Corpus Building for Hate Speech Detection of Gujarati Language

Abhilasha Vadesara and Purna Tanna(✉)

GLS University, Ahmedabad, Gujarat, India
abhilashavadesara@gmail.com

Abstract. Social media is a rapidly expanding platform where users share their thoughts and feelings about various issues as well as their opinions. However, this has also led to a number of issues, such as the dissemination and sharing of hate speech messages. Hence, there is a need to automatically identify speech that uses hateful language. Hate speech refers to the aggressive, offensive language that focuses on a specific people or group as far as their ethnic group or race (i.e., racism), gender (i.e., sexism), beliefs, and religion. The aim of this paper is to examine how hate speech contrasts with non-hate speech. A corpus of Gujarati tweets has been collected from Twitter. The dataset was cleaned and pre-processed by removing unnecessary symbols, URLs, characters, and stop words, and the cleaned text was analyzed. Pre-processed data was annotated by twenty-five people and has achieved Fleiss's Kappa coefficient with 0.87 accuracies for agreement between the annotators.

Keywords: Hate speech · Text mining · Kappa's coefficient · Gujarati language · Sentiment analysis

1 Introduction

Expressions that are harassing, abusive, harmful, urge brutality, make hatred or discrimination against groups, target qualities like religion, race, a spot of beginning, race or community, district, individual convictions, or sexual direction are called hate speech. The Ability to spot hate speech has gotten heaps of attention these days. As a result, hate speech has reached new levels in additional advanced and intellectual types.

Social networking sites make it more direct. To provide honest thoughts and feelings to end-users, Twitter provides a site and microblogging service. In this digital age, social media data is increasing daily, where hate speech detection becomes a challenge in provoking conflict among the countries' voters. However, it's impossible to spot hate speech from a sentence without knowing the context.

As we have seen, much research has been accomplished on European, English, and some Indian languages. In any case, little work has been done in Gujarati as it is the primary language most Gujarati people use in speaking and formulating. The purpose of the analysis is to build the corpus of Gujarati language instead of distinctive hate speech.

K. K. Patel et al. (Eds.): icSoftComp 2022, CCIS 1788, pp. 382–395, 2023.
https://doi.org/10.1007/978-3-031-27609-5_30

After collecting tweets, we pre-processed them with the Natural Language Processing technique [21] and implemented annotation by twenty-five different age groups. To check the inter agreement between annotators, we use Fleiss's kappa. In addition, the range of individuals and their backgrounds, cultures, and beliefs will ignite the flames of hate speech [1]. For the Gujarati region, there's a conspicuous magnification within the utilization of social media platforms.

This paper is structured in different sections. In Sect. 2, we describe the short description of related work. The new dataset and the Methodology, which include the data cleaning and Methodology, are defined in Sects. 3 and 4, respectively. In Sect. 5 we discussed the experiments. Section 6 describes the Result and Discussion of the technique. Section 7 finalizes this paper and suggests possible suggestions for future work (Fig. 1).

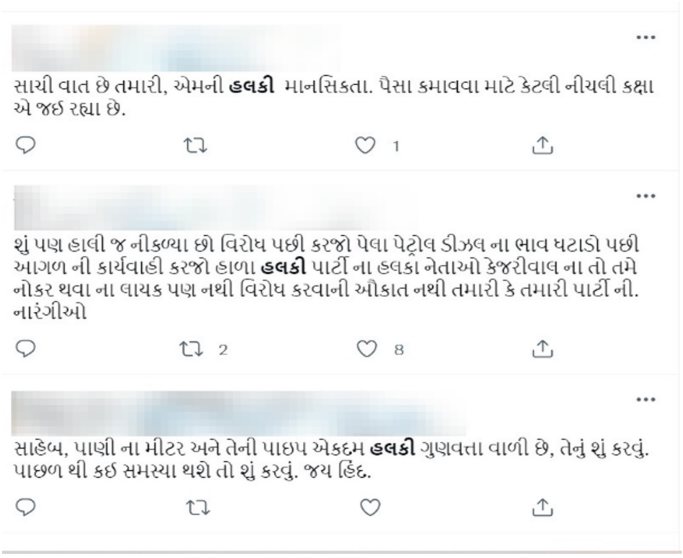

Fig. 1. Post of Twitter tweet

2 Related Forum and Dataset

Collections are an essential quality for any classification method. Several corpora of hate speech were used for analysis. Tremendous work has been done in numerous dialects, particularly for European and Indian. However, standard datasets aren't available for some languages, like Gujarati, and we are trying to make the tagged dataset for such an occasional resource language. Several corpora focus on targets like immigrants, women or racism, religion, politics, celebrities, and community. Others focus on only Hate speech detection or different offensive text types. A recent trend is to classify the data into more fine-grained classification. So, some knowledge challenges need detailed analysis for hate speech like the detection of target, aggressiveness, offensive, stereotype,

