TABHATE: A Target-based Hate Speech Detection Dataset in Hindi

Deepawali Sharma
Banaras Hindu University
Vivek Kumar Singh (✉ vivekks12@gmail.com)
Banaras Hindu University
Vedika Gupta
Jindal Global Business School, O.P. Jindal Global University

Research Article

Keywords: Hate Speech, Hate Speech Corpus, Hate Speech Dataset, Hindi language, Deep Learning

Posted Date: April 20th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2800717/v1

Abstract

Social media has over the years provided a medium for creation and dissemination of opinions and thoughts through online platforms. While it allows users to express their views, sentiments and emotions, some people try to use it to generate and share unpleasant and hateful content. Such content is now referred to as hate speech and it may target an individual, a group, a community, or a country. During the last few years, several techniques have been developed to automatically detect and identify hate speech, offensive and abusive content from social media platforms. However, majority of the studies focused on hate speech detection in English language texts. With social media getting higher penetration across different geographies, there is now a significant amount of content generated in various languages. Though there have been significant advancements in algorithmic approaches for the task, the non-availability of suitable dataset in other languages poses a problem in research advancement in them. Hindi is one such widely spoken language where such datasets are not available. This work attempts to bridge this research gap by presenting a curated and annotated dataset for target-based hate speech (TABHATE) in the Hindi language. The dataset comprises of 2,020 tweets and is annotated by three independent annotators. A multiclass labelling is used where each tweet is labelled as: (i) individual targeting, (ii) community targeting, and (iii) none. Inter annotator agreement is computed. The suitability of dataset is then further explored by applying some standard deep learning and transformer-based models for the task of hate speech detection. The experimental results obtained show that the dataset can be used for experimental work on hate speech detection of Hindi language texts.

1. Introduction

Social media allows us to communicate with others, share our thoughts/opinions/feelings, and build social networks. According to a recent report[2], there are now 4.76 billion active social media users, which is about 59.4% of the total world population. The penetration of social media is found to have increased from 1,720 million users in January 2013 to about 4,760 million users in January 2023. The large number of users has resulted in a huge volume of content on various social media platforms. Social media is now used by people from diverse backgrounds for a variety of purposes. In fact, it is also being used as the new means to protest, and fight for certain social, economic and community-specific causes. Online social media users are known to provide valuable support to those individuals or communities who experience exclusion or have disabilities or chronic illnesses. However, at the same time, it has also been abused as a medium to spread hate and target certain individuals, groups and communities. One such instance of abuse is known as hate speech.

Hate speech has been defined as harmful content that directly encourages or promotes hate against an individual or community on the basis of race, religion, caste, age, sex, sexual orientation, religion, and ethnicity [37]. Those comments or posts on social media that humiliate, insult or abuse a community or a member of the community and organization are termed hate speech on social media [38]. Hate speech and other offensive languages have a difference on the basis of subtle linguistic distinctions [39]. An example of hate speech could be: "*The Jews are again using holohoax as an excuse to spread their agenda. Hitler should have eradicated them*".

The spread of hateful content against any individual or community might pose serious problems including on people's mental health. Therefore, it is very important to have algorithmic approaches to automatically detect such content so that appropriate corrective action can be taken in time. Fortunately, there have been significant advancements in computational approaches for automatic detection of such content. For example, several studies have been carried out on the detection of hate speech in the English language [1][2][3][4]. Similarly, the detection of offensive or abusive content on social media in the English language is presented in [5][6][7][8]. There are several datasets available for hate speech and offensive content detection research in the English language [9][10][11]. However, there is relatively very less research in other languages. One major reason for absence of research in various languages is unavailability of suitable datasets for the purpose. Hindi is one such language which is widely spoken (across 20 countries with 577 million native speakers) and has now significant amount of online content available, but as on date there is not appropriate dataset available for hate speech detection research. Therefore, this work attempts to bridge this research gap by presenting a manually curated and annotated dataset of tweets in Hindi language. The target-based hate speech dataset (referred as TABHATE) is curated by collecting tweets in Hindi and getting them annotated by three independent annotators. The quality of annotation is computed by calculating a standard measure of inter-annotator agreement. Thereafter, various deep learning models (CNN, LSTM, Bi-LSTM) and transformer-based models (mBERT, muRiL BERT, Indic BERT, XLM-Roberta) are applied on the dataset.

The main contribution of this paper can be summarised as follows:

- It provides a curated and annotated dataset of 2,020 tweets in Hindi language for hate speech detection research. Multi-class tagging of tweets is used, where each tweet is labelled as individual targeted, community targeted and none.
- The suitability of the dataset is evaluated by computing standard measure of inter-annotator agreement.
- A set of experiments are carried out by deploying various deep learning and transformer based state-of-the-art computational models for detection of hate speech in the tweets.

The rest of the paper is organized as follows: Section 2 presents the relevant related work on hate speech detection, including the existing datasets for the purpose. Section 3 describes in detail the dataset construction and annotation procedure and quality. Section 4 presents the experimental work carried out on the curated dataset, with results and evaluation described in the Section 5. The paper concludes with a brief summary and contribution of the work in Section 6.

