Towards Safer Communities: Detecting Aggression and Offensive Language in Code-Mixed Tweets to Combat Cyberbullying

Nazia Nafis[1], Diptesh Kanojia[1,2], Naveen Saini[1], Rudra Murthy[1]
[1]Indian Institute of Information Technology Lucknow, India.
[2]Surrey Institute for People-Centred AI, University of Surrey, United Kingdom.
[1]{mcs21004, naveen, rudra}@iiitl.ac.in
[2]d.kanojia@surrey.ac.uk

Abstract

Cyberbullying is a serious societal issue widespread on various channels and platforms, particularly social networking sites. Such platforms have proven to be exceptionally fertile grounds for such behavior. The dearth of high-quality training data for multilingual and low-resource scenarios, data that can accurately capture the nuances of social media conversations, often poses a roadblock to this task. This paper attempts to tackle cyberbullying, specifically its two most common manifestations - *aggression* and *offensiveness*. We present a novel, manually annotated dataset of a total of $10,000$ English and Hindi-English code-mixed tweets, manually annotated for aggression detection and offensive language detection tasks[1]. Our annotations are supported by inter-annotator agreement scores of 0.67 and 0.74 for the two tasks, indicating substantial agreement. We perform comprehensive fine-tuning of pre-trained language models (PTLMs) using this dataset to check its efficacy. Our challenging test sets show that the best models achieve macro F1-scores of 67.87 and 65.45 on the two tasks, respectively. Further, we perform cross-dataset transfer learning to benchmark our dataset against existing aggression and offensive language datasets. We also present a detailed quantitative and qualitative analysis of errors in prediction, and with this paper, we publicly release the novel dataset, code, and models.

1 Introduction

Social media is a group of Internet-based applications that allows the creation and exchange of user-generated content. Lately, it has risen as one of the most popular ways in which people share opinions with each other (Pelicon et al., 2019). With rapid advances in Web 3.0, social media is expected to evolve and emerge as an even more vital and potent means of communication. Simultaneously, there has also been noticed a sharp uptick in bullying behavior - including but not limited to the use of snide remarks, abusive words, and personal attacks, going as far as rape threats (Hardaker and McGlashan, 2016) on such platforms. In this context, by leveraging the technological advancements in machine learning and natural language processing, automatic detection of instances of cyberbullying on social media platforms such as Twitter can help create a safer environment. Here we investigate two forms of cyberbullying - aggression and offensiveness.

OAG	You wont march against kids being raped in country or the endless stream of migrants maybe cheaper energy no but you would march against @username pathetic fking pathetic!
CAG	Wait a few days sir, you are getting used to hearing harsh words. In future, when all the banks will be of Adani or Ambani, then you will also have to listen to abuses for withdrawing your deposits. #demonetisation
NAG	Well this is pure goosebumps, whenever I see him I feel so proud that he is our PM, a living legend for sure

Table 1: Examples of overtly aggressive (**OAG**), covertly aggressive (**CAG**), and non-aggressive (**NAG**) tweets from our dataset.

Aggression has been defined as any behavior enacted with the intention of harming another person who is motivated to avoid that harm (Anderson et al., 2002; Bushman and Huesmann, 2014). Several studies have noted the proliferation of abusive language and an increase in aggressive content on social media (Mantilla, 2013; Suzor et al., 2019)

[1] https://github.com/surrey-nlp/woah-aggression-detection/blob/main/data/New10kData/cyberbullying_10k.csv

The 7th Workshop on Online Abuse and Harms (WOAH), pages 29–41
July 13, 2023 ©2023 Association for Computational Linguistics

OFF	Bhikaris like u, can u first afford watching movie in threatres? Talk about that first. Just coz internet is cheap does not mean u will do open defecation in social media. MDRCHD bhaag BSDK
NOT	Who doesn't enjoy the daily press briefings? They really ease the tension! We have to find some way to keep ourselves entertained

Table 2: Examples of offensive (**OFF**) and non-offensive (**NOT**) tweets from our dataset.

On the other hand, offensiveness has been described as any word or string of words which has or can have a negative impact on the sense of self or well-being of those who encounter it (Molek-Kozakowska, 2022) – that is, it makes or can make them feel mildly or extremely discomfited, insulted, hurt or frightened.

Motivation The dearth of manually-annotated datasets for the tasks of aggression detection and offensive language detection, especially in the Hindi-English code-mixed setting, necessitated us to work in this area.

This paper investigates the tasks of aggression detection and offensive language detection on Twitter data. We curate politically-themed tweets and perform manual annotation to create a dataset for the tasks. Our annotation schema is in line with the existing aggressive and offensive language detection datasets. With the help of pre-trained language models, we fine-tune pre-trained language models for both tasks and discuss the obtained results regarding precision, recall, and macro F1-scores. The **key contributions** of this work are:

- Introduction of a novel, manually-annotated dataset containing English and Hindi-English code-mixed tweets to model aggression and offensiveness in text.

