Cyberbully: Aggressive Tweets, Bully and Bully Target Profiling from Multilingual Indian Tweets

Suman Karan and Suman Kundu[0000−0002−7856−4768]

Dept. of Computer Science and Engineering, IIT Jodhpur, India
{karan.1,suman}@iitj.ac.in

Abstract. The present work proposes an end-to-end solution to identify a potential bully and bully-targets from multilingual aggressive tweets in Indian Twitter-o-sphere. The proposed work uses two LSTM based classifiers in pipeline to detect the tweet's language and aggressiveness. The model was trained with over 150,000 tweets of Hindi, English, Bengali, and Hinglish languages. F1 scores achieved for English, Hindi, Bengali, and Hinglish are 0.73, 0.83, 0.69, and 0.91, respectively. The paper further reported the patterns identified for several different attributes such as followers count, friends count, frequencies of tweets, and percentage of aggressive tweets of such potential bully and target users.

1 Introduction

People's ability to speak their minds freely on social media also leads to an increase in the use of profanity. For instance, tweets such as 'This incident happened in ****** r**di. Majority of muzlim nowadays. What can we expect' or 'I literally want to punch this b**ch' certainly hurt the target person emotionally or psychologically. Instead of having a firm mechanism to deal with such cases immediately, platforms like Twitter only support reporting/flagging. This kind of action takes time and hence becomes ineffective. Thus, automatic detection of aggression will aid in developing a self-quarantine system. Finding a potential bully or their victims can also help build community resilience.

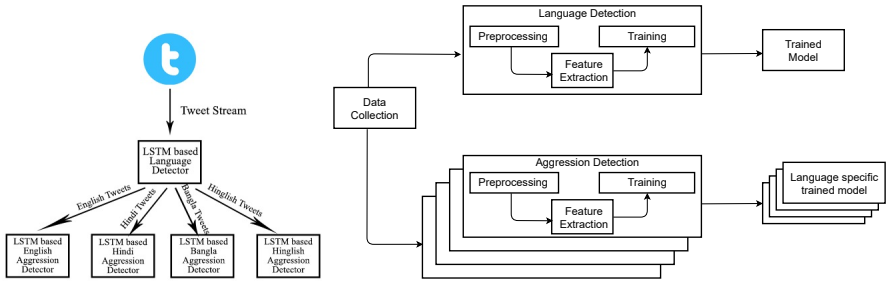
Bullying is a common issue with young children, and research has been conducted on it for a long time [1,2]. The earliest work on understanding the risk of violence in the digital world was studied by [3]. About a decade after, intensified research has been started on cyberbully [4,5,6,7]. The majority of the work is based on text analysis [5,8,9,10]. Major studies on aggression detection are for the English language [4,5,8,9,10] while Work for multilingual texts is limited [11,12]. Existing bi/multilingual algorithms either translate texts to English or manually segregate the documents and use separate language-based classifiers. India is a country of many languages, and its social media produces material in diverse languages. There is significantly less work on Indian languages.

This study proposed an end-to-end solution for identifying potential bullies and vulnerable targets from multilingual Indian Tweets. Experiments include Tweets in English, Hindi, Hinglish (Hindi written in Roman Script), and Bengali. The

procedure detects the language and selects the appropriate aggression detection classifier. Further, we use Twitter's public data to identify the behavior patterns of potential bullies and their victims.

The rest of the paper is organized as follows. The proposed multilingual aggression detection methodologies are presented in Section 2. Section 3 describes the bully and bully target identification and analysis along with experiments and results. Finally, Section 4 concludes the findings of the research.

2 Multilingual Aggressive Tweet Detection



(a) Overall Methodology (b) Block diagram of training

Fig. 1: Working model of the propose method

The proposed end-to-end solution takes a tweet and predicts its language on-the-go. According to the language, the text is passed to the respective language-specific aggression predictor. One trained language detector and four different aggression detector LSTM classifiers are trained (Fig. 1a) for Hindi, Hinglish, English, and Bengali texts. Unlike existing research, no translation is performed. The memory cell of LSTM helps to retain the essential parts of the sentence and reject the insignificant parts. Each LSTM unit has three gates to control what information to keep, and what to get rid of. These are the sigmoid function ($\sigma(.)$) where output 0 means block all information and 1 means keep all information. In order to remember long-term dependencies we have cell state (c_t). The following update equations are used in the experiments:

$$
\begin{aligned}
[l] f_t &= \sigma(W_f[x_t, h_{t-1}] + b_f) \\
i_t &= \sigma(W_i[x_t, h_{t-1}] + b_i) \\
o_t &= \sigma(W_o[x_t, h_{t-1}] + b_o)
\end{aligned}
\qquad
\begin{aligned}
[l] c_t &= f_t * c_{t-1} + i_t * \tilde{c}_t \\
\tilde{c}_t &= tanh(W_c[x_t, h_{t-1}] + b_c)
\end{aligned}
$$

where f_t, i_t, o_t are function of the forget, input, and output gates respectively, h_{t-1}, x_t is the output of previous and input of the current LSTM unit, and W_x, b_x are the weight and bias of the respective neuron. $\tilde{c}_t$ is the candidate cell-state and c_t is the cell-state at timestamp t. The final output pass to the next cell is

calculated by $h_t = tanh(c_t) \times o_t$. The model designed here is scalable in terms of languages, allowing us to add a new language by updating and retraining the language detector model, and prepare a new aggression detector for the concerning language without tinkering with other models. The block diagram of the training is shown in Fig. 1b.

2.1 Data Collection

Different layers of LSTM are trained on different data collected from different sources such as Twitter and Wikipedia. The data and codes are available at https://figshare.com/s/ce91ed033c29f7379cdb.

Data for language detection model: Data for three regular languages, English, Bengali, and Hindi were scrapped from Wikipedia articles. Hinglish which is a code mixed between English and Hindi is not available on Wikipedia, hence, downloaded from [13]. Sample data is shown in Fig. 2a.

মাতৃভাষীর সংখ্যায় বাংলা ইন্দোইউরোপীয় ভাষা পরিবারের চতুর্থ ও বিশ্বের ষষ্ঠ বৃহত্তম ভাষা	Bangla
অনেক বাঙালি এমনকি ইংরেজি কিংবা সংস্কৃত থেকে ধারকৃত শব্দ উচ্চারণের ক্ষেত্রেও এই ধারাটি বজায় রাখে যেমন গ্রাম উচ্চারণ করেন গেরাম, স্কুল উচ্চারণ করেন ইস্কুল হিসেবে'	Bangla
They seem to be a mainstay of condominium resorts like The Margate at Winnepesaukee a Premier Lakes Region Resort which suggests ownership Winterberry a Resort Village and Moon Ridge a Point of View which contains a nice whimsy	English
The White House Inn for example has become The White House a Country Inn	English
अगर आप गेट की परीक्षा पास करके एम टेक करेंगे तो आपको एक अच्छी स्कालरशिप भी मिलेगी	Hindi
इन कॉलेजो में प्रवेश गेट स्कोर साक्षात्कार लिखित परीक्षा तथा अनुभव के आधार पर मिलता है	Hindi
doktar grant mujhe yada kada varneriyan sosayati kee sabhaon men bhi le jate the	Hinglish
san kee garmiyon men mainne apane do doston ko sath liya apaneapane pitthoo thaile lade aur north vels kee sair ko paidal hi nikal gae	Hinglish

(a) For language detection model

ভাগ্য ভাল ভিডিওটা করার সময় বুঝতে পারেনি ওরা না হলে তোমাকে ও চু* দিতো	AG
৮০% পতিতা নিজের ইচ্ছায় এই কাজ করের না মা durga 'র murti এই পতিতাদের বাড়ির মাটি দিয়ে তৈরী হয় মেয়েদের somman করতে শেখ	NAG
Confused! Guys new video is up, do check it out, thank you. https://goo.gl/XuclJV	NAG
What has so far Mr.Yechuri done for this Country. Ask him to shut down his bloody p**hole for good or I if given the chance will crap on his mouth hole.	AG
भारतीय मीडिया को मिला दलाली का प्रमाणविश्वसनीयता पर 180 देशों में 136 वां स्थान !	AG
जेडीयू ने मांग की कि लोकसभा में विपक्ष का नेता कांग्रेस से होना चाहिए	NAG
Pad nahi Mila to bhavuk hona to lazimi he.Amanatullah ne konsa jhooth bola tha	NAG
Chl be fake account jyada na bol Himmat hi ni hJa na apne page me Baar 2 pta ehi hu pakistan k page jaNi h na to bna leTra desh kya ho rha h bhut acche se pta nikl lo and wIt and watch india ku6 t9 krega zarur fir puchenge	AG

(b) For aggression detection model

Fig. 2: Sample Data

Data for aggression detection model: We used TRAC data sets [14,15] with label OAG (Openly Aggressive), CAG (Covertly Aggressive), and NAG (Non-Aggressive) and converted them into binary by considering OAG and CAG as Aggressive (AG). A combined TRAC 1 and 2 data sets are used for training the model to increase accuracy. We collected 4952, 1583, 12210, and 17169 rows of data for Hindi, Bengali, Hinglish, and English respectively. Fig. 2b shows a sample data set used in the model.

