0.86	0.38	0.24	0.29	0.8412
CNN (FT)	0.89	0.85	0.87	0.94	0.87	0.90	0.83	0.89	0.86	0.50	0.59	0.54	0.8524
BiLSTM (FT)	0.86	0.89	0.87	0.89	0.92	0.91	0.90	0.85	0.87	0.61	0.59	0.60	0.8641
C+B (FT)	0.90	0.88	0.89	0.90	0.94	0.92	0.85	0.90	0.87	0.67	0.43	0.53	0.8691
m-BERT (MB)	0.92	0.92	0.92	0.97	0.96	0.96	0.91	0.88	0.90	0.60	0.74	0.66	0.9073
distil-BERT(DB)	0.90	0.92	0.91	0.93	0.92	0.93	0.85	0.92	0.88	0.72	0.39	0.51	0.8794
Bangla-BERT(BB)	0.94	0.96	0.95	0.96	0.95	0.95	0.90	0.92	0.91	0.74	0.61	0.67	0.9176
XLM-R (XR)	0.93	0.93	0.93	0.97	0.97	0.97	0.89	0.92	0.90	0.71	0.59	0.64	0.9146
A-ensemble models													
MB+DB	0.91	0.92	0.91	0.97	0.96	0.96	0.90	0.90	0.90	0.65	0.65	0.65	0.9059
MB+BB	0.93	0.96	0.95	0.98	0.97	0.97	0.91	0.92	0.91	0.73	0.65	0.69	0.9271
MB+XR	0.93	0.93	0.93	0.98	0.97	0.97	0.91	0.91	0.91	0.67	0.70	0.68	0.9210
DB+BB	0.93	0.96	0.94	0.96	0.97	0.97	0.91	0.94	0.92	0.75	0.52	0.62	0.9216
DB+XR	0.93	0.95	0.94	0.96	0.97	0.96	0.89	0.93	0.91	0.75	0.52	0.62	0.9144
BB+XR	0.94	0.96	0.95	0.97	0.97	0.97	0.90	0.92	0.91	0.71	0.63	0.67	0.9241
MB+DB+BB	0.91	0.94	0.93	0.97	0.97	0.97	0.90	0.93	0.91	0.76	0.57	0.65	0.9167
MB+DB+XR	0.93	0.92	0.93	0.98	0.97	0.97	0.90	0.93	0.91	0.70	0.67	0.69	0.9203
DB+BB+XR	0.94	0.95	0.95	0.97	0.97	0.97	0.90	0.95	0.92	0.73	0.59	0.65	0.9266
MB+DB+BB +XR	0.94	0.96	0.95	0.98	0.97	0.97	0.91	0.93	0.92	0.75	0.63	0.69	0.9275
W-ensemble models													
MB+DB	0.92	0.92	0.92	0.97	0.96	0.96	0.90	0.90	0.90	0.67	0.72	0.69	0.9123
MB+BB	0.94	0.96	0.95	0.98	0.97	0.98	0.91	0.92	0.92	0.69	0.67	0.70	0.9301
MB+XR	0.93	0.94	0.93	0.98	0.97	0.97	0.91	0.91	0.91	0.68	0.70	0.69	0.9223
DB+BB	0.93	0.95	0.94	0.96	0.97	0.97	0.91	0.94	0.92	0.75	0.52	0.62	0.9202
DB+XR	0.93	0.94	0.93	0.97	0.97	0.97	0.89	0.93	0.91	0.75	0.59	0.66	0.9172
BB+XR	0.94	0.96	0.95	0.97	0.97	0.97	0.90	0.92	0.91	0.69	0.59	0.64	0.9206
MB+DB+BB	0.91	0.94	0.93	0.97	0.97	0.97	0.90	0.92	0.91	0.74	0.57	0.64	0.9154
MB+DB+XR	0.93	0.92	0.93	0.98	0.97	0.97	0.90	0.93	0.91	0.70	0.67	0.69	0.9203
DB+BB+XR	0.94	0.95	0.95	0.97	0.97	0.97	0.90	0.94	0.92	0.78	0.63	0.69	0.9308
MB+DB+BB +XR	0.94	0.96	0.95	0.98	0.97	0.97	0.92	0.93	0.92	0.74	0.63	0.68	**0.9311**

developed models for coarse-grained classification. Among the ML models, LR acquired the highest weighted f_1-score (WF) of 0.8968. NB and SVM also achieved a higher than 89% WF score. In the case of DL, results revealed that models with Keras embedding achieved better scores in the classification task. Surprisingly, all the DL models' efficiency reduced by $\approx$ 1% after using FastText embedding. CNN with Keras embedding gained the maximal WF score of 0.9110 amid the DL models. A significant rise is observed in the system performance with transformer models. Multilingual BERT and XLM-R models attain a higher than 92% WF score. Initially, we have evaluated a total of 14 models using several statistical measures (such as precision, recall, f1-scores) on the dataset (BAD) and empirically observed each of them. In particular, based on the highest weighted f1-score, we selected four base models (m-BERT, distil-BERT, Bangla-BERT, XLM-R) for the ensemble. All possible combination of these base models are investigated for both average and weighted ensemble technique. As per expectation, we noticed an increase in the performance where the average ensemble method attained maximum WF score of 0.9336. Finally, by employing the proposed weighted ensemble method, the system obtains the highest WF score of **0.9343**, which outperformed all other models.