- Validation of our dataset's efficacy for aggressive and offensive language detection tasks within two subsets of this data, *viz.,* English and Hindi-English code-mixed.

- Cross-validation of dataset efficacy with the help of zero-shot transfer learning-based experiments on existing datasets.

- Quantitative and qualitative analysis of erroneous predictions.

2 Related Work

Our work deals with two different but correlated classification tasks. In the available literature, both have been investigated with the help of various machine learning and deep learning-based methods. Below, we provide a detailed overview of the literature from both tasks in separate subsections.

2.1 Aggression Detection

We model aggression detection as a multi-class classification task where our schema is defined as proposed in the TRAC dataset (Kumar et al., 2018a). However, the earliest approaches used decision trees (Spertus, 1997a) to detect aggression with the help of manual rules. These rules were based on syntactic and semantic features. Later, the focus shifted to feature engineering from text which included features like Bag-of-Words (BOW) (Kwok and Wang, 2013; Liu et al., 2019a), N-grams at the word level (Pérez and Luque, 2019; Liu and Forss, 2014; Watanabe et al., 2018), N-grams at the character level (Gambäck and Sikdar, 2017; Pérez and Luque, 2019), typed dependencies (Burnap and Williams, 2016b), part-of-speech tags (Davidson et al., 2017b), dictionary-based approaches (Molek-Kozakowska, 2022) and lexicons (Burnap and Williams, 2016b; Alorainy et al., 2019).

Word embedding-based approaches for automated extraction of these features from text further improved the detection of aggressive text (Nobata et al., 2016; Zhang et al., 2018; Badjatiya et al., 2017; Kshirsagar et al., 2018; Orăsan, 2018; Pratiwi et al., 2019; Galery et al., 2018). Various deep learning-based architectures proposed in the literature use word embeddings to encode features in the text (Nina-Alcocer, 2019; Ribeiro and Silva, 2019). Authors have proposed the use of Convolutional Neural Networks (Gambäck and Sikdar, 2017; Roy et al., 2018; Huang et al., 2018), Long-Short Term Memory (LSTM) (Badjatiya et al., 2017; Pitsilis et al., 2018; Nikhil et al., 2018), Gated Recurrent Unit (GRU) (Zhang et al., 2018; Galery et al., 2018), or a combination of different Deep Neural Network architectures in an ensemble setting (Madisetty and Sankar Desarkar, 2018).

However, state-of-the-art performance (Bojkovský and Pikuliak, 2019; Ramiandrisoa and

Mothe, 2020; Mozafari et al., 2019) was achieved with the help of pre-trained language models (PTLMs) with encoders like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). Further, we also observe the use of these contextual embeddings in SemEval-2019 Task 6 (Zampieri et al., 2019b) for English tweets, and TRAC (Kumar et al., 2018b) for Hindi and English tweets and Facebook comments; further motivating us to explore the use of multiple pre-trained language models to validate the efficacy of our dataset.

2.2 Offensive Language Identification

We model the offensive language identification task as a binary classification problem. Waseem et al. (2017) proposed a typology for abusive language and synthesize the typology with two-fold siders containing whether the abuse is 'generalized' or 'directed' *vs.* when it is 'explicit' or 'implicit'. They discuss the distinction between explicit and implicit in the context of 'denotation' *vs.* 'connotation' as discussed by (Barthes, 1957). A detailed review of hate speech detection (Schmidt and Wiegand, 2017) task has surveyed various approaches deployed in the past. Spertus (1997b) propose a rule-based framework for identifying hostile messages where they use manually constructed rules to identify profanity, condescension, insults and so on. Razavi et al. (2010) utilize a flame annotated corpus which contains a lexicon of hostile and abusive words to detect offensive language in personal and commercial communication. Dictionaries (Liu and Forss, 2015) and bag-of-words (Burnap and Williams, 2016a) have also been proposed as lexical features to detect offensive language.

The use of machine learning algorithms to detect offensive language has been prevalent in the research community (Davidson et al., 2017a; Waseem and Hovy, 2016). Further, the use of word embeddings learned with the help of word2vec or FastText approaches combined with machine/deep learning improved the performance of offensive language identification by a significant margin (Rakib and Soon, 2018; Herwanto et al., 2019; Badri et al., 2022). However, as we point out in the previous subsection, state-of-the-art performance has been achieved with the help of PTLMs.

Pitenis et al. (2020) perform the task specifically for the low-resource Greek language. Similarly, Ranasinghe and Zampieri (2020) show that the use of cross-lingual embeddings for inter-task and inter-language scenarios is beneficial. The authors first train a multilingual PTLM (XLM-R) on the English data, and then further continue the training using saved weights and *softmax* layer, for other languages *viz.* Hindi, Bengali, and Spanish.