irony, etc. A recent and attention-grabbing diversity is CONAN. It offers Hate Speech and also the reactions to it [2]. It opens opportunities for detecting Hate Speech by analyzing it collectively with consecutive posts. The researcher summarizes the standard Hate speech dataset attainable at various forums. Karim, Md. Rezaul et al. proposed DeepHateExplainer, which detects different sorts of hate speech with an 88% f1-score on several ML and DNN classifiers. For annotation of the dataset, they used the cohesion kappa technique [13]. Alotaibi, B. Alotaibi et al. have provided an approach for detecting aggression and hate speech through short texts. They have used three models: the bidirectional gated recurrent unit (BiGRU), the second transformer block, and the third convolutional neural network (CNN) based on Multichannel Deep Learning. They used the NLP approach and categorized 55,788 datasets into Offensive and Non-Offensive. They achieved 87.99% accuracy upon evaluating it on trained data 75% and testing data 25% [18]. Karim, Md Rezaul et al. proposed hate speech detection for the under-resourced Bengali language. They Evaluate against DNN baselines and yield F1 scores of 84%. They applied approaches to accurately identifying hateful statements from memes and texts [15]. Gohil and Patel generated G-SWN using Hindi SentiWordNet (H-SWN) and IndoWordNet (IWN) by manipulating synonym relations. The Corpus was annotated for negative and positive polarity classes by two annotators. They used Cohen's kappa Statistical measure for inter-annotator agreement between annotators [16]. The GermEval Task2 2019 is the data set of German language that tagged 4000 Twitter tweets to identify the three levels, hate, type, and implicit/explicit, with the macro F1 0.76 [3]. The racism dataset was used to determine binary and racism on 24000 English tweets with the accuracy 0.72 F1 Score [4]. Arabic social media dataset is on the market to identify Arabic tweets where it focuses on identifying obscene and inappropriate data with the f1 score of 0.60 [5]. Table 1 contains the dataset that is offered. Al-Twairesh, Nora et al. [22] presented the collection and construction of the Arabic dataset. They explained the technique of annotation of 17,573 twitter datasets. For inter agreement, they used Fleiss's Kappa and achieved 0.60 kappa's value, considered moderate. Akhtar, Basile et al. [23] Tested three different Twitter social media datasets in English and Italian language. They annotated the dataset with three annotators and measured Fleiss's kappa value of 0.58. They combined the single classifiers into an inclusive model.

Table 1. Collections of research on hate speech.

Paper reference	Dataset	Task	Example	Font size and style
[6]	Twitter	Binary, Hate	14500	English
[7]	Twitter	Hate, aggression, target	19000	Spain, English
[8]	Twitter	3 levels, Hate, targeted and target type	13200	English
[9]	TRAC COLING	3 classes, overtly or covertly Aggressive	15000 each language	English, Hindi
[17]	Facebook	6 classes	5,126	Bengali
[16]	Twitter	2 classes	1120	Gujarati

3 Dataset and Collection

Our main goal was to collect datasets using different techniques. The tweet gathered in the period from January 2020 to January 2021. We gathered the tweets data using the Twitter API with different categories like politics, sports, religion, and celebrity, as shown in Fig. 2. Most of the substance on Twitter isn't offensive, so we attempted different techniques to keep the dissemination of offensive tweets on about 30% of the dataset. Keywords and hashtags used to identify Gujarati hate speech. The Twitter API gives numerous recent tweets with an unprejudiced dataset. Thus, the tweets are acquired with the help of keywords and hashtags containing offensive content. The difficulties during the assessment of hate speech were language registers like irony or indirectness and youth talk, which researchers might not understand. We have collected approximately Twelve thousand tweets on hate and none-hate Gujarati content. The corpus was separated into training and testing categories to perform the classification task (Table 2).