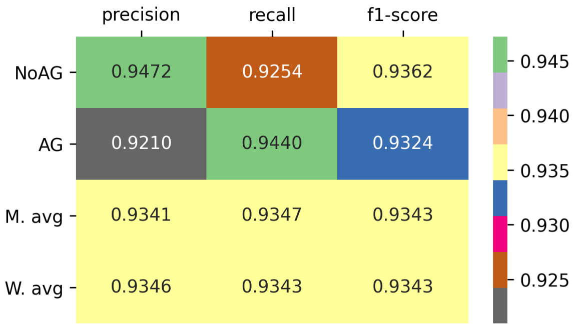
Evaluation results for fine-grained classification presented in Table 16. Like coarse-grained classification, LR also achieved the maximum WF score amid the ML models in fine-grained classification. Interestingly LR outdoes the CNN and BiLSTM models implemented with Keras embedding in fine-grained classification by attaining a 0.8689 WF score. Although in coarse-grained classification, DL models performed poorly with FastText embedding. However, in fine-grained classification, the DL models obtained a higher WF score with FastText. Among the ML and DL models, combined CNN+BiLSTM acquired the maximum of 0.8691 WF score. Transformer based methods also showed noteworthy performance. Bangla-BERT achieved the maximum WF score (0.9176) amid the BERT variants. However, the proposed weighted ensemble method surpasses all other models and achieves the highest WF score of **0.9311** for fine-grained classification. Thus, it is confirmed that the performance of the proposed system has significantly improved on both tasks after employing the weighted ensemble technique. This higher performance might happen because the weighting technique can adjust the softmax probabilities of the base classifiers of the ensemble depending on their prior results.

The final model is cross-validated to acquire better insight regarding the proposed model's performance. For ease of analysis, we only cross-validated the four base transformers and the proposed combination for both the A-ensemble and W-ensemble techniques. A 10-fold cross-validation technique [114] has been carried out on a combined (training + validation) set using *scikit-learn*. Table 17 represents the cross-validation results of the models. The W-ensemble technique has achieved the highest mean weighted f_1-scores of 92.85% (task-A) and 92.21% (task-B). The average standard deviation is approximately 3% for both tasks. The analysis of cross-validation results revealed that the model's performance had not significantly affected by the dataset split.

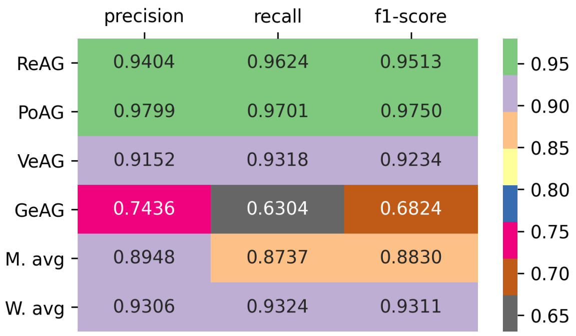
Table 17: 10-fold cross-validation results of the transformer-based models, including the proposed technique on the combined (training + validation) set. The values in the cell represent the weighted f_1-scores, and *Std* denotes standard deviation.

	Task-A		**Task-B**	
Method	**Mean**	**Std**	**Mean**	**Std**
m-BERT(M)	0.9223	0.0291	0.9102	0.0235
dislit-BERT(D)	0.9145	0.0294	0.8777	0.0381
Bangla-BERT(B)	0.9124	0.0248	0.9167	0.0304
XLM-R(X)	0.9262	0.0279	0.9139	0.0355
A(M+D+B+X)	0.9268	0.0267	0.9205	0.0263
W(M+D+B+X)	0.9285	0.0254	0.9221	0.0266

To get more insights, we take a closer look at the proposed model's classification reports shown in Figure 6. In coarse-grained classification, NoAG class has the higher (0.9472) precision while AG has the higher (0.9440) recall value. Since both classes have approximately similar instances, no meaningful difference is observed between the macro and weighted f_1-score. Among the fined-grained classes, GeAG and PoAG obtained the minimum (0.6824) and maximum (0.9750) f_1-scores. The performance of the proposed model (W-ensemble) is lower in the GeAG than in other classes. The limited number of instances in GeAG class has resulted in a reduced performance than others. Moreover, the confusion matrix analysis and the Jaccard similarity index revealed that GeAG class mostly overlaps with the VeAG class. Therefore, the overall misclassification rate increased and hence the performance of the W-



(a) Coarse-grained



(b) Fine-grained

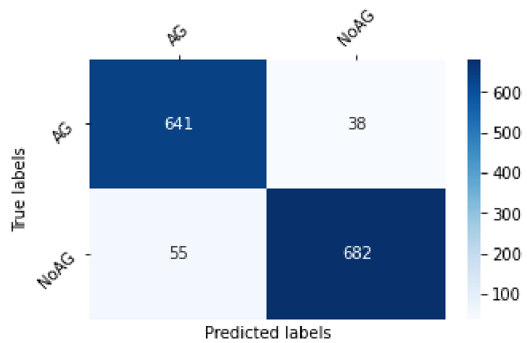
Figure 6: Classification report of the proposed (w-ensemble) model on the test set. M. avg and W. avg denote macro and weighted average values.

ensemble model is decreased in GeAG class. Results noticed a $\approx 5\%$ difference in macro and weighted f_1 score values as the classes are highly imbalanced.

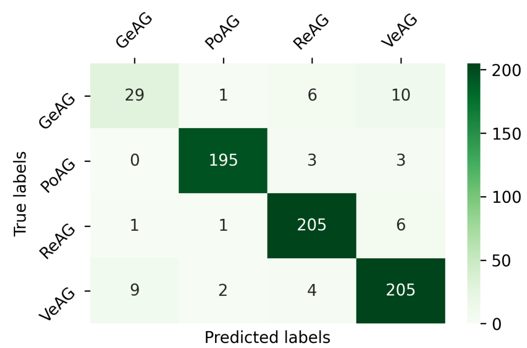
In summary, LR achieves the highest WF score among the ML models in both tasks. CNN and CNN + BiLSTM attain the maximum WF score with Keras and FastText embedding, respectively. It noticed that when the number of classes increases, the efficiency of ML and DL models decreases. However, the performance of the transformer models remains consistent. This consistency occurred due to the massive number of examples usage by pre-trained models, so their generalization capability is much higher. The results showed that the proposed weighted ensemble method outperformed all ML, DL and transformer models in coarse and fine-grained classification. The proposed architecture's ability to emphasize the models' softmax predictions based on their prior results might be the reason behind this superior performance.