Language Specific Slang Words: Slang words or curse words are often used by users to vent their aggression. We did not find any organized collection of

slang words. We created lists of slang words from various websites and personal experiences. The lists consists of 67 English ([16,17,18]), 40 Bengali [19,20], 71 Hindi [21,22,23] slang words. We converted Hindi slang words into code mixed Hinglish which resulted in 125 slang words for Hinglish language.

2.2 Data Preprocessing

We removed punctuation, numbers, and URLs from language and aggression detection data as these do not have any significance. All letters in English and Hinglish texts are converted to lowercase. No stop words were removed from the sentences while used for the language detection model. Stop words play a significant role in language detection, as their frequent presence helps the model better learn what language it is. Apart from the this, we removed language-specific stop words, white spaces, and new line characters from aggression detection data. As it is common to have user mentions in tweets, these names are used latter to identify vulnerable users but removed as part of preprocessing.

2.3 Language Detector Model

Feature Extraction: We first vectorized the texts by text tokenizer function (Python Tensorflow Keras library) with the most frequently used 500K words. A fixed length of 250 words is used for representing any text. We truncated the input sequence when it is more than 250, and padded with 0 if it is less than 250. As Twitter only allows 280 characters, 250 words are sufficient.

The Model: LSTM is used for classification. The embedding layer encodes the input sequence into a sequence of dense vectors of size 100. Dropout and recurrent dropout are set to 0.2, and SoftMax is used as an activation function. We used adam optimizer. Categorical cross entropy $\rho = -\sum_{c=1}^{C} y_{o,c} \log(p_{o,c})$ is used as a loss function where C is the set of classes (e.g., 4 languages here), $y_{o,c}$ is a binary value (0 or 1) if class c is the correct classification for the observation o, and $p_{o,c}$ is the predicted probability for the observation o is of class c

2.4 Aggression Detection

Feature Extraction: We vectorized the text using the tokenizer (Python Tensorflow Keras library), limiting the corpus size to 50K most frequent words for each language. The maximum number of words in a text is kept to 250. Density of slang words, i.e., the number of slang words in a sentence and emojis are considered as features. Density of capital letters is used as another feature as it is generally used for screaming. Similarly, we observed frequent use of question marks and exclamation marks leads to aggressive text. We counted the occurrence of those as a feature. Note that this feature is extracted from the original text rather than the preprocessed text. Polarity in sentiment analysis is identifying sentiment orientation, while subjective expressions are opinions that describe people's feelings towards a specific subject or topic. We used polarity and subjectivity scores of each sentence as features.

The Model: An LSTM classifier is used with embedding dimension, output vector size, dropout, and recurrent dropout set to 100, 128, 0.2, and 0.2. SoftMax is used as the activation function, and cross-entropy is used as a loss function.

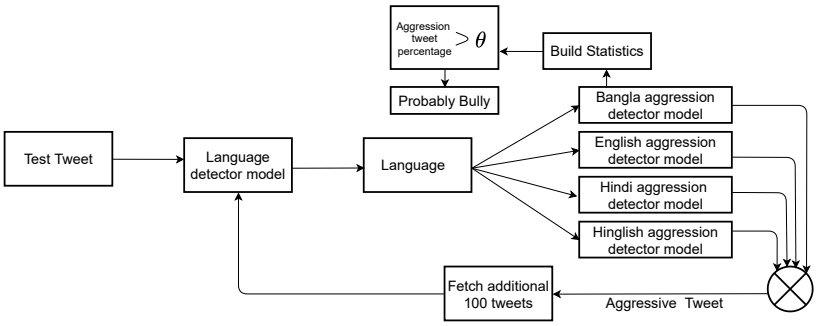


Fig. 3: Block diagram of bully detection

3 Bully and Bully Target Identification and Analysis

Fig. 3 depicts the method by which we detect bullies using the trained model described previously. Once a tweet is determined to be aggressive, the system retrieves 100 of the author's most recent tweets. The model then evaluates and categorizes all of these tweets as aggressive or not. If θ (user-defined) percent, of these tweets, are aggressive, the user is flagged. In addition, we extracted every user mentioned in the aggressive tweet and designated them as potential targets. Our experiment evaluates the user profiles, number of friends, and followers of bully and bully targets by identifying hidden patterns in it.

Table 1: Precision, Recall and F1 score for aggression detection.