Table 2. Collections of research on hate speech

Categories	Different Target	Example of Gujarati tweet	Translation in English
Religion	Religious people	આ મુ**ઓ ભ**ઓ પોતે દેશ વિરોધી કામ કરે છે.	These mul***s themselves do anti-national work.
Sports	Sports people, Sports	ક્રિકેટ તો માત્ર સટ્ટા વાળ માટે છે.	Cricket is only for betting hair.
Politics	Political party, People	શિ**નાએ કહ્યું, દેશમાં ઘૂસેલા પા***ની, બાં***શી મુસ્લિમોને બહાર ફેંકી દેવા જોઈએ.	The Shi***na said Pa**ni, Ban**shi Muslims who entered the country should be thrown out.
Celebrity	People, Movie, Celebrity	રા* *વંત એવી અટકચાળી અભિનેત્રી જેને કોઇને કોઇ વાતે સોશિયલ મીડિયા પર ચર્ચામાં રહેવાનો શોખ છે.	Ra** **want is a speculation actress who is fond of being in the news on social media for some means.

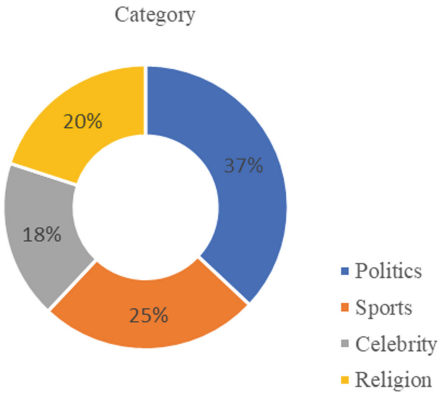

Fig. 2. Distribution diagram of hate speech data in each category

4 Methodology

In this section, we discussed the proposed approach in detail, discussion of preprocessing techniques and its example in the Gujarati language, and the annotation task and its process in detail.

4.1 Data Preprocessing

The dataset from Twitter is very noisy because it is not processed. To extract the model into a better feature, we need to perform the text processing on the actual dataset [21]. Initially, our data was in UTF-8 because of Twitter API responds to it in encoded form. We converted the data into Gujarati with the help of python decode method. Although the data was not clean, it contained many extra characters and Hindi, and English alphabets mixed, so it was necessary to clean it. With the help of python, we implement the pre-processing task which contain removal of URL, hashtags, user mentions, punctuation, Numbers, stop words, tokenizing etc. with the help of python libraries like pandas, re (regular expression), nltk. Below we describe each steps in detail.

Removal of URLs, Hashtags, User Mentions, Other Characters and Noise. Undesirable strings are considered extra information, which creates noise in the data. The tweet contains much extra information, such as URLs (http://www.imrobo.com) which refers to extra information, hashtags symbol (#socialmedia), which denotes the tweet is associated with some particular topic. User mentions (@name) means the post links to a particular user's profile. This information is helpful to human beings, but for a machine, it is considered noise that requires to be handled. Many researchers have presented different techniques to deal with such content [13, 23].

An example is given below:

Before: આગામી છ મહિના સુધી ક્રિકેટ અથવા અન્યકોઈ રમત શક્ય નથી. \nhttps://t.co/DHGGnlLGOi'b' @dhwansdave #cricket
After: આગામી છ મહિના સુધી ક્રિકેટ અથવા અન્ય કોઈ રમત શક્ય નથી.

Removal of Emoticons. Social media users use emojis such as ☹, 😊, 😉, etc., to express their sentiments. Such content is not helpful for tweet classification, so it needs to be removed from the tweet. An example is given below:

Before: અસ્મિતા પર હમાણા તમારો ઇંટરવ્યું જોયો. મજા પડી ગઇ તમને સાંભળવાની😍
After: અસ્મિતા પર હમાણા તમારો ઇંટરવ્યું જોયો .મજા પડી ગઇ તમને સાંભળવાની

Removal of Numbers. The Dataset ordinarily contains undesirable numbers and provides essential information, but they don't provide the information that helps in classification. So many researchers altogether remove it from the corpus. However, Eliminating the number from the Dataset may lead to a loss of information, but it does not impact much on the classification task. So, we eliminate all the numbers from the Dataset. An example is given below:

Before: હિન્દુ ધર્મ ના લોકો એ ફરજીયાત મિસકોલ કરવો88662 88662 પર અને બીજા 10 હિન્દૂ ને આ મેસેજ વિડિઓ સાથે મોકલવો
After: હિન્દુ ધર્મ ના લોકો એ ફરજીયાત મિસકોલ કરવો પર અને બીજા હિન્દૂ ને આ મેસેજ વિડિઓ સાથે મોકલવો