Further, there have been a lot of efforts to create datasets for the detection of offensive language and hate speech[2] on social media. Çöltekin (2020) presents a dataset for the Turkish language with a specified target for offense. Díaz-Torres et al. (2020) build the same for Mexican Spanish. A clear majority of studies deal with the English language. *While other Indian language datasets have been proposed, there is a clear dearth of English-Hindi datasets which also address code-mixing*, in the available literature (Chakravarthi et al., 2021, 2022) except a few (Mathur et al., 2018; Saroj and Pal, 2020).

3 Dataset Creation

We create a dataset containing a mix of English and Hindi-English sentences, to ensure that sufficient data is available for our research. We used the official Twitter API to obtain data from Twitter. Initially, we collected 15,000 tweets based on the search results for one of the 52 keywords (listed in Table 10 in Appendix) in our list pertaining to recent political events and popular political personalities.

We filtered out tweets that were in any language other than English or Hindi (or containing a mix of both) using XLM Roberta-base with a classification head on top (Conneau et al., 2020). Next, with the help of HingBERT-LID code-mixed language identification model (Nayak and Joshi, 2022), we created subsets of tweets belonging to one of the two aforementioned categories.

We preprocessed the tweets by masking all usernames to minimize the introduction of bias to the annotators. Finally, after cleaning, we were left with 5,452 English monolingual and 4,548 Hindi-English code-mixed tweets.

3.1 Annotation Setup

The following **guidelines** were supplied to the annotators, which outline the definition and provide a few sample tweets for each Aggression and Offensive Language label.

Task I: Aggression Detection

[2] hatespeechdata.com - a catalog of hate speech datasets

Aggression focuses on the user's intention to be aggressive and harmful, or to incite, in various forms, violent acts against a target. The aggression level in the text is categorized into three classes:

- **Overtly Aggressive (OAG):** This type of aggression shows a direct verbal attack pointing to a particular individual or group.

  For example, in the sample tweet for OAG in Table 1, the person expresses frustration over issues such as child sexual abuse, immigration, and high gas prices while also condemning the apathy of others towards these issues. The aggression here is overt, as also seen by the use of words *"fking"* and *"pathetic"* in the tweet.

- **Covertly Aggressive (CAG):** In this type of aggression, the attack is not direct but hidden, subtle, and more indirect while being stated politely in most cases.

  For example, in the sample tweet for CAG in Table 1, the person harbors angst against the process of demonetization of the Indian currency and privatization of banks, but chooses to display it covertly while conversing over Twitter.

- **Not Aggressive (NAG):** Generally, these types of text lack any kind of aggression. It is used to state facts, express greetings and good wishes occasionally, and show agreeableness and support.

  For example, in the sample tweet for NAG in Table 1, the person does not display any aggression at all - on the contrary, they praise the PM by calling them a *"living legend"*.

Task II: Offensive Language Detection

Offensiveness focuses on the potentially hurtful effect of the tweet content on a given target. Text can be identified as belonging to either of the two offensiveness classes:

- **Offensive (OFF)** This category of text often contains offensive words such as sarcastic remarks, insults, slanders, and slurs.

  For example, in the sample tweet for OFF in Table 2, the person uses words such as *"bhikaris"* ("beggars") for others, while also availing outright derogatory Hindi slang to address them.

- **Not Offensive (NOT)** In this category, there is either a thorough use of positive and uplifting language, such as salutations or homage, or a neutral tone.

  For example, in the sample tweet for NOT in Table 2, the person makes a remark about how everybody enjoys the daily press briefings, and how they ease tension and keep everybody entertained. There is no offensive tone in this instance.

	Aggression			**Offensiveness**	
	OAG	**CAG**	**NAG**	**OFF**	**NOT**
Monolingual	1134	1715	2599	1323	4125
Code-mixed	1150	1322	2080	1749	2803
Combined	2284	3037	4679	3072	6928

Table 3: Aggression and Offensive language statistics of our dataset.

Setup Our team of annotators consisted of two undergraduate students fluent in both Hindi (native) and English as their second language. The selection of annotators was objective and unbiased. The aforementioned guidelines were made available to them, to refer to while deciding upon the labels for the tweets. This was done to ensure that their political beliefs/loyalties do not play a role in the annotation process. We also recorded their highest level of education and medium of schooling to ensure that the annotations would be of the desired quality, and we informed them about the collection of this data.

All usernames in the data were masked, so at any given point, only the tweet content was visible to the annotators whereas the target personality/organization was hidden from their purview. To ensure the confidentiality of data and to check biases, any metadata too, such as the tweet senders' demographic identity, was not made available to the annotators.

Moreover, since the tweets often contained aggressive and highly abusive language, the annotators were also given a choice to quit whenever they felt uncomfortable with the task.

3.2 Inter Annotator Agreement

While labeling, each annotator had to decide independently which category the comment belonged to, with the help of a set of guidelines. It can be inferred that all the annotators clearly understood the guidelines for annotation, as in most cases, they arrived at the same annotation freely. To quantify how good the annotation decisions were, we calculated **Cohen's Kappa** score to measure the inter-annotator agreement. It may be noted that a high score on this statistical metric does not mean the annotations are accurate. It only shows the homogeneity of agreement among the annotators about the chosen label.