7.3. Error Analysis

It is evident from Tables 15 and 16 that the weighted ensemble is the best performing model to classify aggressive texts in Bengali. Here, a detailed error analysis is carried out quantitatively and qualitatively to acquire in-depth insights into individual model's performance.



(a) Coarse-grained



(b) Fine-grained

Figure 7: Confusion matrix of the proposed (w-ensemble) model on the test set

Quantitative analysis: Figure 7a shows the confusion matrix for coarse-grained classification. It indicates that 38 instances of AG class wrongly classified as NoAG whereas 55 comments of NoAG class labelled as AG by the W-ensemble model. Some aggressive texts express aggression implicitly, which is very difficult to identify. Figure 7b indicated that the false-negative rate in GeAG and PoAG classes are higher than the false positive rate. In contrast, false-positive values in ReAG and VeAG classes are higher. The model classifies 195, 205 and 205 instances correctly among 201, 213 and 220 instances of PoAG, ReAG and VeAG classes. In the case of the GeAG class, the proposed model performed poorly. It incorrectly classifies 17

Table 18: Error rate of various models in Task-A (coarse-grained) and Task-B (fined-grained).

Method	Task-A (%)	Task-B (%)
LR	10.31	12.50
RF	16.17	18.24
NB	10.88	16.62
SVM	10.45	14.71
CNN (C)	8.9	14.71
BiLSTM (B)	9.39	14.12
C+B	9.32	15.15
CNN (FT)	9.46	15.00
BiLSTM (FT)	10.03	13.53
C+B (FT)	10.03	12.56
m-BERT (MB)	7.77	9.26
distil-BERT (DB)	8.55	11.46
Bangla-BERT (BB)	8.76	8.09
XLM-R (XR)	7.27	8.38
A-ensemble models		
MB+DB	8.33	9.41
MB+BB	7.48	7.21
MB+XR	6.72	7.93
DB+BB	8.12	7.5
DB+XR	7.34	8.24
BB+XR	6.71	7.5
MB+DB+BB	7.13	8.09
MB+DB+XR	6.76	7.94
DB+BB+XR	6.92	7.43
MB+DB+BB +XR	6.64	7.06
W-ensemble models		
MB+DB	8.26	8.82
MB+BB	7.42	6.98
MB+XR	6.72	7.79
DB+BB	7.91	7.64
DB+XR	7.19	8.09
BB+XR	6.64	7.78
MB+DB+BB	7.13	8.24
MB+DB+XR	6.76	7.93
DB+BB+XR	6.84	6.90
MB+DB+BB +XR	6.56	6.76

texts among 46 test texts. All the classes mostly make confusing with VeAG classes. The presence of outrageous words in all the aggressive classes may cause this confusion. The error rate for all models in coarse and fine-grained classification is presented in Table 18. The proposed weighted ensemble technique is achieved the lowest error rate of 6.56% (task-A) and 6.76% (task-B).

Qualitative analysis: Table 19 shows a few examples which exhibit the contrasting nature of the transformer models. Although the models quantitatively achieve similar scores, their class predictions are qualitatively different. One model can classify a test sample correctly while another can not. The proposed transformer-based weighted ensemble method can be helpful to deal with this contrasting nature. For better understanding, outputs of the ensemble models are further investigated. Table 20 illustrates some examples of incorrect classification on test data.

Analysis of the incorrect predictions revealed that it is arduous to identify those texts that implicitly propagate or express aggression. Such instances do not contain any aggressive references or words; therefore, it is difficult to flag them. On the other hand, some texts sarcastically use aggressive words with no intention to harm or do evil, but the model wrongly classifies them as aggressive. It is challenging to identify and classify such text samples from the surface level analysis without understanding the context. Moreover, some words are frequent in both aggressive and non-aggressive classes. The presence of such words in a text creates confusion and makes the task more complicated. Contextual analysis of aggressive texts, adding their meta-information, and more training data might improve the classification performance of the proposed model.

7.4. Comparison between the proposed and existing methods

As per this work exploration, no significant work has been conducted to categorize aggressive texts into fine-grained classes, including dataset development in Bengali. Therefore, this research adopted several recent techniques that have been explored on similar tasks in other language's datasets. For consistency, previous methods [115, 57, 116, 117] have implemented on the developed dataset (i.e., BAD) and compared their performance with the proposed technique. Table 21 shows the comparison in terms of weighted f_1-score for coarse-grained and fined-grained classification.

Kumari et al. [115] develop a model on TRAC-2 dataset [32] using LSTM and FastText embedding to classify aggressive Bengali texts. One layer of LSTM with 192 unit is used where dropout and recurrent dropout value set to 0.2. We obtained a WF score of 90.54% (coarse-grained) and 81.20% (fine-grained) by mimicking their architecture. On the same dataset, Ranasinghe et al. [57] applied inter-language transfer strategy along with XLM-R. After employing XLM-R, the system achieved a

Table 19: Instances exhibiting the contrasting nature of the transformer models. MB, DB, BB and XR denotes predicted labels for multilingual-BERT, distil-BERT, Bangla-BERT, XLM-R models. **A** indicates the actual labels and the wrong predictions are marked in bold.