Removal of Stop Words. The tweet contains common words like ' તો,' ' અને,' ' છે,' 'આવે,' ' થતા' etc. are known as stop words in the Gujarati language [19]. It doesn't have complete, meaningful information, which helps in classification. One of the significant advantages of eliminating stop words in NLP text-based handling is decreasing the text by 30–40% from the corpus [20]. In our analysis, we created the list of stop words and eliminated them from the corpus. An example is given below:

Before: આખી દુનિયા ભગવાન ની મરજી થી ચાલે છે
After: આખી દુનિયા ભગવાન મરજી ચાલે

Tokenizing. In this step, tweets are separated using spaces to find the boundaries of words. Splitting a sentence into meaningful parts and recognizing the individual entities in the sentence is called Tokenization. We implement word Tokenization for classification tasks after the annotation of data. An example is given below:

Before: આખી દુનિયા ભગવાન મરજી ચાલે
After: આખી, દુનિયા, ભગવાન, મરજી, ચાલે

4.2 Data Annotation

After collecting the data, the second stage consists of annotating the Gujarati corpus. Before building the corpus, a review of techniques used to detect hate speech [12]. However, we eliminated many tweets from the corpus because of data duplication and off-topic content. At present, the amount of annotated data consists of 10000 tweets. Hate speech is a complex and multi-level concept. The annotation task is tricky and subjective, so we have taken all the initial steps to ensure that all annotators have a general basic

knowledge about the task starting with the definition. The annotation process includes a multi-step process. After a fundamental step, it was carried out by 25 annotators manually who the people of different age groups are shown in Fig. 3.

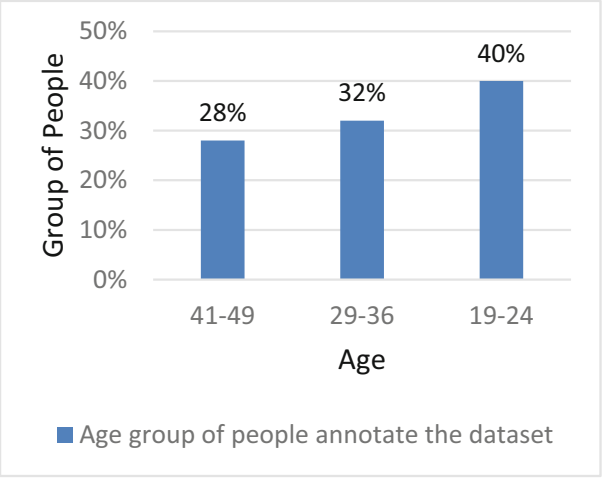

Fig. 3. Annotation different age group people wise

They labeled the corpus based on the relevant definitions, rules, regulations, and examples. The annotators were given the instructions as guidelines to classify the tweet into hate and none-hate category. The following factors are considered for hate tweets.

The first factor considered as a target means that the tweet should address, or refer to, one of the minority groups previously known as hate speech targets or the person considered for membership in that category.

The second is action, or more explicitly pronounced offensive force, in which it is capable of spreading, inciting, promoting, or justifying violence against a target. At whatever point the two factors happen in a similar tweet, we consider it a hate speech case, as in the example below (Table 3):

Table 3. Example of hate and none-hate annotation in tweets

Tweets	Label
કેટલાક મૂર્ખ કહે છે કે મુસ્લિમોને દેશમાં ડરાવવામાં આવે છે.	none-hate
તુ પાકિસ્તાન વયો જા હરામી	hate
આવું જ્ઞાન સરકાર ના નેતા ઓ ને પણ આપો . સાલા બેસર્મો ના લીધે જ આજે આ દિવસો પાછા આવ્યા છે.	hate
સાલા સરકારી કર્મચારી ને મફત નો પગાર જોઈ છે.	hate
હવે શિક્ષકોને શાળાના પુસ્તકો વિદ્યાર્થી પાસે પહોંચાડવાનું પણ ભાડુ જોઇએ છે !	none -hate