[2] https://datareportal.com/reports/digital-2023-global-overview-report

2. Related Work

Several studies have been carried out on the detection of hate speech in the English language texts [1][2][3][4]. Similarly, the detection of offensive or abusive content on social media in the English language has also been explored [5][6][7][8]. There are several datasets available for hate speech and offensive content detection research in the English language. For example, a dataset for explainable hate speech detection known as HateXplain is available in the English language [9]. The dataset was collected from Twitter and Gab, and the posts are classified into three categories: hate, offensive and normal. Some studies experimented with the HateXplain dataset to classify the posts into given categories [17][18][19]. Another dataset ETHOS is contributed in the English language to detect online hate speech [10]. The dataset was collected from YouTube and Reddit comments. The dataset has two variants: binary classification and multi-label classification. In binary classification, comments contain hate or not. For multi-label classification, there are eight labels: violence, directed_vs_generalized, race, national_origin, disability, sexual_orientation, and religion. Similarly, authors in [11] presented a benchmark dataset for learning to intervene in online hate speech. This dataset was collected from Reddit and Gab. The posts or comments are classified as hate speech or not. Thus, there are several datasets for hate speech research in English language. However, when it comes to other languages, such resources are scare.

There are only few datasets related to offensive content detection in Hindi and code-mixed Hindi-English language. One such dataset is on hate speech detection from social media text is in Hindi-English code-mixed language [13]. Here, the tweets are classified as hate speech or normal speech. Some studies used this dataset to carry out experiments [24][25][26]. Similarly, conversational hate speech detection in code-mixed language dataset is available in Hindi and English language [14]. The dataset was collected from Twitter. About 7000 posts in code-mixed Hindi and English are classified as non-hate offensive (NOT) and hate and offensive (HOF). Few studies used this dataset to perform the subtask to classify the posts as NOT and HOF in code mixed language [27][28][29]. Another dataset is there to detect hate speech and offensive content identification in Indo-European languages [15]. This dataset is in English, Hindi and German languages. In this dataset, binary classification is used with tweets classified as hate and not offensive. For fine-grained classification, the tweets are classified into three classes: hate speech, offensive and profanity. Several studies experimented on this dataset [30][31][32][33]. The Dravidian code-mixed offensive span identification dataset is another dataset in regional languages [16]. This dataset is available in under-resourced Tamil-English and Kannada-English code-mixed. The dataset was annotated for offensive spans. A summary of all the major datasets available in different languages for hate and offensive speech detection is provided in Table 1.

Table 1
Major Datasets on Hate and Offensive Speech Detection

Name of the Dataset	Description				Link
	Source	Size	Class/Labels	Language	
Explainable Hate Speech Detection (HateXplain) [9]	Twitter and Gab	20,148	Hate, offensive and normal	English	GitHub - hate-alert/HateXplain
Online Hate Speech Detection (ETHOS) [10]	YouTube and Reddit		Binary classification: Hate or not Multi-label classification: violence, directed_vs_generalized, race, national_origin, disability,sexual_orientation, and religion	English	https://github.com/intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset
Online Hate Speech [11]	Gab and Reddit	22,324 comments-Reddit 11,825 comments-Gab	Hate speech and Not hate Speech	English	GitHub - jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech
Hindi-English Code-Mixed Social Media Text for Hate Speech Detection [13]	Twitter	4,575	Hate speech or Normal speech	Hindi-English code-mixed	GitHub - deepanshu1995/HateSpeech-Hindi-English-Code-Mixed-Social-Media-Text
Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC-2019) [15]	Twitter, Facebook	2,963-Hindi 2,373-German, 3,708-English	Non-Hate-Offensive, Hate and Offensive.	Hindi, English and German	https://hasocfire.github.io/hasoc/2019/dataset.html
Offensive language Identification-DravidianCodeMix (HASOC-Dravidian-CodeMix)[40]	Twitter, YouTube	4000 comments and tweets	Offensive and Not-offensive	(Tamil and Malayalam) code-mixed	https://dravidian-codemix.github.io/HASOC-2021/datasets.html
Dravidian Code-Mixed Offensive Span Identification (DOSA) [16]	YouTube	4786-Tamil-English 1097-Kannada-English	Offensive or not offensive spans	(Tamil and Kannada) code-mixed	GitHub - manikandan-ravikiran/DOSA: Dravidian Code-Mixed Offensive Span Identification Dataset

There is relatively higher set of resources for hate and offensive speech detection in English and code-mixed languages. The Hindi language, which is a widely spoken language has received less attention, with only few datasets being there for related tasks. The major existing datasets on hate speech detection in the Hindi language uses binary labelling of hate and non-hate speech. There are no existing datasets that label the posts/ comments based on target of the hate speech in the Hindi language. Detection of hate speech and its target is very important as such information can be used to prevent adverse effects on society. If the comments/posts or tweets that spread hate are detected as the individual targeted or community targeted, then they may be flagged so and appropriate corrective action can be taken. This work presents a curated and annotated dataset on target-based hate speech detection, which is first of its kind. The tweets are annotated as individual targeted, community targeted and none.

3. Dataset Curation

This section provides the details of the data collection and annotation strategy along with details of inter-annotator agreement and a summary of the main statistics of the dataset. The Fig. 1 presents an illustration of the various steps in the dataset creation.

