Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Automatic Hate Speech Detection on Gujarati Language Using Machine Learning

Abhilasha Vadesara[1], Dr. Purna Tanna[2]

[1] *Research Scholar, GLS University, Gujarat, India*

[2] *Assistant Professor, GLS University, Gujarat, India*

Abstract:- Any communication that disparages a target group of people on the basis of a trait like race, color, gender, sexual orientation, ethnicity, nationality or other characteristic is usually referred to as hate speech. There is a steady growth in hate speech as a result of the immense rise in user-generated online content on social media. Along with the phenomenon's growing social effect, interest in online hate speech detection and, in particular, the automation of this task has developed over the past several years. Identification and monitoring of hate speech is becoming an increasingly difficult issue for individuals and society. The objective of this paper is to identify hate speech detection using Natural language processing and Machine learning classifier on gujarati language. This paper compares the four different classifiers like SVM, Naïve bayes, Decision tree and logistic regression with Bag of Word and TF-IDF feature extraction technique. The proposed system pre-preprocesses the twelve thousand tweets and then extract the important features using feature extraction technique to classify into hate and none-hate category using machine learning classifier. Among all classifier naïve bayes classifier and bag of word technique achieved the highest F1-score 91% of hate category and 87.54% accuracy for whole Gujarati corpus including hate and none hate.

Keywords*: Hate speech, text mining, kappa's coefficient, Gujarati Language, Sentiment Analysis.*

1. Introduction

In the current Internet period, multi-modular substance increments on the Internet at high speed. Progresses in Internet Technologies (ITs) and informal online communities have given more advantages to humankind. Social media such as Facebook, Twitter, YouTube, WhatsApp, Instagram, Snapchat, LinkedIn, etc. brought about the age of huge metadata for the mining of data [1]. Twitter, possibly the most well-known web-based social medium and small writing for a blog administration, is a highly famous technique for offering viewpoints and collaborating with others in the internet-based world. Approximately 8.5% of twitter user are bots, according to a survey on Twitter [2], which means it act as agent for spreading hate speech [3].

The speech is a nontrivial instrument to impart thoughts, convictions, sentiments, and other forms of information from one to another. The right to speak freely may be one of the causes of hate speech. Hate speech utilizes hostile, harmful, or offending language towards an individual or a minority of individuals [4]. Various guidelines in various nations deal with hate speech. The target of hate speech is spreading scorn and segregation dependent on the grounds of religion, sex, race, or disability. Each nation has its own definition of hate speech.

According to Indian law, if a person shows the citizen disrespect on the grounds of religion, race, place of birth, residence, language, caste, sexual orientation, gender identity, or any other ground whatsoever is punishable and considered hate speech, as we have seen that bunches of examination found for English and some Indian languages but lake of work has been done on gujarati language [9]. Detecting hate speech on social networking sites is done by NLP and machine learning algorithm. Classification sometimes becomes very complex as the Gujarati language contains adjectives (**વિશેષણ)** and adverbs (**ક્રિયાવિશેષણ).** The gujarati tweet example of hate and None hate speech with the translation.

Hate: પેલા સાવલિયાનું તો મો કાળુ કરવું જોઈએ. (That savaliya should be blacked out.)

None Hate: રાશીબેન એ રાજીનામું આપ્યું. (Rashiben resigned.)

In, this paper expects to identify hate speech on tweet's textual features by ma-chine learning algorithm with Natural Language Processing (NLP) [10,11]. NLP is concerned with applying various statistical pre-processing procedures. The reason for NLP strategies is changing the text-based datasets into datasets that are feasible by ML algorithms. NLP processes include data normalization, stemming, tokenization, and highlights extractions [12,34]. In any case, NLP processes face a few im-pediments when taking care of complex language. We have already built the Gujarati dataset, which focuses on hate speech [34]. The data was gathered using several keywords such as news, racism, Sports, religion, etc. The dataset is categorized into two classes that are Hate and None hate. The word embedding technique like TF-IDF and Bag of Words [13,14] are utilized for extracting a bunch of words includes that can catch the hidden relations of expressions of the dataset and then classifies tweets using SVM [12], Linear Regression, Naïve Bayes, and Decision tree algorithm of Machine Learning [10,15].