Figure 4 illustrate the procedure for the annotation of the corpus. First step is to check the tweets is belong to which category ex. religion, political, ethnicity etc. Then it should be analyzed by few questions Like “Is there any intention to offend someone?” If the answer will be no than it would be considered as none hate. Because that tweet considers normal tweet ex. જન્માષ્ટમી ના દીવસે શ્રદ્ધાળુઓ ભાવુક બન્યા. (Devotees became emotional on the day of Janmashtami.) which doesn’t contain any offend towards any religious or person. But if it is yes than the next question would be asked like “Is there any swearing word or expression?” if the answer is yes than it would be consider as hate speech because the swearing word can be used harm the feelings of particular person or religion or group. ex. મુ** સાલા કસાઈ છે(The mullahs are butchers). If it is no means we required to analyzed it in depth like next question will be “Is the post contains any target or any action?” If the answer is yes then the tweets consider as hate ex. તું એક વાર મારા હાથમાં આવ હું તને મારી નાખીશ(you once come into my hands I will kill you) such type of tweets contains some action towards person so it’s considered as hate speech. Otherwise, it is non-hate ex. તું એક વાર મારા હાથમાં આવ(you once come into my hands).

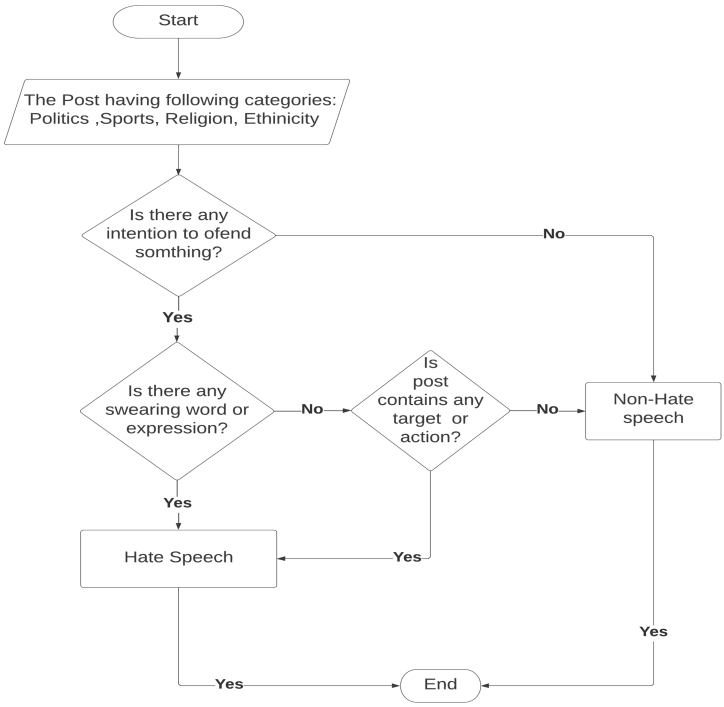

Fig. 4. Process of annotation for hate speech detection

5 Experiments

The 12k Gujarati dataset was raw and mixed with punctuations, URLs, non-Gujarati characters, emoticons, special characters, and stop words and tokenized after the annotation task. We removed the punctuations, stopwords, emoticons, URLs, symbols, non-Gujarati