3.1 Data Collection

The data is collected from Twitter which is a widely used platform in the world by various people to express opinions on a particular topic, incident, person, or community. To collect data, the following approach was used:

- Twitter is chosen as a platform to collect the data. The Twitter API[3] is used to collect the tweets. The tweets are collected using a set of keywords and the names of potential victims of hate speech such as women, LGBTIQ+, gay, lesbian, hijab and some social issues (such as Indian farmer protest, domestic violence) and the name of some politicians etc. Some specific hashtags are used to collect the tweets on target-based hate speech, such as #savegirlchild, #stopviolence, #metoo, #savefarmer, #LGBTIQ, #Hijab Row, #womeninstem, #lgbtcommunity, #BreakTheBias, #feminism, #islam, #boycottMuslim, #kisanektazindabad, #speakupforfarmers, #emotionalabuse.
- The duplicate tweets in the data so collected are removed and the reposts are not considered. The non-text entities like links, videos and pictures are removed. However, emojis are not removed from the text as they might carry information for the target-based hate speech labelling task. The mentions '@' are removed to hide the identity of a particular person. Thus, 2,020 tweets are collected based on these keywords and specific hashtags.

The tweets of the dataset are mostly related to the minority groups, such as women, LGBTIQ+, and many more. Therefore, the individual identity is removed by removing the mentions and personal information from the dataset. However, since the dataset is on a target-based hate speech, the information related to gender, race, ethnic origin, religion, and other beliefs are needed to be retained in the dataset.

3.2 Annotation

The tweets in the dataset are manually annotated by three independent different annotators. The annotators were postgraduate students in the field of computer science, having Hindi as their native language. The annotators were provided with sufficient time for the annotation. Before starting the annotation task, the annotators were clearly instructed about the goal of the annotation task and how to annotate the tweets considering the definition for each category. They were also informed that the annotation task will involve reading hateful and offensive content.

The annotators were informed that the tweets need to be categorized into three classes: individual targeted, community targeted and none. Following instruction was provided for this purpose:

a. Those tweets that are against any person or targeted any individual based on their race, religion, gender, or any psychological beliefs are to be labelled as the individual targeted.
b. Those tweets that are against or that attack any community/religion/organization or political parties and spread hate or false beliefs about a particular community are to be labelled as community targeted.
c. Those tweets that do not spread hate against any individual, community, organization or political party are to be labelled as none.

Each annotator was asked to assign the class label of "0", or "1" or "2" to each tweet in the dataset. A value '0' indicated 'none' class, value of '1' indicated 'individual targeted' and value of '2' indicated 'community targeted'. The annotators read the tweets and assigned the class label accordingly.

One randomly selected example tweet for each of the three classes, and the English language translation of the tweet, is provided in the Table 2 below.

Table 2
Examples of annotated tweets in three classes

S.No.	Tweets	Label
1.	“ , “ https://t.co/C8RI3t6CaS Translation: The embarrasement of Iran's President muslim islam Ibhrahim Raisi in America. The anchor did not wear hijab, then the Iranian President did not give an interview, after which there was trouble.	Individual targeted
2.	“ “ #Women https://t.co/50Wn2KsvGu Translation: Teach daughter, fight with each other.	Community targeted
3.	“ : - DNA ?” Translation: Mohan Bhagwat said that the DNA of Hindu-Muslim is the same. The country is not complete without Muslims. What do you think?	None

3.3 Inter-Annotator Agreement

Before the dataset can be used for the task of hate speech research, the quality of the annotation needs to be evaluated. Therefore, after obtaining the annotated files from the three independent annotators, the inter-annotator agreement was computed. The Fleiss Kappa measure [34] was computed to measure the inter-annotator agreement. The Fleiss Kappa is used to measure the inter-annotator agreement between three or more annotators on a categorical scale. The curated dataset is labelled into three categories by the annotators and hence the annotation produces categorical data. Therefore, the Fleiss kappa is used as a measure to determine the level of agreement between the annotators.

The Fleiss Kappa measure can be computed as follows. Let N be the total number of documents to be annotated (2020 tweets in this case); and n be the number of annotators of the documents to be annotated (three annotators); and k be the number of categories (three categories, where 0-> none, 1->

individual targeted and 2-> community targeted) Let i = 1,……N represents each document and j = 1,……k represent the categories of the annotation. Let n_{ij} be the number of raters who assign the i^th document to the j^th category, then p_j (which represents the proportion of all assignments that were made to the j^th category) is given by,

$$p_j = \left(\frac{1}{Nn}\right)\sum_{i=1}^{N} n_{ij}$$

1

The scope of agreement among the n annotators for the i^th document is given by the proportion of all the n(n-1) possible pairs of assignment. This proportion P_i is given by, The Overall extent of agreement $\bar{P}$ can be given by the mean of all the $P_i\ (s)$ calculated in the Eq. (2**)**

$$P_i = \left(\frac{1}{n(n-1)}\right)\sum_{i=1}^{k} n_{ij}\left(n_{ij}-1\right) = \left(\frac{1}{n(n-1)}\right)\left(\sum_{j=1}^{k} - n\right)$$

(2)

$$\bar{P} = \left(\frac{1}{N}\right)\sum_{i=1}^{N} P_i$$

(3)

The mean proportion of agreement $\bar{P_e}$ would be given by,

$$\bar{P_e} = \sum_{j=1}^{k} {p_j}^2$$

(4)

The Fleiss Kappa (k) is given by the formula,

$$k = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}}$$

(5)

where, $1 - \bar{P_e}$ denotes the degree of agreement among the annotators attainable, while the value $\bar{P} - \bar{P_e}$ denotes the degree of agreement among the annotators that is actually achieved. The Fleiss Kappa statistic ranges from − 1 to 1, where a value of less than or equal to 0 indicates no agreement, value between 0.01–0.20 indicate none to slight agreement, value between 0.21–0.40 indicates fair agreement, value between 0.41–0.60 indicates moderate agreement, value between 0.61–0.80 indicates substantial agreement, and value 0.81–1.00 indicates almost perfect agreement [35]. In the present case, the value obtained was 0.77, which indicates a substantial agreement among annotators and hence makes the dataset suitable for use.