Example	MB	DB	BB	XR	A
এই অপরাধ এর একটাই সাজা প্রকাশে ফাঁসি দেয়া (The only punishment for this crime is hanging)	**PoAG**	VeAG	VeAG	VeAG	VeAG
মানুষ এখন হিংস্র জানোয়ার (Humans are now ferocious beasts)	**ReAG**	GeAG	**ReAG**	**VeAG**	GeAG
জয় মা কালি বলা, এটা একটা রোগ (Saying joy ma Kali is a disease)	ReAG	ReAG	ReAG	**GeAG**	ReAG
ফ্যাসিবাদের মূখে গনতন্ত্রের কথা মানায় না (In the face of fascism, democracy is not acceptable)	**GeAG**	**ReAG**	PoAG	PoAG	PoAG

Table 20: Few examples that are incorrectly classified by the proposed weighted ensemble model. P and A denotes predicted and actual labels respectively.

Text	A	P
নিজের মা বোনদের সম্মান কর (Respect your mother and sisters)	NoAG	GeAG
পুলিশ পাহারার বাহিরে এসে কথা বলে দেখ তখন বুঝবে আমরা কারা (Come out of the police guard and talk then you will understand who we are.)	VeAG	PoAG
ধর্ম মানুষে মানুষে বিভেদ সৃষ্টি করে সব অশান্তির পেছনে কারন ধর্ম (Religion is the cause of all the unrest that divides people)	ReAG	NoAG
এসব ফালতু মেয়েদের জন্য রাস্তাঘাটে সমস্যা হয় (These bad girls create problems in the road)	GeAG	VeAG
বাংলাদেশে নির্বাচন আর প্রহসন একই কথা। সরকার ই সব ক্ষমতার মালিক। (Election and farce are the same thing in Bangladesh. The government owns all the power)	PoAG	NoAG

Table 21: Comparison between proposed and existing techniques in terms of weighted f_1-score of the models on BAD.

Technique	Coarse-grained	Fine-grained
Kumari et al. [115]	90.54	81.20
Ranasinghe et al. [57]	92.71	91.45
Baruah et al. [116]	89.31	84.01
Nayel et al. [117]	89.89	85.98
Proposed	93.43	93.11

WF score of 92.71% and 91.45%. The other two works [116, 117] used SVM with tf-idf and other parameter combination to classify aggressive and offensive languages. These methods gained lower accuracy on BAD than other methods in both classification tasks. The comparative analysis shows that the proposed technique outperformed the existing techniques by acquiring the highest weighted f_1-score of 93.43% and 93.11% in coarse and fine-grained classification, respectively.

8. Conclusion

This paper presents a manually annotated novel Bengali aggressive text dataset ('BAD') and empirically validates it. The BAD comprises 14158 texts accumulated from various social media sources and labelled adopting a two-level hierarchical annotation schema. Level-A has two coarse-grained (AG, NoAG), and level-B has four fine-grained (ReAG, PoAG, VeAG, GeAG) classes. Various machine learning (LR, RF, NB, SVM), deep learning (CNN, BiLSTM, CNN+BiLSTM) and transformer (m-BERT, distil-BERT, Bangla-BERT, XLM-R) models are applied on BAD to examine their performance. After analyzing these models' outcomes, this work proposed a weighted ensemble architecture. The proposed technique has the ability to adjust the softmax probabilities of the participating models depending on their previous outcomes on the dataset. This technique outperformed the average ensemble and other baselines by obtaining the maximum weighted f_1-score of 0.9343 in coarse-grained classification. It also achieved the highest weighted f_1-score in fine-grained classes: ReAG (0.95), PoAG (0.97), VeAG(0.92) and GeAG (0.68). Quantitative and qualitative error analysis reveal that it is difficult to identify aggression that expressed implicitly or sarcastically. In future, this work plan to identify mixed aggression by adding more diverse data in fine-grained categories. It will be interesting to investigate how the models perform if we transfer knowledge from resource-rich language's. Other aspects to explore are code-mixing and code-switching of Bengali and English/other languages. Moreover, we also aim to investigate the proposed model's performance with Twitter data with more classes such as racial and geographic aggression.

CRediT authorship contribution statement

Omar Sharif: Conceptualization, Data curation, Methodology design and Implementation, Writing - Original draft, Experiments, Analysis. **Mohammed Moshiul Hoque:** Conceptualization, Methodology, Writing - review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by ICT Innovation Fund, ICT Division, Ministry of Posts, Telecommunications and Information Technology, Bangladesh. This work also supported by the Directorate of Research and Extension (DRE), Chittagong University of Engineering & Technology.

References

[1] D. U. Patton, J. S. Hong, M. Ranney, S. Patel, C. Kelley, R. Eschmann, T. Washington, Social media as a vector for youth violence: A review of the literature, Computers in Human Behavior 35 (2014) 548–553. doi:https://doi.org/10.1016/j.chb.2014.02.043.
URL https://www.sciencedirect.com/science/article/pii/S0747563214001101

[2] R. Bannink, S. Broeren, P. M. van de Looij – Jansen, F. G. de Waart, H. Raat, Cyber and traditional bullying victimization as a risk factor for mental health problems and suicidal ideation in adolescents, PLOS ONE 9 (4) (2014) 1–7. doi:10.1371/journal.pone.0094026.
URL https://doi.org/10.1371/journal.pone.0094026

[3] R. A. Bonanno, S. Hymel, Cyber bullying and internalizing difficulties: Above and beyond the impact of traditional forms of bullying, Journal of youth and adolescence 42 (5) (2013) 685–697.
URL https://doi.org/10.1007/s10964-013-9937-1

[4] Z. Waseem, W. H. K. Chung, D. Hovy, J. Tetreault (Eds.), Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017. doi:10.18653/v1/W17-30.
URL https://www.aclweb.org/anthology/W17-3000

[5] J. Salminen, H. Almerekhi, M. Milenković, S. gyo Jung, J. An, H. Kwak, B. Jansen, Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media, 2018.
URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17885

[6] B. Haddad, Z. Orabe, A. Al-Abood, N. Ghneim, Arabic offensive language detection with attention-based deep neural networks, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, European Language Resource Association, Marseille, France, 2020, pp. 76–81.
URL https://www.aclweb.org/anthology/2020.osact-1.12

[7] M. Ravikiran, A. E. Muljibhai, T. Miyoshi, H. Ozaki, Y. Koreeda, S. Masayuki, Hitachi at semeval-2020 task 12: Offensive language identification with noisy labels using statistical sampling and post-processing (2020). arXiv:2005.00295.