We obtained an agreement score of 0.67 for Task I, and a score of 0.74 for Task II, both of which indicate *"substantial agreement"* (p >0.05). In case of disagreement on any instance, we obtained a label on such instances with the help of a third annotator.

3.3 Dataset Statistics

Table 3 shows the exploratory statistics on our dataset for aggression and offensiveness, respectively. We have a total of $10,000$ data instances in the form of tweets. Out of this, $2,284$ are overtly aggressive (OAG), $3,073$ are covertly aggressive (CAG), and $4,679$ are not aggressive (NAG). Similarly, there are $3,072$ offensive (OFF) and $6,928$ not offensive (NOT) instances in the dataset.

Additionally, the monolingual vs. code-mixed statistics are also mentioned for each class in both tables. We have $1,134$ monolingual and $1,150$ code mixed tweets in the OAG category, $1,715$ monolingual and $1,322$ code mixed tweets in the CAG category, and $2,599$ monolingual and $2,080$ code mixed tweets in the NAG category. Similarly, there are $1,323$ monolingual and $1,749$ code mixed tweets in the OFF category and $4,125$ monolingual and $2,803$ code mixed tweets in NOT.

4 Approach

In recent times, sequence classification via fine-tuning of pre-trained language models has become a standard approach for performing various NLP tasks. We take a similar approach and fine-tune some pre-trained language models for the two tasks, and report the results in the next section. We work with two general-purpose English models, one multilingual model, one model trained specifically on Hindi-English code-mixed data, and one trained exclusively on Twitter data.

Every tweet containing a sequence of words is tokenized into a sequence of sub-words using the model-specific tokenizer. This sequence of subword tokens is the input to the model that passes through the Transformer's encoder layers. An encoder representation for each token in the sequence is the output from the transformer. We take the encoder representation of the [CLS] token in the case of BERT or the last encoder hidden states for other models. The output layer is a linear layer followed by the softmax function, which takes in the above representation. The model is trained by optimizing for a custom weighted cross-entropy loss value which we explain in detail in an upcoming subsection.

Experimental Setup We fine-tune for the aforementioned two tasks the following pre-trained language models: BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b) which are trained on English data, XLM-R (Conneau et al., 2020) base which is trained over multilingual data containing both Hindi and English, HingRoBERTa (Nayak and Joshi, 2022), a multilingual language model specifically built for Hindi-English code-mixed language as seen in the Indian context, and Bernice (DeLucia et al., 2022), a multilingual language model trained exclusively on Twitter data.

Data Split and Evaluation Criteria We report macro F1-scores on our complete dataset, as well as on its code-mixed and non-code-mixed subsets individually. For the train/validation/test splits, we choose uniform 80% / 10% / 10% from each dataset to perform the experiments.

Experiment Settings We perform experiments using the Huggingface Transformers library (Wolf et al., 2020). We monitor the validation set's macro F1-scores to find the best hyperparameter values, using the following range of values for selecting the best hyperparameter:

- **Batch Size**: 8, 16, 32
- **Learning Rate**: 1e-5, 1e-6, 3e-5, 3e-6, 5e-5, 5e-6

We repeat each training five times with different random seeds and report the mean macro F1-scores along with their standard deviation. Our experiments were performed using 2 x Nvidia RTX A5000 and a single training run usually takes approximately 1 hour on the dataset. For the subsets, however, the runtime is approximately 30 minutes. The models generated during our experiments see the number of trainable parameters varying from 100M to 200M depending upon the language model used.

	Aggression Detection			Offensive Language Detection		
PTLM	**Monolingual**	**Code mixed**	**Combined**	**Monolingual**	**Code mixed**	**Combined**
BERT$_{base}$	63.58±0.51	65.22±0.77	64.98±0.28	60.99±0.43	61.94±0.14	62.05±0.25
RoBERTa$_{base}$	**66.63±0.12**	65.42±0.61	62.13±0.89	**63.46±0.75**	62.06±0.48	60.21±0.30
XLM-R$_{base}$	65.49±0.73	66.85±0.22	**67.87±0.05**	61.24±0.31	64.42±0.02	65.41±0.73
HingRoBERTa	64.01±0.53	**66.94±0.53**	66.47±0.53	61.92±0.26	**64.97±0.13**	**65.45±0.21**
Bernice	63.49±0.15	61.13±0.43	62.75±0.82	60.88±0.57	59.01±0.38	60.58±0.16

Table 4: Mean macro F1-scores obtained from pre-trained language models on our dataset and its two subsets - English monolingual and Hindi-English code-mixed. The values in **bold** highlight the best-performing language model on each dataset.