3.4 Dataset Statistics

The dataset contains 2,020 tweets of which 182 belong to the individual targeted, 243 belong to the community targeted and 1595 belong to none category. Table 3 represents the distribution of tweets into three categories: individual targeted, community targeted and none.

Table 3
Category-wise Data Distribution

Category	No. of Tweets Before Oversampling
Individual Targeted	182
Community Targeted	243
None	1595

[3] https://developer.twitter.com/en/docs/twitter-api

4. Experimental Setup

Prior to implementing the various models, the dataset is pre-processed as explained earlier. Then tweets are tokenized and the input of uniform size need to fed to the model. Therefore, padding is applied on the sequences to make the shorter sequences to the same length as the sequence of maximum length. The tokens are fed into the embedding layer. For the implemented deep learning models (CNN, LSTM, Bi-LSTM) GloVe embedding is used. The dataset is imbalanced, containing 182 tweets for the 'individual target', 243 tweets for the 'community target' contains and 1595 tweets for 'none' category. Therefore, oversampling of data is performed using Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is an algorithm that solves the problem of an imbalanced data set [36]. It performs data augmentation by creating synthetic data points that are based on the original data points. SMOTE does not generate duplicates, it creates synthetic data points that are a bit different from the original. Table 4 represents the distribution of tweets before and after oversampling into three categories: individual targeted, community targeted and none. The annotated dataset is split into the train, validation, and test set. The training set contains 80%, the validation set contains 10% and the test set contains the remaining 10% of the data as shown in Table 5.

Table 4
Distribution of tweets before and after oversampling

Category	No. of Tweets Before Oversampling	No. of Tweets After Oversampling
Individual Targeted	182	1595
Community Targeted	243	1595
None	1595	1595

Table 5
Train-Valid-Test data distribution

	No. of Tweets Before Oversampling	No. of Tweets After Oversampling
Training	1618	3830
Validation	200	479
Testing	202	476

4.1 Models

Convolutional Neural Network (CNN)

After pre-processing, the data is tokenized and these tokens are fed to the embedding layer and the tokens are converted into their corresponding vectors. The embedding vectors are fed to the model. For text classification, 1D convolutional neural network is used. In this model, three layers convolutional layers are added with ReLU as an activation function. 1D max pooling layer is added for downsampling by diving the input. Finally, its output is fed to the final or dense layer. Now, the model is trained on 50 epochs with softmax as an activation function and classification is done into the given categories: Individual targeted, community targeted and none. Figure 2 presents the model architecture.

Long Short-term memory (LSTM)

In this implementation, the data is tokenized after pre-processing and pad_sequences() is used for padding the sequences to make the shorter sequence of same length as the length of the longest sequence. These tokens are fed to the embedding layer which is first layer with 300-length vectors to represent each word. The SpatialDropout1D used to perform variational dropout. The model uses the three LSTM layers with ReLU as an activation function. The last layer is the dense layer with softmax as an activation function for multi-class classification. Figure 3 presents the model architecture. The categorical_crossentropy loss is used as a loss function for multi-class classification.

Bidirectional LSTM (Bi-LSTM)

In bi-directional Long short-term memory, the information flows in both directions; forward and backward. After tokenization, the tokens are fed into the first layer which is the embedding layer. The tokens are represented as vectors. The next layer is Bi-LSTM in which vectors are passed as input. The final layer is the dense layer with an activation function to classify the comments. Figure 4 presents the model architecture. The model is trained on 100 epochs using the categorical_crossentropy loss is used of batch size 32.

Multilingual BERT (mBERT)

mBERT is multilingual BERT trained on 104 languages using masked language modelling (MLM). Firstly, the data is pre-processed, and the pre-processed data is tokenized using Bertokenizer. The two special tokens [CLS] and [SEP] are added at the start and end of the sequence, respectively. The tokenized data is fed to the BERT model as an input. Figure 5 presents the model architecture. The pre-trained BERT-base-multilingual-cased model used to build the model. The model is trained on 10 epochs and set the learning rate is set to be 3e-5 with "AdamW" as an optimizer. The Hindi tweets are classified as individual targeted, community targeted and none.

XLM-RoBERTa

XLM-RoBERTa is pre-trained on 100 languages and has a powerful vocabulary. It is a multilingual model and based on the RoBERTa. The architecture of RoBERTa is same as the BERT architecture. RoBERta is pre-trained on a large corpus in a self-supervised manner. After pre-processing the data, the data is tokenized using the XLM-RoBERTa tokenizer. The token_ids is generated. The XLMRoBERTaForSequenceClassification function calls using the pre-trained xlm-roberta base model. The model is trained for 10 epochs using the “AdamW” as an optimizer and the learning rate is set to be 3e-5. The tweets are classified as individual targeted, community targeted and none.