[8] A. Bhattacharjee, T. Hasan, K. Samin, M. S. Rahman, A. Iqbal, R. Shahriyar, Banglabert: Combating embedding barrier for low-resource language understanding (2021). arXiv:2101.00204.

[9] O. Sharif, E. Hossain, M. M. Hoque, 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, Association for Computational Linguistics, Kyiv, 2021, pp. 255–261.
URL https://aclanthology.org/2021.dravidianlangtech-1.35

[10] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1–11.
URL https://www.aclweb.org/anthology/W18-4401

[11] N. Nikhil, R. Pahwa, M. K. Nirala, R. Khilnani, LSTMs with attention for aggression detection, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 52–57.

URL https://www.aclweb.org/anthology/W18-4406

[12] N. Safi Samghabadi, P. Patwa, S. PYKL, P. Mukherjee, A. Das, T. Solorio, Aggression and misogyny detection using BERT: A multi-task approach, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 126–131.
URL https://www.aclweb.org/anthology/2020.trac-1.20

[13] P. Fortuna, J. Soler-Company, L. Wanner, How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?, Information Processing & Management 58 (3) (2021) 102524. doi:https://doi.org/10.1016/j.ipm.2021.102524.
URL https://www.sciencedirect.com/science/article/pii/S0306457321000339

[14] L. Gao, R. Huang, Detecting online hate speech using context aware models, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, INCOMA Ltd., Varna, Bulgaria, 2017, pp. 260–266. doi:10.26615/978-954-452-049-6_036.
URL https://doi.org/10.26615/978-954-452-049-6_036

[15] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Forum for Information Retrieval Evaluation, FIRE 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 29–32. doi:10.1145/3441501.3441517.
URL https://doi.org/10.1145/3441501.3441517

[16] S. T. Roberts, J. Tetreault, V. Prabhakaran, Z. Waseem (Eds.), Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019.
URL https://www.aclweb.org/anthology/W19-3500

[17] E. W. Pamungkas, V. Patti, Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 363–370. doi:10.18653/v1/P19-2051.
URL https://www.aclweb.org/anthology/P19-2051

[18] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, H. Margetts, Challenges and frontiers in abusive content detection, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 80–93. doi:10.18653/v1/W19-3509.
URL https://www.aclweb.org/anthology/W19-3509

[19] A. G. D'Sa, I. Illina, D. Fohr, Towards non-toxic landscapes: Automatic toxic comment detection using DNN, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 21–25.
URL https://www.aclweb.org/anthology/2020.trac-1.4

[20] M. Karan, J. Šnajder, Preemptive toxic language detection in Wikipedia comments using thread-level context, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 129–134. doi:10.18653/v1/W19-3514.
URL https://www.aclweb.org/anthology/W19-3514

[21] S. Bhattacharya, S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, B. Lahiri, A. K. Ojha, Developing a multilingual annotated corpus of misogyny and aggression, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 158–168.
URL https://www.aclweb.org/anthology/2020.trac-1.25

[22] S. Sharifirad, S. Matwin, When a tweet is actually sexist. a more comprehensive classification of different online harassment categories and the challenges in nlp (2019). arXiv:1902.10584.

[23] T. Mihaylov, G. Georgiev, P. Nakov, Finding opinion manipulation trolls in news community forums, in: Proceedings of the Nineteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Beijing, China, 2015, pp. 310–314. doi:10.18653/v1/K15-1032.
URL https://www.aclweb.org/anthology/K15-1032

[24] L. G. Mojica de la Vega, V. Ng, Modeling trolling in social media conversations, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018.
URL https://www.aclweb.org/anthology/L18-1585

[25] M. Dadvar, D. Trieschnigg, F. de Jong, Experts and machines against bullies: A hybrid approach to detect cyberbullies, in: M. Sokolova, P. van Beek (Eds.), Advances in Artificial Intelligence, Springer International Publishing, Cham, 2014, pp. 275–281. doi:https://doi.org/10.1007/978-3-319-06483-3_25.

[26] M. Dadvar, D. Trieschnigg, R. Ordelman, F. de Jong, Improving cyberbullying detection with user context, in: Proceedings of the 35th European Conference on Advances in Information Retrieval, ECIR'13, Springer-Verlag, Berlin, Heidelberg, 2013, p. 693–696. doi:10.1007/978-3-642-36973-5_62.
URL https://doi.org/10.1007/978-3-642-36973-5_62

[27] J. Pavlopoulos, N. Thain, L. Dixon, I. Androutsopoulos, ConvAI at SemEval-2019 task 6: Offensive language identification and categorization with perspective and BERT, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 571–576. doi:10.18653/v1/S19-2102.
URL https://www.aclweb.org/anthology/S19-2102

[28] G. Wiedemann, S. M. Yimam, C. Biemann, UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1638–1644.
URL https://www.aclweb.org/anthology/2020.semeval-1.213

[29] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 task 6: Identify-

ing and categorizing offensive language in social media (OffensEval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 75–86. doi:10.18653/v1/S19-2010.
URL https://www.aclweb.org/anthology/S19-2010

[30] S. T. Aroyehun, A. Gelbukh, 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), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 90–97.
URL https://www.aclweb.org/anthology/W18-4411