	Aggression Detection		Offensive Language	
	D1–>D2	**D2–>D1**	**D1–>D2**	**D2–>D1**
BERT$_{base}$	55.63±0.21	52.98±0.56	48.69±0.11	46.49±0.53
RoBERTa$_{base}$	52.13±0.74	50.99±0.47	46.02±0.31	43.64±0.49
XLM-R$_{base}$	**56.81±0.84**	55.33±0.60	50.94±0.55	49.27±0.75
HingRoBERTa	56.29±0.71	54.04±0.10	**51.51±0.28**	49.01±0.24
Bernice	52.05±0.87	49.65±0.57	46.16±0.18	45.88±0.05

Table 5: Cross-dataset Test Set F1-Scores from various language models. **D1** represents our dataset. For Aggression detection, **D2** is the TRAC dataset, whereas for Offensive language detection, **D2** is the OLID dataset.

Custom Weighted Loss As our dataset exhibits class imbalance, we use weighted cross-entropy loss (Lee and Liu, 2003) in all our experiments. We assign a weight to the loss of every instance depending on the class label. Then, we find the percentage of examples by class belonging to each class from the train split and take the inverse of the probability values as the weight for the particular class. In this way, we give more importance to the instances belonging to the minority class.

5 Results

We report the results obtained via fine-tuning pre-trained language models in this section. Table 4 reports the test set macro F1-scores from pre-trained language models for the two tasks of aggression detection and offensive language detection on our dataset. In addition to this, we also present the scores on English monolingual and Hindi-English code-mixed subsets of our dataset.

For aggression, we observe that XLM-R$_{base}$ outperforms other pre-trained language models on our overall dataset, achieving the highest macro F1-score of 67.87. On the English subset, we observe that RoBERTa$_{base}$ performs better than other models with a macro F1-score of 66.63, whereas for the Hindi-English code-mixed subset, Hing-RoBERTa gives the best macro F1-score of 66.94.

For offensive language detection, we observe that Hing-RoBERTa outperforms other pre-trained language models on our overall dataset, achieving the highest macro F1-score of 65.45. On the English subset, we observe that RoBERTa$_{base}$ outperforms other models with a macro F1-score of 63.46. For the Hindi-English code-mixed subset, Hing-RoBERTa once again gives the best performance with a macro F1-score of 64.97.

Cross-dataset Transfer Learning We perform transfer learning experiments to benchmark our dataset against some existing datasets for the same tasks. Results from our transfer learning setup are presented in Table 5.

For the task of aggression detection, we benchmark our dataset against a curated subset of the TRAC (Trolling, Aggression, and Cyberbullying) dataset (Bhattacharya et al., 2020). This subset, (discussed in Table 8 in section 9), contains instances in Hindi (Roman script) and English, and is annotated for aggression (OAG: overtly aggressive, CAG: covertly aggressive, NAG: not aggressive). For the task of Offensive language detection, we use OLID (Offensive Language Identification Dataset) (Zampieri et al., 2019a) to benchmark our dataset. OLID is an English language dataset and we make use of its Level-A labels (OFF: offensive, NOT: not offensive), discussed in Table 9 in section 9. We chose these two datasets because their annotation schema aligned with that of ours, for aggression detection and offensive language detection tasks respectively.

For Aggression detection, the columns D1–>D2 and D2–>D1 in Table 5 present a cross-dataset setup within which we observe the performance of models fine-tuned on D1 (our dataset) and tested on D2 (TRAC dataset), and vice versa. We observe

Tweets | Task: Aggression Detection	**GT**	**M1**	**M2**	**M3**	**Error Cause**
Romanticizing open defecation under heavy rain to enjoy the melancholy	**CAG**	NAG	CAG	NAG	**Sarcasm**
@username The way Rahul Gandhi changed his DP to Nehru holding a tricolour, I want to change it to Savarkar or Golwalkar holding the Flag. Can anyone help and share pictures of theirs holding the tricolor..	**CAG**	NAG	NAG	NAG	**Real-world context**
@username You use words like waqf, muslims, mullahs, terrorists, radicals and you will get a block message from twitter, hope Elon buys twitter very soon.	**CAG**	CAG	NAG	NAG	**Hidden Aggression**

Table 6: Prediction on test set instances from resultant models for aggression detection. **GT**: Ground Truth label, **M1**: XLM-R$_{base}$, **M2**: RoBERTa$_{base}$, **M3**: Hing-RoBERTa.

that models trained on our dataset obtain better F1-scores than those trained on the TRAC dataset. Further, we observe that the best performance achieved in this setup is with the help of XLM-R$_{base}$, the same multilingual model which also performs the best on the combined dataset in Table 4.

For Offensive language detection, we examine the results in columns D1–>D2 and D2–>D1 in Table 5, which note the performance of models fine-tuned on D1 (our dataset) and tested on D2 (OLID dataset), and vice versa. As was true with the first task, it is observed here too that models trained on our dataset obtain better F1-scores as compared to the models trained on the OLID dataset. Additionally, we observe a similar correlation between both sets of results. The model fine-tuned on code-mixed data, Hing-RoBERTa, performs the best in this scenario, as was the case with the combined dataset performance in Table 4.