Language	Precision	Recall	F1 Score
English	0.7255	0.7382	0.7318
Hindi	0.8427	0.8255	0.8341
Bengali	0.8125	0.6046	0.6933
Hinglish	0.8735	0.9512	0.9107

3.1 Experiments and Results

All the experiments are conducted on Windows 10 PC with Intel Core i5, 2.2GHz processor, and 16GB ram. Python 3.6 with TensorFlow is used. The batch size

of 64, and 15 epoch is used for both the language and aggression detection. The epoch size is kept low as the loss is found to be converged before 15 and no further significant reduction in loss value is observed. The data set for language detection contains 176681 samples, 57991 English, 54920 Hindi, 50954 Bengali, and 12816 Hinglish data. The overall accuracy achieved is 99.97% on the test data set. On the other hand, the data set for aggression detection has 35,914 data samples with 4952 Hindi, 1583 Bengali, 12210 Hinglish, and 17169 English data. For both cases, we used 80% for training, 10% for validation, and the rest 10% for testing. Table 1 shows the comparison matrices for all four languages.

3.2 Analysis of Bully and Vulnerable Targets

We ran our algorithm on approximately 50000 tweets and discovered 1515 users to be aggressive. These users are suspected of being the bully. However, it may be a one-time outburst instead of a personality trait, so additional 100 recent tweets are retrieved. In addition, we extracted all mentioned users from aggressive tweets. These are potential targets. 100 tweets from target accounts were also extracted. In case users have less than 100 tweets we retrieved them all. Further, we collected public information about users, including follower and following counts. We calculated tweet frequency per day using tweets that we retrieved.

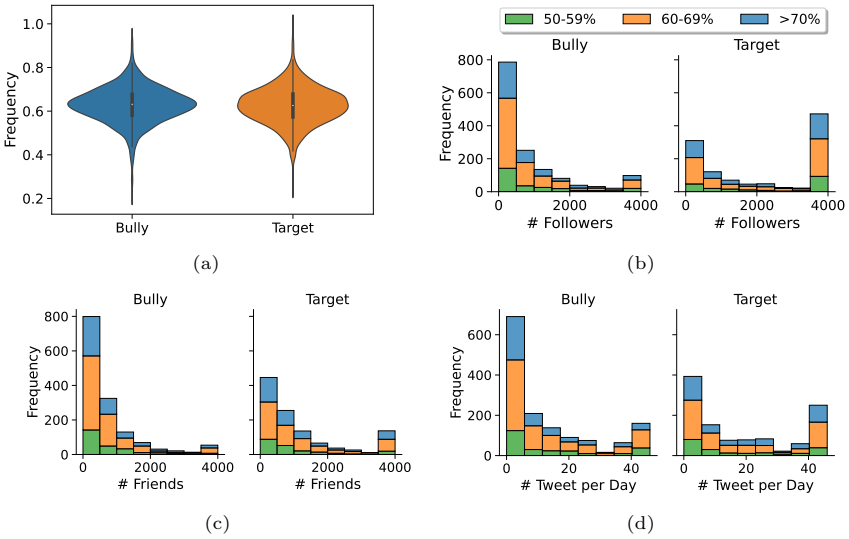


Fig. 4: Different statistics for predicted bully and target users.

Patterns in Bully: Experimental results show that many users post high (60% or more) aggressive tweets (Fig. 4a). We further study these suspected

bully users. Most suspected bullies have followers (Fig. 4b) below 500. There are, however, a significant number of users with followers of more than 4000. Similarly, most bullies follow (Fig. 4c) less than 1000 people, specifically, with less than 500 following. It is also found that most of the suspected bully tweets between 0 to 10 tweet (Fig. 4d) per day. However, there are a significant number of users who have tweets between 60 to 80 per day.

Patterns in Predicted Targets: Our experiment found about 1100 mentions (target) in the aggressive tweets. Interestingly, many of these targets themselves post aggressive tweets (Fig. 4a). Hence, we also analyzed target users in the same glass as of the bully. Similar to the prospective bully, where the majority have 50% or more aggressive tweets, most of their targets have aggressive tweets between more than 50% (Fig. 4a). Users having a follower count greater than 4000 is more for target users (Fig. 4b). It indicates that these targets, with more than 4000 followers, are celebrity users and are subject to aggressive targets. As we can see, the targets of aggressive tweets may also have bully characteristics, so a deeper understanding of the profile is required apart from just classifying a post as aggressive or bully-post. It may not be that a target user is always a victim.