[31] J. Risch, R. Krestel, Bagging BERT models for robust aggression identification, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 55–61.
URL https://www.aclweb.org/anthology/2020.trac-1.9

[32] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Evaluating aggression identification in social media, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 1–5.
URL https://www.aclweb.org/anthology/2020.trac-1.1

[33] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1415–1420. doi:10.18653/v1/N19-1144.
URL https://www.aclweb.org/anthology/N19-1144

[34] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of twitter abusive behavior, Proceedings of the International AAAI Conference on Web and Social Media 12 (1) (Jun. 2018).
URL https://ojs.aaai.org/index.php/ICWSM/article/view/14991

[35] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, Proceedings of the International AAAI Conference on Web and Social Media 11 (1) (May 2017).
URL https://ojs.aaai.org/index.php/ICWSM/article/view/14955

[36] P. Mathur, R. Shah, R. Sawhney, D. Mahata, Detecting offensive tweets in Hindi-English code-switched language, in: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 18–26. doi:10.18653/v1/W18-3504.
URL https://www.aclweb.org/anthology/W18-3504

[37] R. Kumar, A. N. Reganti, A. Bhatia, T. Maheshwari, Aggression-annotated corpus of Hindi-English code-mixed data, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018.
URL https://www.aclweb.org/anthology/L18-1226

[38] M. Bhardwaj, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Hostility detection dataset in hindi (2020). arXiv:2011.03588.

[39] H. Mulki, H. Haddad, C. Bechikh Ali, H. Alshabani, L-HSAB: A Levantine Twitter dataset for hate speech and abusive language, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 111–118. doi:10.18653/v1/W19-3512.
URL https://www.aclweb.org/anthology/W19-3512

[40] H. Mubarak, K. Darwish, W. Magdy, Abusive language detection on Arabic social media, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 52–56. doi:10.18653/v1/W17-3008.
URL https://www.aclweb.org/anthology/W17-3008

[41] S. Hassan, Y. Samih, H. Mubarak, A. Abdelali, ALT at SemEval-2020 task 12: Arabic and English offensive language identification in social media, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1891–1897.
URL https://www.aclweb.org/anthology/2020.semeval-1.249

[42] M. Á. Á. Carmona, E. Guzmán-Falcón, M. Montes-y-Gómez, H. J. Escalante, L. V. Pineda, V. Reyes-Meza, A. R. Sulayes, Overview of MEX-A3T at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J. C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, Vol. 2150 of CEUR Workshop Proceedings, CEUR-WS.org, 2018, pp. 74–96.
URL http://ceur-ws.org/Vol-2150/overview-mex-a3t.pdf

[43] M. Graff, S. Miranda-Jiménez, E. S. Tellez, D. Moctezuma, V. Salgado, J. Ortiz-Bejar, C. N. Sánchez, INGEOTEC at MEX-A3T: author profiling and aggressiveness analysis in twitter using μtc and evomsa, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J. C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, Vol. 2150 of CEUR Workshop Proceedings, CEUR-WS.org, 2018, pp. 128–133.
URL http://ceur-ws.org/Vol-2150/MEX-A3T_paper6.pdf

[44] M. Wiegand, Overview of the germeval 2018 shared task on the identification of offensive language, online available: https://epub.oeaw.ac.at/?arp=0x003a10d2 - Last access:11.3.2021 (2018).
URL https://epub.oeaw.ac.at/?arp=0x003a10d2

[45] J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, Overview of germeval task 2, 2019 shared task on the identification of offensive language, Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg, German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg, München [u.a.], 2019, pp. 352 – 363.
URL http://nbn-resolving.de/urn:nbn:de:bsz:mh39-93197

[46] J. A. Leite, D. Silva, K. Bontcheva, C. Scarton, 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, Association for Computational Linguistics, Suzhou, China, 2020, pp. 914–924.
URL https://www.aclweb.org/anthology/2020.aacl-main.91

[47] R. de Pelle, V. Moreira, Offensive comments in the brazilian web: a dataset and baseline results, in: Anais do VI Brazilian Workshop on Social Network Analysis and Mining, SBC, Porto Alegre, RS, Brasil, 2017. doi:10.5753/brasnam.2017.3260.
URL https://sol.sbc.org.br/index.php/brasnam/article/view/3260

[48] P. Fortuna, J. Rocha da Silva, J. Soler-Company, L. Wanner, S. Nunes, A hierarchically-labeled Portuguese hate speech dataset, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 94–104. doi:10.18653/v1/W19-3510.
URL https://www.aclweb.org/anthology/W19-3510

[49] S. Mishra, S. Prasad, S. Mishra, Multilingual joint fine-tuning of transformer models for identifying trolling, aggression and cyberbullying at TRAC 2020, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 120–125.
URL https://www.aclweb.org/anthology/2020.trac-1.19

[50] D. Gordeev, O. Lykova, BERT of all trades, master of some, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 93–98.
URL https://www.aclweb.org/anthology/2020.trac-1.15

[51] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1425–1447.
URL https://www.aclweb.org/anthology/2020.semeval-1.188

[52] S. Wang, J. Liu, X. Ouyang, Y. Sun, Galileo at SemEval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1448–1455.
URL https://www.aclweb.org/anthology/2020.semeval-1.189

[53] H. Ahn, J. Sun, C. Y. Park, J. Seo, NLPDove at SemEval-2020 task 12: Improving offensive language detection with cross-lingual transfer, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1576–1586.
URL https://www.aclweb.org/anthology/2020.semeval-1.206

[54] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018., IberEval@ SEPLN 2150 (2018) 214–228.