The overall decrease in F1-scores observed across models for the Offensive language detection task can be attributed to the dissimilarities in the composition of the OLID dataset and our dataset, despite both being annotated for the offensive language identification task with the same annotation schema. While our dataset contains English and Hindi-English tweets pertaining specifically to the Indian political scenario, OLID is an English-language dataset with no instances of Hindi-English code-mixing, and little to no emphasis on regional or national politics.

On the contrary, the TRAC dataset contains English and Hindi-English sentences with a clear focus on the conversational data generated within India, which explains why we see a greater harmony in Table 5 between the TRAC dataset and our dataset, as compared to the OLID dataset.

Error Analysis

For error analysis, we pick the best-performing models for monolingual, code-mixed, and combined datasets, which as per our experiments have been RoBERTa$_{base}$, Hing-RoBERTa, and XLM-R$_{base}$ respectively. We report some of the most common error patterns in Table 6 and Table 7.

For the task of aggression detection, instances carrying sarcasm that make heavy use of oxymoronic/ironic language were misclassified the most by all three models. An example of this is the first tweet in Table 6, where the person who made the tweet observes discontent with the practice of open defecation not by attacking it directly but with sarcasm. Another common error we observed was among instances, that seemed to have a neutral tone ostensibly but required some real-world knowledge to understand the context of aggression within. The second tweet in Table 6 is an excellent example of this. By itself, the tweet does not appear to be aggressive, but its true meaning unveils when read along with context. A few wrongful predictions can also be observed because of the aggression being very covert or hidden, as seen in the third tweet in Table 6 where under the garb of advocating for Elon Musk's free speech, the person is expressing an intent to, in fact, be disrespectful and use words on the platform that spread disharmony.

For the task of offensive language detection, the most common error type was observed due to the presence of offensive named entities. The first tweet in Table 7 is an example of this, where the use of *"Khujliwal"* (a pun on *"Kejriwal"* - which is the name of an Indian politician), is the cause of offense, as labeled by our annotators. Another common error was in instances that required real-world knowledge to understand their full context. For

Tweets | Task: Offensive Language Detection	GT	M1	M2	M3	Error Cause
@username It was always very clear that **Khujliwal** is a Godse Lover	**OFF**	NOT	NOT	NOT	**Named entity**
@username @username Dogs are at least loyal bro ..not these **rice bags**	**OFF**	NOT	NOT	NOT	**Real-world context**
@username @username Nah, you need to do Ghar Wapasi to find real Moksha. Else you will remain a **mlechha**	**OFF**	NOT	NOT	OFF	**Code-mixed**

Table 7: Prediction on test set instances from resultant models for offensive language detection. **GT**: Ground Truth label, **M1**: XLM-R$_{base}$, **M2**: RoBERTa$_{base}$, **M3**: Hing-RoBERTa.

example, in the second tweet in Table 7, *"rice bags"* is actually a derogatory slur used quite commonly in the Indian political context. Finally, we also observe misclassified instances due to the code-mixed nature of tweets, as seen in the third tweet in Table 7 where the word *"mlechha"* has derogatory connotations.

6 Conclusion and Future Work

This paper presents a novel dataset to model aggressiveness and offensiveness in text. We analyze this dataset using approaches such as fine-tuning pre-trained language models for the task of aggression detection and offensive language detection and report the results. Our analysis also takes into account the code-mixing phenomenon observed on social media platforms as we report additional results for this task. Since aggression and offense can be subtle, and their identification in the text can sometimes be subjective, it is important to note the limitations of such a study - which we discuss in the next section. We release any data (including any raw data, but only in the form of tweet IDs and their respective labels for the two tasks), code, and models produced during this study publicly for further research by the community. We license this release under CC-BY-SA 4.0.

In the near future, we aim to annotate this data for tasks such as sarcasm detection - to develop a deeper understanding of how it is related to aggression and offensiveness. Additionally, the motivation for collecting the same data instances marked with aggression and offense labels is for a multi-task learning-based model also to be able to identify when 1) the tone of a text is aggressive without being offensive *vs.,* 2) the text is offensive, despite it not being overtly aggressive. We also aim to collect more data and annotate it using weak supervision. Finally, we also aim to expand on the theoretical underpinnings of sublime aggression and offense by attempting to identify these within other more tangential domains, *viz.,* comedy.

7 Limitations

Our work can be considered to have the following limitations:

1. The dataset we introduce contains $10,000$ text instances sampled from a single social media platform. However, we acknowledge this limitation and as noted in section 6, we aim to extend this work by collecting more political data across various social media platforms and using it to model aggressive behavior.