4 Conclusion

In this paper, we presented a multilingual cyberbully detection from Twitter. The work is done on the Indian text of Hindi, English, Bengali, and Hinglish languages. The model is found to be effectively detecting aggressive tweets. The model automatically detects the language of the tweet and accordingly selects the corresponding aggression detection model. Although the present work only consider 4 languages, it can be extended by attaching any language specific aggression detection module in the pipeline. The paper also presented the patterns in the users' public profiles who like to post aggressive tweets. One interesting fact reported here is that mentions in the bully posts are also very aggressive in their own tweets. It would be an interesting research work for the future to analyze further the reciprocity of bully posts and its contagiousness.

References

1. F. Chang and B. M. Burns, "Attention in preschoolers: Associations with effortful control and motivation," *Child Development*, vol. 76, pp. 247–263, jan 2005.
2. P. C. Rodkin, T. W. Farmer, R. Pearl, and R. V. Acker, "They're Cool: Social Status and Peer Group Supports for Aggressive Boys and Girls," *Social Development*, vol. 15, pp. 175–204, may 2006.
3. I. R. Berson, M. J. Berson, and J. M. Ferron, "Emerging risks of violence in the digital age: Lessons for educators from an online study of adolescent girls in the United States," *Journal of School Violence*, vol. 1, pp. 51–71, mar 2002.
4. D. Yin, Z. Xue, H. Liangjie, B. D. Davison, A. Kontostathis, L. Edwards, and L. Edu, "Detection of Harassment on Web 2.0," in *Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*, (Madrid), 2009.

5. K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," *Proc of the International AAAI Conference on Web and Social Media*, vol. 5, no. 3, pp. 11–17, 2021.
6. S. Hinduja and J. W. Patchin, "Social Influences on Cyberbullying Behaviors Among Middle and High School Students," *Journal of Youth and Adolescence*, vol. 42, no. 5, pp. 711–722, 2013.
7. A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: Query terms and techniques," in *Proc of the 5th Annual ACM Web Science Conference*, p. 195–204, 2013.
8. C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, "Detection and fine-grained classification of cyberbullying events," in *International Conference Recent Advances in Natural Language Processing (RANLP)*, pp. 672–680, 2015.
9. B. S. Nandhini and J. I. Sheeba, "Cyberbullying detection and classification using information retrieval algorithm," in *Proc: International Conf. on Advanced Research in Computer Science Engineering & Technology*, pp. 1–5, 2015.
10. H. Zhong, D. J. Miller, and A. Squicciarini, "Flexible Inference for Cyberbully Incident Detection," in *Proc of Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*, (Wurzburg), pp. 356–371, 2019.
11. A. Malte and P. Ratadiya, "Multilingual cyber abuse detection using advanced transformer architecture," in *TENCON IEEE Region 10 Conference*, pp. 784–789, 2019.
12. S. Si, A. Datta, S. Banerjee, and S. K. Naskar, "Aggression Detection on Multilingual Social Media Text," in *Proc: International Conf. on Computing, Communication and Networking Technologies*, pp. 1–5, IEEE, 2019.
13. KaggleDataset, *Code Mixed (Hindi-English) Dataset*, 2018.
14. R. Kumar, A. N. Reganti, A. Bhatia, and T. Maheshwari, "Aggression-annotated Corpus of Hindi-English Code-mixed Data," in *Proc of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association (ELRA), 2018.
15. S. Bhattacharya, S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, B. Lahiri, and A. K. Ojha, "Developing a multilingual annotated corpus of misogyny and aggression," in *Proc of the 2nd Workshop on Trolling, Aggression and Cyberbullying*, (Marseille), pp. 158–168, European Language Resources Association (ELRA), 2020.
16. "26 english swear words that you should use very very carefully." `https://www.rypeapp.com/blog/english-swear-words`. Accessed: 2021-02-02.
17. "Category:english swear words." `https://en.wiktionary.org/wiki/Category:English_swear_words`. Accessed: 2021-02-02.
18. "A definitive ranking of every swear word from worst to best." `https://tinyurl.com/ranking-swear`. Accessed: 2021-02-02.
19. "Urban thesaurus." `https://tinyurl.com/bnslang`. Accessed: 2021-02-02.
20. "Bengali slang words with meaning (bengali slang dictionary)." `https://tinyurl.com/bengali-slang`. Accessed: 2021-02-02.
21. "40+ hindi gaaliyan in english that you have to know being indian!." `https://tinyurl.com/Hindi-Gaaliyan`. Accessed: 2021-02-02.
22. "Hindi language blog." `https://tinyurl.com/slang-in-hindi-i`. Accessed: 2021-02-02.
23. "Slay the chats with the most popular hindi texting slang!." `https://tinyurl.com/hindi-text-slang`. Accessed: 2021-02-02.