Table 4. Steps of data cleaning using preprocessing technique

Number	Preprocessing Techniques	Raw data	Clean Data
1.	Removal of URLs, hashtags, user mentions, other characters and noise	@sandeshnews અમદાવાદ મુંબઈ નેશનલ હાઇવે ઉપર અંકલેશ્વર નજીક બે ટ્રક વચ્ચે સર્જાયો ગંભીર અકસ્માત, 2 લોકોના ઘટનાસ્થળે મોત. ☹ #accident #ankleshwar #highway #truckaccident	અમદાવાદ મુંબઈ નેશનલ હાઇવે ઉપર અંકલેશ્વર નજીક બે ટ્રક વચ્ચે સર્જાયો ગંભીર અકસ્માત, 2 લોકોના ઘટનાસ્થળે મોત. ☹
2.	Removal of Emoticons	અમદાવાદ મુંબઈ નેશનલ હાઇવે ઉપર અંકલેશ્વર નજીક બે ટ્રક વચ્ચે સર્જાયો ગંભીર અકસ્માત, 2 લોકોના ઘટનાસ્થળે મોત. ☹	અમદાવાદ મુંબઈ નેશનલ હાઇવે ઉપર અંકલેશ્વર નજીક બે ટ્રક વચ્ચે સર્જાયો ગંભીર અકસ્માત, 2 લોકોના ઘટનાસ્થળે મોત.
3.	Removal of punctuations	અમદાવાદ મુંબઈ નેશનલ હાઇવે ઉપર અંકલેશ્વર નજીક બે ટ્રક વચ્ચે સર્જાયો ગંભીર અકસ્માત,2 લોકોના ઘટનાસ્થળે મોત.	અમદાવાદ મુંબઈ નેશનલ હાઇવે ઉપર અંકલેશ્વર નજીક બે ટ્રક વચ્ચે સર્જાયો ગંભીર અકસ્માત 2 લોકોના ઘટનાસ્થળે મોત
4.	Removal of number	અમદાવાદ મુંબઈ નેશનલ હાઇવે ઉપર અંકલેશ્વર નજીક બે ટ્રક વચ્ચે સર્જાયો ગંભીર અકસ્માત 2 લોકોના ઘટનાસ્થળે મોત	અમદાવાદ મુંબઈ નેશનલ હાઇવે ઉપર અંકલેશ્વર નજીક ટ્રક વચ્ચે સર્જાયો ગંભીર અકસ્માત લોકોના ઘટનાસ્થળે મોત
5.	Removal of Stop-words	અમદાવાદ મુંબઈ નેશનલ હાઇવે ઉપર અંકલેશ્વર નજીક ટ્રક વચ્ચે સર્જાયો ગંભીર અકસ્માત લોકોના ઘટનાસ્થળે મોત	અમદાવાદ મુંબઈ નેશનલ હાઇવે અંકલેશ્વર ટ્રક સર્જાયો ગંભીર અકસ્માત લોકોના ઘટનાસ્થળે મોત
6.	Tokenizing	અમદાવાદ મુંબઈ નેશનલ હાઇવે અંકલેશ્વર ટ્રક સર્જાયો ગંભીર અકસ્માત લોકોના ઘટનાસ્થળે મોત	અમદાવાદ, મુંબઈ, નેશનલ, હાઇવે અંકલેશ્વર, ટ્રક, સર્જાયો, ગંભીર અકસ્માત, લોકોના, ઘટનાસ્થળે, મોત

characters, and tokenization to increase the accuracy of the classification model. Now the dataset is entirely ready to train the model. Table 4 shows the step-by-step process of data cleaning using preprocessing technique.

After the preprocessing task the data we have annotated by 25 different age group people. The training data was hand-coded and manually annotated and admits the potential for hard-to-trace bias within the hate speech categorization [3]. To prove the reliability between annotators we adopt some measures. In addition to the annotation rules, the Kappa call agreement based on the Cohen's letter data points that estimate the data constant between $0 \leq \kappa \leq 1$ is additionally used for the two annotators [11, 21]. For measuring IAA between more than two annotator we used Fleiss's Kappa [26, 27]. Fleiss's Kappa were implemented on ten thousand tweets annotated by twenty-five annotators with classes hate and non-hate. For implementing in python, the algorithm requires the numeric values for that value of non-hate and hate considered as 1 and 0. The kappa's score was measured 0.86. There is no such guideline to assess the value of kappa 0.86 is (i.e., measure the level of agreement between annotators). The Cohen's kappa has been suggested to measure how strong level of agreement annotator have. Table 5 illustrate the lowest value of kappa is between 0 to 20 which is considered as none level of agreement where the value between above 90 considered as almost perfect agreement between annotator. Between 0 to 90 the ranges like 21 to 39, 40 to 59, 60 to 79,80 to 90 are minimal, weak, moderate and strong level of agreement respectively. According to the Table 5 our Fleiss's kappa value is 0.87 which is considered almost perfect agreement between annotator [27].

Table 5. Cohesion Kappa's level of agreement [11]

Value of Kappa	Level of agreement	% of data that are reliable
0–.20	None	0–4%
.21–.39	Minimal	4–15%
.40–.59	Weak	15–35%
.60–.79	Moderate	35–63%
.80–.90	Strong	64–81%
Above.90	Almost Perfect	82–100%

After annotation task, we found 69.3% of all tweets have been considered as hate speech, whereas 30.7% of tweets are none hate across the whole corpus as mentioned in Fig. 5.

As per Table 4, we get the 6930 tweets which belongs to hate speech and 3070 none hate speech among the whole corpus. To implement the classification task, we will be keeping the 80–20 ratio of whole corpus for train and test the model.