2. We obtained this dataset by crawling for tweets based on 52 keywords (as shown in Table 10). We acknowledge that these keywords may have limited the domains in which political aggression can occur. That being said, we also hope that task generalizability is not compromised due to the presence of pretrained language models at the helm of our experiments.

8 Ethics Statement

Our dataset of tweets was obtained by scraping Twitter. All tweets have been anonymized, and metadata such as senders' demographic identity is never included in the data used to train our models. We plan to release only the tweet ids and their respective labels for the two tasks as part of our dataset.

References

Wafa Aloraing, Pete Burnap, Han Liu, and Matthew L. Williams. 2019. "the enemy among us": Detecting cyber hate speech with threats-based othering language embeddings. *ACM Trans. Web*, 13(3).

Craig A Anderson, Brad J Bushman, et al. 2002. Human aggression. *Annual review of psychology*, 53(1):27–51.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. ACM Press.

Nabil Badri, Ferihane Kboubi, and Anja Habacha Chaibi. 2022. Combining fasttext and glove word embedding for offensive and hate speech text detection. *Procedia Computer Science*, 207:769–778. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 26th International Conference KES2022.

Roland Barthes. 1957. Mythologies, le seuil. *Points, Paris*.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

Michal Bojkovský and Matúš Pikuliak. 2019. STUFIIT at SemEval-2019 task 5: Multilingual hate speech detection on Twitter with MUSE and ELMo embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 464–468, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Pete Burnap and Matthew L Williams. 2016a. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15.

Peter Burnap and Matthew Leighton Williams. 2016b. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *Epj Data Science*, 5.

Brad J Bushman and L Rowell Huesmann. 2014. Twenty-five years of research on violence in digital games and aggression revisited: A reply to elson and ferguson (2013).

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Parameswari Krishnamurthy, Elizabeth Sherly, and Sinnathamby Mahesan, editors. 2022. *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Dublin, Ireland.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017a. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017b. Automated hate speech detection and the problem of offensive language. In *ICWSM*, pages 512–515.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

María José Díaz-Torres, Paulina Alejandra Morán-Méndez, Luis Villasenor-Pineda, Manuel Montes-y Gómez, Juan Aguilera, and Luis Meneses-Lerín. 2020. 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*, pages 132–136, Marseille, France. European Language Resources Association (ELRA).

Thiago Galery, Efstathios Charitos, and Ye Tian. 2018. Aggression identification and multi lingual word embeddings. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 74–79, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Claire Hardaker and Mark McGlashan. 2016. “real men don’t hate women”: Twitter rape threats and group identity. *Journal of Pragmatics*, 91:80–93.

Guntur Herwanto, Annisa Ningtyas, Kurniawan Nugraha, and I Nyoman Prayana Trisna. 2019. Hate speech and abusive language classification using fast-text. pages 69–72.

Qianjia Huang, Diana Inkpen, Jianhong Zhang, and David Van Bruwaene. 2018. Cyberbullying intervention based on convolutional neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 42–51, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32, Brussels, Belgium. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 1–11.

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

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI’13, page 1621–1622. AAAI Press.

Wee Sun Lee and Bing Liu. 2003. Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML’03, page 448–455. AAAI Press.

Han Liu, Pete Burnap, Wafa Alorainy, and Matthew L. Williams. 2019a. A fuzzy approach to text classification with two-stage training for ambiguous instances. *IEEE Transactions on Computational Social Systems*, 6(2):227–240.

Shuhua Liu and Thomas Forss. 2014. Combining n-gram based similarity analysis with sentiment analysis in web content classification. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval - Volume 1: SSTM, (IC3K 2014)*, pages 530–537. INSTICC, SciTePress.

Shuhua Liu and Thomas Forss. 2015. New classification models for detecting hate and violence web content. In *2015 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K)*, volume 1, pages 487–495. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Sreekanth Madisetty and Maunendra Sankar Desarkar. 2018. Aggression detection in social media using deep neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 120–127, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Karla Mantilla. 2013. Gendertrolling: Misogyny adapts to new media. *Feminist studies*, 39(2):563–570.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Katarzyna Molek-Kozakowska. 2022. Recenzja/review: Jim o’driscoll (2020). offensive language: Taboo, offence and social control. london: Bloomsbury academic. isbn 9781350169678. *Res Rhetorica*, 9:166–169.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

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

Victor Nina-Alcocer. 2019. HATERecognizer at SemEval-2019 task 5: Using features and neural networks to face hate recognition. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 409–415, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Constantin Orăsan. 2018. Aggressive language identification using word embeddings and sentiment features. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 113–119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andraž Pelicon, Matej Martinc, and Petra Kralj Novak. 2019. Embeddia at semeval-2019 task 6: Detecting hate with neural network and transfer learning approaches. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 604–610.

Juan Manuel Pérez and Franco M. Luque. 2019. Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 64–69, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48:4730–4742.