The related works in the Section 2 Literature review continue the paper's body. the methodologies and information about the datasets used is provided in Section 3, The performance, observations and analysis are presented in Section 4, and the conclusion and future work are presented in Section 5.

2. Related Work

This overview of the literature discusses significant research on the detection of hate speech. People all around the world are using social media more and more, which has led to a rise in hate speech and other issues that the present research aims to address. While there has been a lot of study on identifying hate speech in general, nothing has been done especially to identify Gujarati hate speech.

Alsafari, Safa, et.al [13] used the deep learning technology to detect hate speech in Arabic. They studied the influence of word-embedding models and neural network topologies on predict accuracy using 2-class, 3-class, and 6-class classification tasks. They train numerous neural networks for each detection task using pre-trained word embedding on Arabic hate and offensive speech dataset. They trained and compared five word-embedding models using CNN, GRU, BILSTM, and a hybrid CNN+BILSTM neural network architecture. They evaluated the performance of each word embedding-classifier pair for three classification tasks, 2-class, 3-class, and 6-class, using an Arabic hate speech dataset. Skip-gram models created more effective representations than other word embeddings, according to the results.

Alatawi, Hind S., et al. [16] looked at the possibility of employing deep learning and natural language processing methods to automatically identify white suprema-cist hate speech on Twitter. They utilized two approaches: the first uses domain-specific embeddings collected from a white supremacist corpus with a bidirectional LSTM deep learning network to determine the relevance of this white nationalist lingo. This method yielded an F1 score of 0.74890. The second technique, the BERT model, is cutting-edge for many NLP applications. It received an F1 score of 0.79605.

Mossie, Zewdie, and JenqHaur Wang [17] proposed using Apache Spark in hate speech identification to reduce the challenges. To categorize Amharic Facebook posts and remarks into hate and not hate, researchers used an Apache Flash-based model. The authors used Word2Vec and TF-IDF for feature selection and Random Forest and Naive Bayes for learning. The model based on word2vec embedding was received best with 79.83% accuracy when tested by 10-fold cross validation.

Suman Rani et al. [12] emphasized their work on extracting emotions or feelings in tweets about Indian politicians using SVM. The performance of the suggested technique is estimated in terms of accuracy, precision, recall, and f-measure using the feature extractors Unigram and TF-IDF.

Wang, Bin, et al. [14] discussed the desirable characteristics of word models and assessment methods while showcasing the well-known word embedding models. Then they divided evaluators into intrinsic and extrinsic

two categories. Extrinsic evaluators utilize word embeddings as input elements to a downstream job and determine changes in execution metrics particular to that activity, whereas intrinsic evaluators test the nature of a representation irrespective of specific natural language processing activities. They provide the trial outcomes of inside and outward assessors on six-word embedding models. It is demonstrated that different word embedding model components are the focus of different investigators, and some of these components are connected to tasks involving natural language processing. Finally, they use correlation analysis to examine how consistently extrinsic and intrinsic assessors perform.

Abro, Sindhu, et al. [15] compared the effectiveness of three feature extraction approaches and eight machine learning algorithms, and evaluated the performance of a publicly accessible dataset with three distinct classes. With the use of the feature and support vector machine technique, they were able to reach 79% accuracy.

The table 1 represent the available dataset and the NLP and ML technique which is used in different language to detect hate speech.