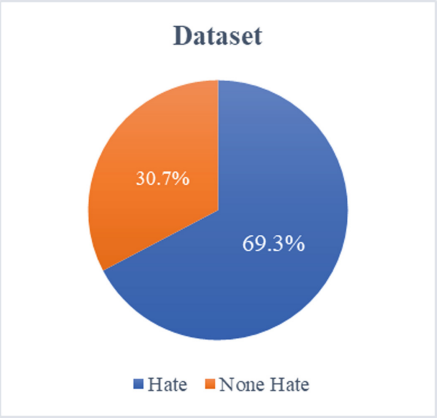

Fig. 5. Distribution of Gujarati hate and none hate tweets in the dataset

Table 6. Total no. of hate and none hate tweets from dataset

Numerical representation	Class	Total instance
0	Hate	6930
1	Non-hate	3070
Total		10000

6 Result and Discussion

Hate speech identification is not just a simple word identification task; it has some complexity, particularly in Gujarati text. As tweets are short texts with a decent number of characters, the capacity to distinguish the hashtags helps enormously in recognizing the subject of the tweet. The collection of tweets was in different categories like politics, sports, religion, and celebrity. We collected the majority of tweets in the politics category, where we get the highest hate data. The initial data was mixed and not clean, so to increase the performance of the dataset, we cleaned the data using preprocessing technique. As discussed in the previous section, extraction and cleaning the tweet is pretty challenging. But We can observe based on Table 4 that the goal was achieved using the preprocessing technique. It shows how data became clean using the different preprocessing techniques. The preprocessing technique executed with the NLTK library is the one that is broadly utilized for preprocessing the other languages texts like Hindi, English, Arabic, etc. We used the RE library for cleaning the Gujarati data, which was quite helpful for the task. Here, we compared the results of preprocessing technique in Waseem et al. and Davidson et al. datasets. We observed on the dataset of Waseem et al. that the performance of the SVM (trigram), CNN, and LR (bigrams) classifier increased using the pre-pressing technique [29]. Because of the username, hashtags, and URLs, the performance of classifiers doesn't increase accuracy in the dataset of Davidson et al. [28]. Therefore, removing URLs, hashtags, and user mentions is required. Removal

of punctuation gave the significant performance of the dataset of Davidson et al. The number is not required for the detection of hate speech. In terms of the result of the LSTM classifier, it achieved a good score in Waseem et al. Removal of stop words is the general baseline approach that increases the performance of all the datasets. After the implementation of preprocessing technique, we got the data for annotation. The 25 annotators annotated the whole corpus manually based on the given guideline. We used Fleiss kappa to check their inter agreement and achieved the 0.87 value of k. According to Cohesion Kappa's measure, it is considered a perfect agreement between annotators. After the implementation of the annotation task, we had a clear picture of hate and non-hate data shown in Table 6. The total no of tweets is 10000 after preprocessing, and the 69.3% hate and 30.7% non-hate data after annotation ask. Based on this dataset, we can implement various datasets.

7 Conclusion

Twitter serves as a useful starting point for social media analysis. Through Twitter, people often express their feelings, ideas, and opinions. The major focus of the current contribution is developing and testing a novel schema for hate speech in Gujarati. About 12,000 Gujarati tweets were gathered for the suggested study using the Twitter API. The data was unclean, so we used Python to explore preprocessing methods. After that, twenty-five people of various ages completed the annotating work as class hate and non-hate. We used cohesion kappa's to test the inter-agreement of annotated tweets, and we were able to reach a k value of 0.86, which indicates extremely strong inter-annotator agreement.

In future work, we will extract the features using different NLP technique and implement the machine learning algorithm for the Identifying of Gujarati hate speech. Additionally, we are expanding the annotation process to gather more annotations for one single post and to expand the corpus size.