Nur Indah Pratiwi, Indra Budi, and Ika Alfina. 2019. Hate speech detection on indonesian instagram comments using fasttext approach. In *2018 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018*, 2018 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018, pages 447–450, United States. Institute of Electrical and Electronics Engineers Inc. Publisher Copyright: © 2018 IEEE.; 10th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018 ; Conference date: 27-10-2018 Through 28-10-2018.

Tazeek Bin Abdur Rakib and Lay-Ki Soon. 2018. Using the reddit corpus for cyberbully detection. In *Asian Conference on Intelligent Information and Database Systems*.

Faneva Ramiandrisoa and Josiane Mothe. 2020. Aggression identification in social media: a transfer learning based approach. In *TRAC*.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence: 23rd Canadian Conference on Artificial Intelligence, Canadian AI 2010, Ottawa, Canada, May 31–June 2, 2010. Proceedings 23*, pages 16–27. Springer.

Alison Ribeiro and Nádia Silva. 2019. INF-HatEval at SemEval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 420–425, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Arjun Roy, Prashant Kapil, Kingshuk Basak, and Asif Ekbal. 2018. An ensemble approach for aggression identification in English and Hindi text. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 66–73, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Anita Saroj and Sukomal Pal. 2020. An Indian language social media collection for hate and offensive speech. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 2–8, Marseille, France. European Language Resources Association (ELRA).

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Ellen Spertus. 1997a. Smokey: Automatic recognition of hostile messages. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, page 1058–1065. AAAI Press.

Ellen Spertus. 1997b. Smokey: Automatic recognition of hostile messages. In *Aaai/iaai*, pages 1058–1065.

Nicolas Suzor, Molly Dragiewicz, Bridget Harris, Rosalie Gillett, Jean Burgess, and Tess Van Geelen. 2019. Human rights by design: The responsibilities of social media platforms to address gender-based violence online. *Policy & Internet*, 11(1):84–103.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

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

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835. Publisher Copyright: © 2018 IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Z Zhang, D Robinson, and J Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In A Gangemi, R Navigli, M-E Vidal, P Hitzler, R Troncy, L Hollink, A Tordai, and M Alam, editors, *The Semantic Web: Proceedings of the 15th European Semantic Web Conference (ESWC 2018), Heraklion, Crete, Greece, 3-7 June 2018*, volume 10843 of *Lecture notes in computer science*, pages 745–760. Springer, Cham, Switzerland.

9 Appendix

We note the language-wise class distribution for aggression and offensiveness classes, in the publicly available TRAC and OLID datasets respectively, in Table 8 and Table 9. Next, we list the keywords used for data scraping, during the creation of our novel dataset.

	OAG	CAG	NAG
Monolingual	1114	1693	1820
Code mixed	1058	1581	2529
Combined	2172	3279	4349

Table 8: Aggression statistics of the TRAC dataset.

TRAC Dataset Statistics Table 8 shows the exploratory statistics on TRAC dataset for aggression. There are a total of $9,800$ data instances. Out of this, $2,172$ are overtly aggressive (OAG), $3,279$ are covertly aggressive (CAG), and $4,349$ are not aggressive (NAG).

Additionally, the monolingual vs. code mixed statistics are also mentioned for each class. We have $1,114$ monolingual and $1,058$ code mixed tweets in the OAG category, $1,693$ monolingual and $1,581$ code mixed tweets in the CAG category, and $1,820$ monolingual and $2,529$ code mixed tweets in the NAG category.

	OFF	NOT
Monolingual	2034	3578
Code mixed	1134	2786
Combined	3168	6364

Table 9: Offensive language statistics of the OLID dataset.

OLID Dataset Statistics Table 9 shows the exploratory statistics on OLID dataset for offensive language. There are a total of $9,532$ data instances. Out of this, $3,168$ are offensive (OFF) and $6,364$ are not offensive (NOT).

Additionally, here too we mention the monolingual vs. code mixed statistics for each class. We have $2,034$ monolingual and $1,134$ code mixed tweets in the OFF category and $3,578$ monolingual and $2,786$ code mixed tweets in the NOT category.

Keywords for Scraping Tweets Table 10 contains a list of 52 keywords that were used in the initial scraping of tweets, for creation of our novel dataset. These keywords were obtained from Twitter's top trending keywords list of the previous two years.

nationalism	open defecation	Muslims	marxists	JNU	UPA
demonetization	Farmer’s Bill	hijab	maoists	RSS	NDA
inflation	UAPA	triple talaq	Uyghur	PFI	Modi
unemployment	IPL	ghar wapasi	Pakistan	Gandhi	Rahul Gandhi
rape	ISRO	love jihad	Kashmir	Godse	Kejriwal
marital rape	migrants	CAA	China	Nehru	Emergency
secularism	lockdown	Shaheen Bagh	north-east	Sardar Patel	Indira Gandhi
urban floods	Covid-19	undertrials	drugs	Bhagat Singh	
lynching	Dalits	Adivasis	nepotism	Golwalkar	

Table 10: Keywords used for scraping tweets.