IndicBERT

IndicBERT model is trained on 12 Indian languages: Tamil, Bengali, Marathi, Punjabi, English, Hindi, Gujarati, Malayalam, Assamese, Oriya, Telugu and Kannada. IndicBERT has a smaller number of parameters as compared to the other multilingual models like mBERT and XLM-RoBERTa. In IndicBERT, the number of tokens is 1.84B in pretraining data for the Hindi language. This pre-trained IndicBERT model is based on the ALBERT model. The Autotokenizer is used to tokenize the text and [CLS] and [SEP] are the two tokens that are added at the beginning and end of the sequence, respectively, similar to the BERT model. IndicBERT also has the ability to consider more than one sentence into a single sequence for input. The hyperparameters are fine-tuned: learning rate is set to be 3e-5, epochs = 10 and batch size is 32. At the end, the model classifies the tweets into three categories: Individual targeted, community targeted and None.

MuRIL

MuRIL model pre-trained on 17 languages, 16 Indian languages and English []. Indian languages are: Assamese, Gujarati, Kashmiri, Bengali, Hindi, Kannada, Malayalam, Nepali, Marathi, Punjabi, Oriya, Sanskrit, Tamil, Sindhi, Telugu and Urdu. This model based on the BERT base architecture. In training, translation and transliteration segments are included. The MuRIL model is pre-trained on monolingual segments as well as parallel segments. There are two types of parallel data: translated data and transliterated data. The model is trained on 10 epochs and the learning rate is set to be 3e-5 and batch size is 32. The Hindi tweets are classified into three categories.

5. Experimental Results

The different deep-learning and transformer-based models are implemented to classify the tweets as an individual targeted, community targeted, and none. The deep learning models (CNN, LSTM, Bi-LSTM) and transformer-based models (mBERT, MuRiL, IndicBERT, XLM-RoBERTa) are implemented. Table 5 shows the performance matrices of all implemented deep learning and transformer models before and after oversampling the data. The deep learning-based models (CNN, LSTM, and Bi-LSTM) before oversampling, reported the weighted average F1-score of 0.72, 0.79 and 0.80 respectively. Similarly, after oversampling the data using SMOTE, the resulting weighted F1 score of CNN is 0.73, LSTM is 0.81 and Bi-LSTM is 0.82. The transformer-based models (mBERT, MuRiL, IndicBERT, and XLM-RoBERTa), before oversampling, reported the weighted average F1-score of 0.82, 0.83, 0.85, and 0.89, respectively. After using SMOTE, the weighted average F1-score for transformer-based models (mBERT, MuRiL, IndicBERT, and XLM-RoBERTa) are found to be 0.83, 0.84, 0.86, and 0.91, respectively. As we observed that the models performed better after using SMOTE and give improved results. The XLM-RoBERTa outperforms the other implemented models in both cases (before and after oversampling). The reported macro average F1 score is 0.75 and the weighted average F1 score is 0.89 before oversampling the data. After using SMOTE, the reported macro average f1-score is 0.77 and the weighted average F1-score is 0.91.

Table 5
Classification report for implemented models to classify the tweets in the Hindi language

Models	**Individual Target**			**Community Target**			**None**			**Macro Average F1-score**	**Weighted Average F1-score**
	P	**R**	**F1-score**	**P**	**R**	**F1-score**	**P**	**R**	**F1-score**		
CNN (BS)	0.71	0.10	0.18	1.00	0.27	0.42	0.78	0.99	0.88	0.49	0.72
CNN (AS)	0.73	0.13	0.22	0.70	0.31	0.42	0.82	0.95	0.88	0.50	0.73
LSTM (BS)	0.75	0.17	0.27	0.56	0.38	0.45	0.84	0.96	0.89	0.54	0.79
LSTM(AS)	0.80	0.22	0.35	0.42	0.58	0.49	0.90	0.93	0.91	0.58	0.81
Bi-LSTM (BS)	0.65	0.35	0.45	0.86	0.32	0.46	0.84	0.96	0.89	0.60	0.80
Bi-LSTM (AS)	0.42	0.28	0.33	0.54	0.54	0.54	0.90	0.94	0.92	0.60	0.82
mBERT (BS)	0.40	0.33	0.36	0.59	0.54	0.57	0.89	0.92	0.90	0.61	0.82
mBERT (AS)	0.76	0.44	0.55	0.81	0.35	0.48	0.90	0.93	0.91	0.64	0.83
MuRiL (BS)	0.73	0.45	0.56	0.78	0.37	0.50	0.86	0.96	0.91	0.65	0.83
MuRiL (AS)	0.74	0.47	0.57	0.80	0.39	0.52	0.87	0.94	0.90	0.66	0.84
IndicBERT (BS)	0.76	0.48	0.58	0.81	0.38	0.51	0.89	0.96	0.92	0.67	0.85
IndicBERT (AS)	0.78	0.49	0.59	0.89	0.37	0.52	0.93	0.94	0.93	0.68	0.86
XLM-RoBERTa (BS)	0.51	0.42	0.46	0.93	0.95	0.94	0.85	0.82	0.84	0.75	0.89
XLM-RoBERTa (AS)	0.55	0.41	0.46	0.98	0.93	0.95	0.89	0.95	0.91	**0.77**	**0.91**

***** AS = After Oversampling, BS = Before Oversampling

6. Conclusion

This paper presents a curated and annotated corpus of Hindi tweets for target-based hate speech detection. The dataset comprises of 2,020 tweets labelled in three categories. This is the first dataset of this kind in the Hindi language in which the type of content is characterized based on target of the hate speech, i.e., either the tweet directly targets an individual (individual targeted hate speech) or the tweet targets a particular community, group, or organization (community targeted hate speech). Those tweets that do not show any hate are labelled as none. The dataset classified the tweets based on the target to make the hate speech detection task more accurate and specific. The high value of inter annotator agreement shows that its high quality annotated dataset. In order the assess the usefulness of the dataset for hate speech research, several deep learning and transformer-based models were applied for the classification of tweets in the three categories. The best results are obtained with XLM-RoBERTa., with a macro averaged F1 score and weighted average F1 score of 0.77 and 0.91 respectively. The experimental results show that the dataset is suitable for hate speech detection research in Hindi language.