References

1. Watanabe, H., Bouazizi, M., Ohtsuki, T.: Hate speech on Twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access **6**, 13825–13835 (2018)
2. Chung, Y.L., Kuzmenko, E., Tekiroglu, S.S., Guerini, M.: Conan–counter narratives through niche sourcing: a multilingual dataset of responses to fight online hate speech. arXiv preprint arXiv:1910.03270 (2019)
3. Struß, J.M., Siegel, M., Ruppenhofer, J., Wiegand, M., Klenner, M.: Overview of GermEval Task 2, 2019 shared task on the identification of offensive language (2019)
4. Kwok, I., Wang, Y.: Locate the hate: detecting tweets against blacks. In: Twenty-Seventh AAAI Conference on Artificial Intelligence (2013)
5. Mubarak, H., Darwish, K., Magdy, W.: Abusive language detection on Arabic social media. In: Proceedings of the First Workshop on Abusive Language Online, pp. 52–56 (2017)
6. Wang, B., Ding, Y., Liu, S., Zhou, X.: YNU Wb at HASOC 2019: ordered Neurons LSTM with attention for identifying hate speech and offensive language. In: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, December 2019

7. Basile, V., et al.: Semeval-2019 task 5: multilingual detection of hate speech against immigrants and women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 54–63 (2019)
8. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. In: Proceedings of NAACL (2019)
9. Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T.: Aggression-annotated corpus of Hindi-English code-mixed data. In: Proceedings of the 11th Language Resources and Evaluation Conference (LREC), Miyazaki, Japan, pp. 1–11 (2018)
10. Viera, A.J.: Understanding inter observer agreement: the Kappa statistic, from the Robert Wood Johnson Clinical Scholars Program, University of North Carolina (2005)
11. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Comput. Linguist. **34**(4), 555–596 (2008)
12. Abhilasha, V., Tanna, P., Joshi, H.: Hate speech detection: a bird's-eye view. In: Kotecha, K., Piuri, V., Shah, H., Patel, R. (eds.) Data Science and Intelligent Applications, pp. 225–231. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-4474-3_26
13. Karim, Md.R., et al.: DeepHateExplainer: explainable hate speech detection in under-resourced Bengali language. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10, IEEE (2021). https://doi.org/10.1109/DSAA53316.2021.9564230
14. Chen, B., Zaebst, D., Seel, L.:A macro to calculate kappa statistics for categorizations by multiple raters. In: Proceeding of the 30th Annual SAS Users Group International Conference, pp. 155–230. Citeseer (2005)
15. Karim, Md.R., et al.: Multimodal hate speech detection from Bengali memes and texts. arXiv: 2204.10196 [Cs], April 2022
16. Gohil, L., Patel, D.: A sentiment analysis of Gujarati text using Gujarati senti word net. Int. J. Innov. Technol. Explor. Eng. **8**(9), 2290–2292 (2019). https://doi.org/10.35940/ijitee.I8443.078919
17. Ishmam, A.M., Sharmin, S.: Hateful speech detection in public Facebook pages for the Bengali language. In: Proceedings of the 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019, pp. 555–560 (2019). https://doi.org/10.1109/ICMLA.2019.00104
18. Alotaibi, M., Alotaibi, B., Razaque, A.: A multichannel deep learning framework for cyberbullying detection on social media
19. Rakholia, R.M., Saini, J.R.: A Rule-based approach to identify stop words for Gujarati language. In: Satapathy, S.C., Bhateja, V., Udgata, S.K., Pattnaik, P.K. (eds.) Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications. AISC, vol. 515, pp. 797–806. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-3153-3_79
20. Ladani, D.J., Desai, N.P.: Automatic stopword Identification Technique for Gujarati text. In: 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), 2021, pp. 1–5 (2021) https://doi.org/10.1109/AIMV53313.2021.9670968
21. Effrosynidis, D., Symeonidis, S., Arampatzis, A.: A comparison of pre-processing techniques for Twitter sentiment analysis. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) TPDL 2017. LNCS, vol. 10450, pp. 394–406. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67008-9_31
22. Al-Twairesh, N., et al.: AraSenTi-tweet: a corpus for arabic sentiment analysis of Saudi tweets. Procedia Computer Science **117**, 63–72 (2017). https://doi.org/10.1016/j.procs.2017.10.094
23. Akhtar, B., et al.: Modeling annotator perspective and polarized opinions to improve hate speech detection. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing **8**(1), 151–154 (2020)

24. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159 (1977). https://doi.org/10.2307/2529310
25. Ramachandran, D., Parvathi, R.: Analysis of Twitter specific preprocessing technique for tweets. Procedia Computer Science **165**, 245–251 (2019). https://doi.org/10.1016/j.procs.2020.01.083
26. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**, 378 (1971)
27. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
28. Davidson, T., Warmsley, D., Macy, M.W., Weber, I.: Automated hate speech detection and the problem of offensive language
29. Hovy, D., Waseem, Z.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL Student Research Workshop (2016)