[55] D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, J. Wernimont (Eds.), Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, Brussels, Belgium, 2018.
URL https://www.aclweb.org/anthology/W18-5100

[56] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. doi:10.18653/v1/S19-2007.
URL https://www.aclweb.org/anthology/S19-2007

[57] T. Ranasinghe, M. Zampieri, Multilingual offensive language identification with cross-lingual embeddings, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 5838–5844. doi:10.18653/v1/2020.emnlp-main.470.
URL https://www.aclweb.org/anthology/2020.emnlp-main.470

[58] M. R. Karim, S. K. Dey, B. R. Chakravarthi, Deephateexplainer: Explainable hate speech detection in under-resourced bengali language (2021). arXiv:2012.14353.

[59] N. Romim, M. Ahmed, H. Talukder, M. S. Islam, Hate speech detection in the bengali language: A dataset and its baseline evaluation (2020). arXiv:2012.09686.

[60] O. Sharif, M. M. Hoque, Automatic detection of suspicious bangla text using logistic regression, in: P. Vasant, I. Zelinka, G.-W. Weber (Eds.), Intelligent Computing and Optimization, Springer International Publishing, Cham, 2020, pp. 581–590. doi:https://doi.org/10.1007/978-3-030-33585-4_57.

[61] O. Sharif, M. M. Hoque, A. S. M. Kayes, R. Nowrozy, I. H. Sarker, Detecting suspicious texts using machine learning techniques, Applied Sciences 10 (18) (2020). doi:10.3390/app10186527.
URL https://www.mdpi.com/2076-3417/10/18/6527

[62] E. A. Emon, S. Rahman, J. Banarjee, A. K. Das, T. Mittra, A deep learning approach to detect abusive bengali text, in: 2019 7th International Conference on Smart Computing Communications (ICSCC), 2019, pp. 1–5. doi:10.1109/ICSCC.2019.8843606.

[63] P. Chakraborty, M. H. Seddiqui, Threat and abusive language detection on social media in bengali language, in: 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1–6. doi:10.1109/ICASERT.2019.8934609.

[64] O. Sharif, M. M. Hoque, Identification and classification of textual aggression in social media: Resource creation and evaluation, in: T. Chakraborty, et al. (Eds.), Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer Nature Switzerland AG, 2021, pp. 1–12. doi:https://doi.org/10.1007/978-3-030-73696-5_2.

[65] C. A. Anderson, B. J. Bushman, Human aggression, Annual Review of Psychology 53 (1) (2002) 27–51. doi:10.1146/annurev.psych.53.100901.135231.
URL https://doi.org/10.1146/annurev.psych.53.100901.135231

[66] Facebook, Violence and incitement, available online: https://www.facebook.com/communitystandards/ (accessed on 2 October 2020).

[67] M. J. Díaz-Torres, P. A. Morán-Méndez, L. Villasenor-Pineda, M. Montes-y Gómez, J. Aguilera, L. Meneses-Lerín, Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 132–136.
URL https://www.aclweb.org/anthology/2020.trac-1.21

[68] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th International Conference on World Wide Web, WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2016, p. 145–153. doi:10.1145/2872427.2883062.
URL https://doi.org/10.1145/2872427.2883062

[69] Youtube, Harmful or dangerous content policy, available online: https://support.google.com/youtube/answer/2801939/ (accessed on 2 October 2020).

[70] COE, Hate speech and violence, available online: https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence/ (accessed on 3 October 2020).

[71] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Comput. Surv. 51 (4) (Jul. 2018). doi:10.1145/3232676.
URL https://doi.org/10.1145/3232676

[72] A. Roy, P. Kapil, K. Basak, A. Ekbal, An ensemble approach for aggression identification in English and Hindi text, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 66–73.
URL https://www.aclweb.org/anthology/W18-4408

[73] R. A. Baron, D. R. Richardson, Human aggression, Springer Science & Business Media, 2004.

[74] A. H. Buss, The psychology of aggression, Wiley, 1961.

[75] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 78–84. doi:10.18653/v1/W17-3012.
URL https://www.aclweb.org/anthology/W17-3012

[76] R. Kumar, B. Lahiri, A. K. Ojha, Aggressive and offensive language identification in hindi, bangla, and english: A comparative study, SN Computer Science 2 (1) (2021) 1–20. doi:10.1007/s42979-020-00414-6.
URL https://doi.org/10.1007/s42979-020-00414-6

[77] S. Weingartner, L. Stahel, Online aggression from a sociological perspective: An integrative view on determinants and possible countermeasures, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 181–187. doi:10.18653/v1/W19-3520.
URL https://www.aclweb.org/anthology/W19-3520

[78] S. Srivastava, P. Khurana, Detecting aggression and toxicity using a multi dimension capsule network, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 157–162. doi:10.18653/v1/W19-3517.
URL https://www.aclweb.org/anthology/W19-3517

[79] X. Zhou, M. Sap, S. Swayamdipta, N. A. Smith, Y. Choi, Challenges in automated debiasing for toxic language detection (2021). arXiv:2102.00086.

[80] S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, V. P. Plagianakos, Convolutional neural networks for toxic comment classification (2018). arXiv:1802.09957.

[81] P. Fortuna, J. Soler, L. Wanner, Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6786–6794.
URL https://www.aclweb.org/anthology/2020.lrec-1.838

[82] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, p. 759–760. doi:10.1145/3041021.3054223.
URL https://doi.org/10.1145/3041021.3054223

[83] P. Kapil, A. Ekbal, A deep neural network based multi-task learning approach to hate speech detection, Knowledge-Based Systems 210 (2020) 106458. doi:https://doi.org/10.1016/j.knosys.2020.106458.
URL https://www.sciencedirect.com/science/article/pii/S0950705120305876

[84] S. Akiwowo, B. Vidgen, V. Prabhakaran, Z. Waseem (Eds.), Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020.
URL https://www.aclweb.org/anthology/2020.alw-1.0

[85] M. O. Ibrohim, I. Budi, Multi-label hate speech and abusive language detection in Indonesian Twitter, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 46–57. doi:10.18653/v1/W19-3506.
URL https://www.aclweb.org/anthology/W19-3506