Declarations

The authors declare that the manuscript complies with the ethical standards of the journal and there no financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

Author Contribution

Deepawali Sharma participated in design of the study and carrying out the experimental work and writing of the paper.

Vivek Kumar Singh participated in conceptualizing the study, guided the annotation process and wrote and reviewed the article.

Vedika Gupta participated in guidance of the experimental work and review of the paper.

References

1. Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230.
2. Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88–93).

3. Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study.*International Journal of Advanced Computer Science and Applications*, *11*(8).
4. Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. Z. (2020). A framework for hate speech detection using deep convolutional neural network. *Ieee Access : Practical Innovations, Open Solutions*, *8*, 204951–204962.
5. Koufakou, A., Pamungkas, E. W., Basile, V., & Patti, V. (2020). HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the fourth workshop on online abuse and harms* (pp. 34–43). Association for Computational Linguistics.
6. Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2020). Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.
7. Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (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* (pp. 16–27). Springer Berlin Heidelberg.
8. Vargas, F., de Góes, F. R., Carvalho, I., Benevenuto, F., & Pardo, T. (2021, September). Contextual-lexicon approach for abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 1438–1447).
9. Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021, May). Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 17, pp. 14867–14875).
10. Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2021). ETHOS: A multi-label hate speech detection dataset.Complex Intell, Syst. 8, 4663–4678 (2022).
11. Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019, November). A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4755–4764).
12. Bhardwaj, M., Akhtar, M. S., Ekbal, A., Das, A., & Chakraborty, T. (2020). Hostility detection dataset in Hindi. *arXiv preprint arXiv:2011.03588*.
13. Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June). A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media* (pp. 36–41).
14. Satapara, S., Modha, S., Mandl, T., Madhu, H., & Majumder, P. (2021). *Overview of the hasoc subtrack at fire 2021: Conversational hate speech detection in code-mixed language*. Working Notes of FIRE.
15. Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019, December). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation* (pp. 14–17).
16. Ravikiran, M., & Annamalai, S. (2021, April). DOSA: Dravidian code-mixed offensive span identification dataset. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages* (pp. 10–17).
17. Ludwig, F., Dolos, K., Zesch, T., & Hobley, E. (2022, July). Improving Generalization of Hate Speech Detection Systems to Novel Target Groups via Domain Adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (pp. 29–39).
18. Mehta, H., & Passi, K. (2022). Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI). *Algorithms*, *15*(8), 291.
19. Dascălu, Ș., & Hristea, F. (2022). Towards a Benchmarking System for Comparing Automatic Hate Speech Detection with an Intelligent Baseline Proposal. *Mathematics*, *10*(6), 945.
20. Bhatnagar, V., Kumar, P., & Bhattacharyya, P. (2022). Investigating Hostile Post Detection in Hindi. *Neurocomputing*, *474*, 60–81.
21. Sai, S., Jacob, A. W., Kalra, S., & Sharma, Y. (2021). Stacked embeddings and multiple fine-tuned XLM-roBERTa models for enhanced hostility identification. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1* (pp. 224–235). Springer International Publishing.
22. Bhattacharyya, P. (2021, April). Divide and Conquer: An Ensemble Approach for Hostile Post Detection in Hindi. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers* (Vol. 1402, p. 244). Springer Nature.
23. Bhardwaj, M., Chakraborty, T., & Akhtar, M. (2021). *HostileNet: multi-label hostile post detection in Hindi* (Doctoral dissertation, IIIT-Delhi).
24. Santosh, T. Y. S. S., & Aravind, K. V. S. (2019, January). Hate speech detection in hindi-english code-mixed social media text. In *Proceedings of the ACM India joint international conference on data science and management of data* (pp. 310–313).
25. Sreelakshmi, K., Premjith, B., & Soman, K. P. (2020). Detection of hate speech text in Hindi-English code-mixed data. *Procedia Computer Science*, *171*, 737–744.
26. Rani, P., Suryawanshi, S., Goswami, K., Chakravarthi, B. R., Fransen, T., & McCrae, J. P. (2020, May). A comparative study of different state-of-the-art hate speech detection methods in Hindi-English code-mixed data. In *Proceedings of the second workshop on trolling, aggression and cyberbullying* (pp. 42–48).
27. Farooqi, Z. M., Ghosh, S., & Shah, R. R. (2021). Leveraging Transformers for Hate Speech Detection in Conversational Code-Mixed Tweets. *arXiv preprint arXiv:2112.09986*.
28. Mundra, S., Singh, N., & Mittal, N. (2021). Fine-tune BERT to Classify Hate Speech in Hindi English Code-Mixed Text. In *Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org*.