[86] N. Safi Samghabadi, A. Hatami, M. Shafaei, S. Kar,

T. Solorio, Attending the emotions to detect online abusive language, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 79–88. doi:10.18653/v1/2020.alw-1.10.
URL https://www.aclweb.org/anthology/2020.alw-1.10

[87] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, V. Hoste, Detection and fine-grained classification of cyberbullying events, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 2015, pp. 672–680.
URL https://www.aclweb.org/anthology/R15-1086

[88] B. Vidgen, L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, PLOS ONE 15 (12) (2021) 1–32. doi:10.1371/journal.pone.0243300.
URL https://doi.org/10.1371/journal.pone.0243300

[89] Z. Waseem, Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter, in: Proceedings of the First Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Austin, Texas, 2016, pp. 138–142. doi:10.18653/v1/W16-5618.
URL https://www.aclweb.org/anthology/W16-5618

[90] E. M. Bender, B. Friedman, Data statements for natural language processing: Toward mitigating system bias and enabling better science, Transactions of the Association for Computational Linguistics 6 (2018) 587–604. doi:10.1162/tacl_a_00041.
URL https://www.aclweb.org/anthology/Q18-1041

[91] R. Binns, M. Veale, M. Van Kleek, N. Shadbolt, Like trainer, like bot? inheritance of bias in algorithmic content moderation, in: G. L. Ciampaglia, A. Mashhadi, T. Yasseri (Eds.), Social Informatics, Springer International Publishing, Cham, 2017, pp. 405–415.

[92] L. Derczynski, K. Bontcheva, I. Roberts, Broad Twitter corpus: A diverse named entity recognition resource, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 1169–1179.
URL https://www.aclweb.org/anthology/C16-1111

[93] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, M. Wojatzki, Measuring the reliability of hate speech annotations: The case of the european refugee crisis, arXiv preprint arXiv:1701.08118 (2017).

[94] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1) (1960) 37–46. arXiv:https://doi.org/10.1177/001316446002000104, doi:10.1177/001316446002000104.
URL https://doi.org/10.1177/001316446002000104

[95] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79. doi:https://doi.org/10.1016/j.neucom.2017.11.077.
URL https://www.sciencedirect.com/science/article/pii/S0925231218302911

[96] T. Tokunaga, I. Makoto, Text categorization based on weighted inverse document frequency, in: Special Interest Groups and Information Process Society of Japan (SIG-IPSJ, Citeseer, 1994.

[97] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages (2018). arXiv:1802.06893.

[98] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651 (2016).

[99] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[100] P. Kapil, A. Ekbal, D. Das, NLP at SemEval-2019 task 6: Detecting offensive language using neural networks, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 587–592. doi:10.18653/v1/S19-2105.
URL https://www.aclweb.org/anthology/S19-2105

[101] O. Sharif, E. Hossain, M. M. Hoque, Combating hostility: Covid-19 fake news and hostile post detection in social media (2021). arXiv:2101.03291.

[102] S. Madisetty, M. Sankar Desarkar, Aggression detection in social media using deep neural networks, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 120–127.
URL https://www.aclweb.org/anthology/W18-4415

[103] D. J. C. MacKay, Hyperparameters: optimize, or integrate out?, in: Maximum Entropy and Bayesian Methods: Santa Barbara, California, U.S.A., 1993, Vol. 62, Springer, Dordrecht, 1996, pp. 43–60. doi:10.1007/978-94-015-8729-7_2.

[104] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F. E. Alsaadi, A survey of deep neural network architectures and their applications, Neurocomputing 234 (2017) 11–26. doi:https://doi.org/10.1016/j.neucom.2016.12.038.
URL https://www.sciencedirect.com/science/article/pii/S0925231216315533

[105] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
URL https://www.aclweb.org/anthology/N19-1423

[106] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2020). arXiv:1910.01108.

[107] S. Sarker, Banglabert: Bengali mask language model for bengali language understading (2020).
URL https://github.com/sagorbrur/bangla-bert

[108] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.
URL https://www.aclweb.org/anthology/2020.

acl-main.747
[109] A. S. Maiya, ktrain: A low-code library for augmented machine learning (2020). arXiv:2004.10703.
[110] V. Bhatnagar, P. Kumar, S. Moghili, P. Bhattacharyya, Divide and conquer: An ensemble approach for hostile post detection in hindi (2021). arXiv:2101.07973.
[111] S. Tawalbeh, M. Hammad, M. AL-Smadi, KEIS@JUST at SemEval-2020 task 12: Identifying multilingual offensive tweets using weighted ensemble and fine-tuned BERT, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 2035–2044.
URL https://www.aclweb.org/anthology/2020.semeval-1.269
[112] S. Gundapu, R. Mamidi, Transformer based automatic covid-19 fake news detection system (2021). arXiv:2101.00180.
[113] S. M. S.-U.-R. Shifath, M. F. Khan, M. S. Islam, A transformer based approach for fighting covid-19 fake news (2021). arXiv:2101.12027.
[114] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. doi:10.18653/v1/N16-2013.
URL https://aclanthology.org/N16-2013
[115] K. Kumari, J. P. Singh, 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, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 113–119.
URL https://www.aclweb.org/anthology/2020.trac-1.18
[116] A. Baruah, K. Das, F. Barbhuiya, K. Dey, Aggression identification in English, Hindi and Bangla text using BERT, RoBERTa and SVM, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 76–82.
URL https://www.aclweb.org/anthology/2020.trac-1.12
[117] H. Nayel, NAYEL at SemEval-2020 task 12: TF/IDF-based approach for automatic offensive language detection in Arabic tweets, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 2086–2089.
URL https://www.aclweb.org/anthology/2020.semeval-1.276