29. Bölücü, N., & Canbay, P. (2021). Hate speech and offensive content identification with graph convolutional networks. In *Forum for information retrieval evaluation (working notes)(FIRE), CEUR-WS. org*.

30. Anusha, M. D., & Shashirekha, H. L. (2020). An Ensemble Model for Hate Speech and Offensive Content Identification in Indo-European Languages. In *FIRE (Working Notes)* (pp. 253–259).

31. Kumari, K., & Singh, J. P. (2020, December). AI_ML_NIT_Patna@ HASOC 2020: BERT Models for Hate Speech Identification in Indo-European Languages. In *FIRE (Working Notes)* (pp. 319–324).

32. Mishra, A. K., Saumya, S., & Kumar, A. (2020). IIIT_DWD@ HASOC 2020: Identifying offensive content in Indo-European languages. In *FIRE (Working Notes)* (pp. 139–144).

33. Mohtaj, S., Woloszyn, V., & Möller, S. (2020). TUB at HASOC 2020: Character based LSTM for Hate Speech Detection in Indo-European Languages. In *FIRE (Working Notes)* (pp. 298–303).

34. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

35. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

36. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

37. Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88–93).

38. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media* (Vol. 11, No. 1, pp. 512–515).

39. Kwok, I., & Wang, Y. (2013, June). Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 27, No. 1, pp. 1621–1622).

40. Chakravarthi, B. R., Kumaresan, P. K., Sakuntharaj, R., Madasamy, A. K., Thavareesan,S., Navaneethakrishnan, S. C., … Mandl, T. (2021). Overview of the HASOC-DravidianCodeMix shared task on offensive language detection in Tamil and Malayalam. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation*. CEUR.

Figures

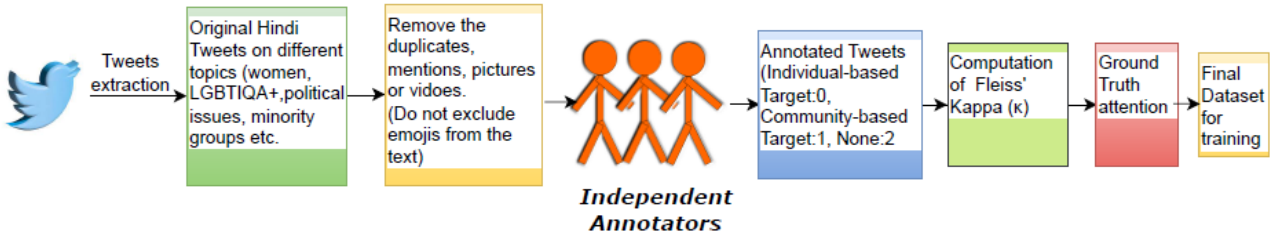
Tweets extraction
Original Hindi Tweets on different topics (women, LGBTIQA+,political issues, minority groups etc.
Remove the duplicates, mentions, pictures or vidoes. (Do not exclude emojis from the text)
Independent Annotators
Annotated Tweets (Individual-based Target:0, Community-based Target:1, None:2
Computation of Fleiss' Kappa (κ)
Ground Truth attention
Final Dataset for training

Figure 1

Illustration of steps for dataset creation

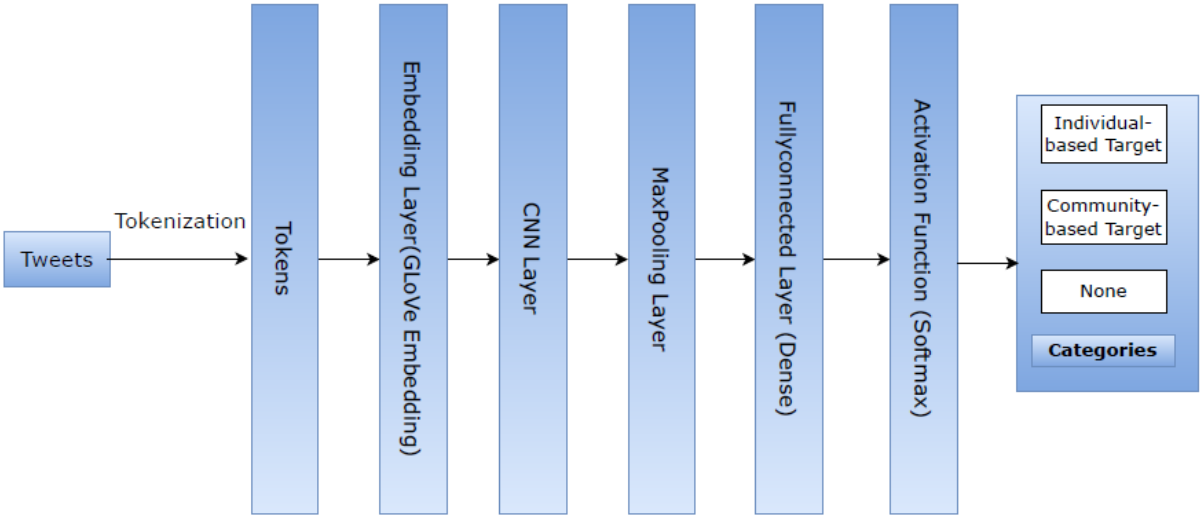

Figure 2

Implementation of CNN to detect target-based hate speech

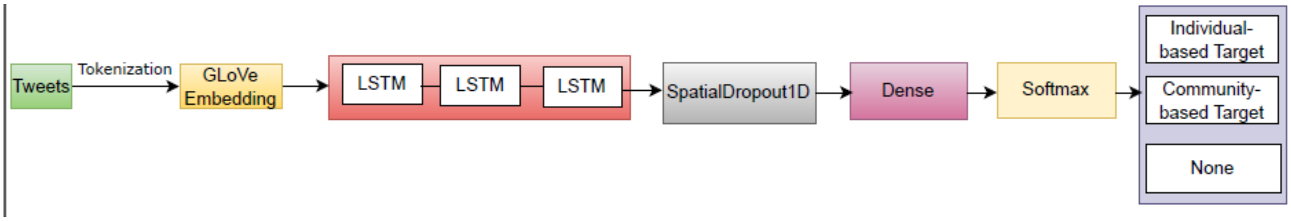

Figure 3

Block diagram of LSTM to classify the Hindi tweets

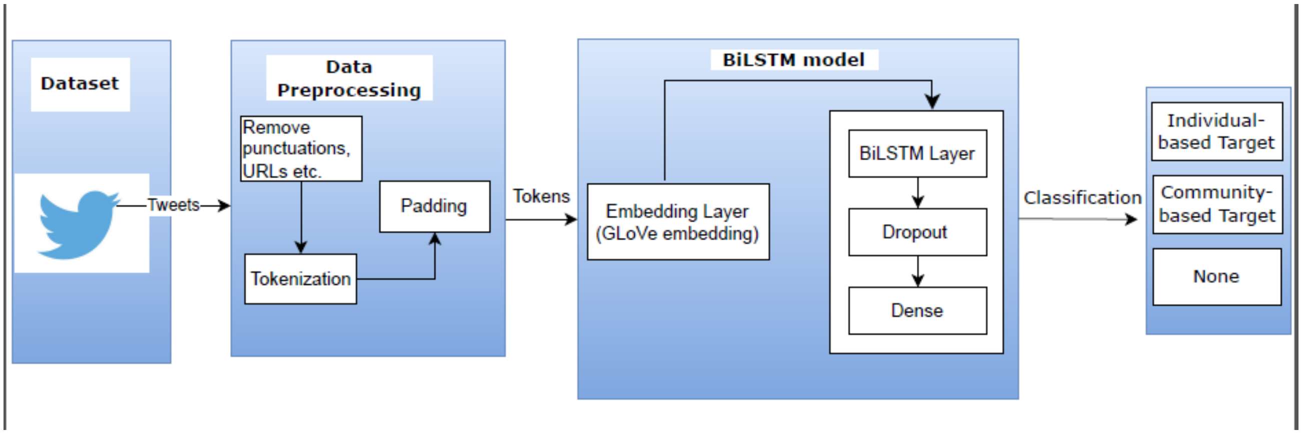

Figure 4

Implementation of Bi-LSTM to detect target-based hate speech

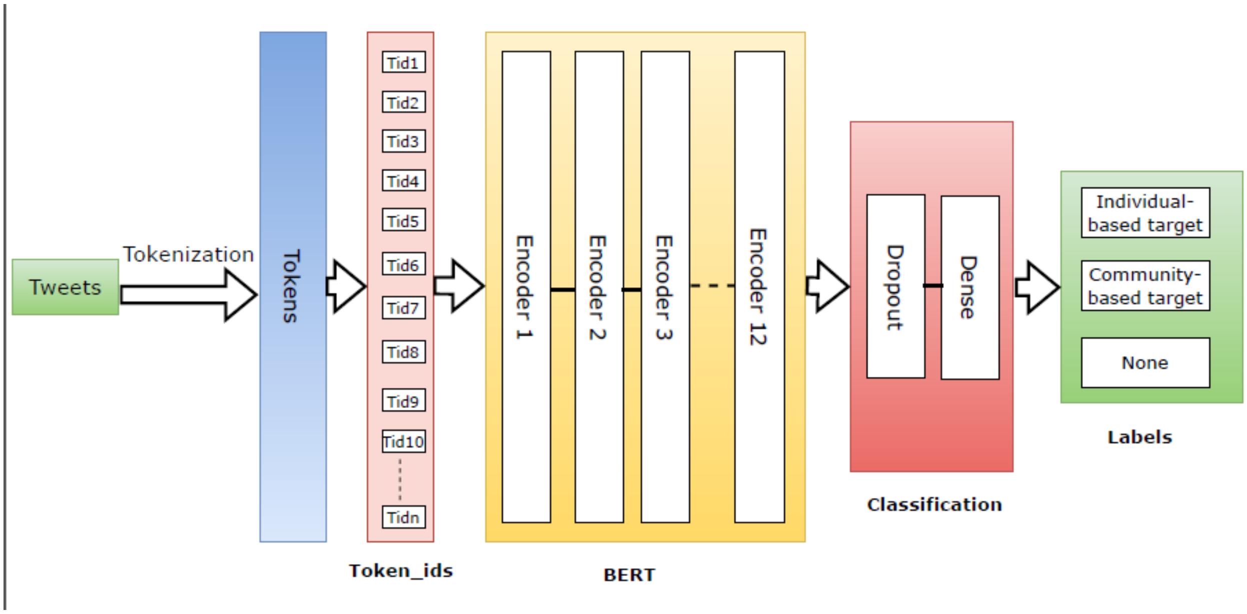

Figure 5

mBERT approach to classify the Hindi tweets