Table 1. Related dataset of hate speech in different language

Paper	Language	Dataset	Class	Dataset size	Features	Algorithm
[18]	English	Web Content	4	165,000	TF-IDF, N-grams, topic similarity, sentiment analysis	Naive Bayes
[19]	Dutch	Facebook	3	5759	Dictionaries	SVM
[20]	English	Facebook and Google	2	300,000	BOW, dictionary, typed dependencies	SVM, Random Forest, Decision Tree
[21]	English	Twitter	2	450,000	N-gram, typed dependencies	Random Forest Decision Tree, SVM, Bayesian Logistic Regression, Ensemble
[22]	Englis, Spanish	Twitter	3	19600	Word Embedding	Machine Learning, Deep Learning

3. Proposed Methodology and Data Collection

All This section explains the proposed studies that we used to divide tweets into the categories "Hate" and "None Hate." The methodology we use for this study is depicted in Fig. 1 The methodology includes five key steps: data collection of Gujarati hate speech twitter data, data pre-processing to clean the data, feature extraction to extract important features from the corpus, data splitting to implement classification model construction to detect hate speech, and classification model evaluation to measure accuracy. The following sections go into great depth about each stage.

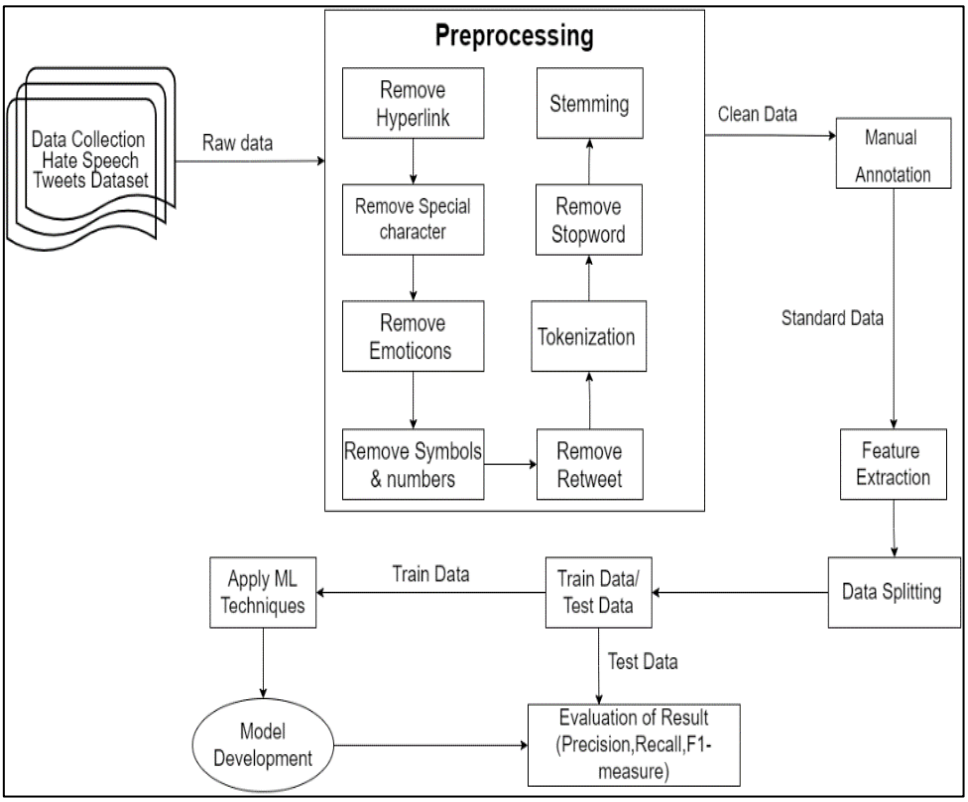

Fig.1. A Sample Process of hate speech detection

3.1. Data Collection

Our major objective was to gather datasets utilizing various methods. The tweets were obtained between January 2020 and January 2021. Using the Twitter API, we gathered information on the percentages of different categories like sports and general (25%), political (37%), community (20%), and the film industry (18%). Since most of the content on twitter is not offensive, we tried a variety of methods to avoid the spread of offensive tweets on around 30% of the dataset. The challenges encountered during Hate Speech evaluation were language types such as sarcasm or indirectness, as well as youth speak, which researchers may not recognize. We have gathered over 12,000 tweets about hateful and non-hateful gujarati tweet. The data set is divided into a training dataset and a test dataset in order to accomplish the classification job.

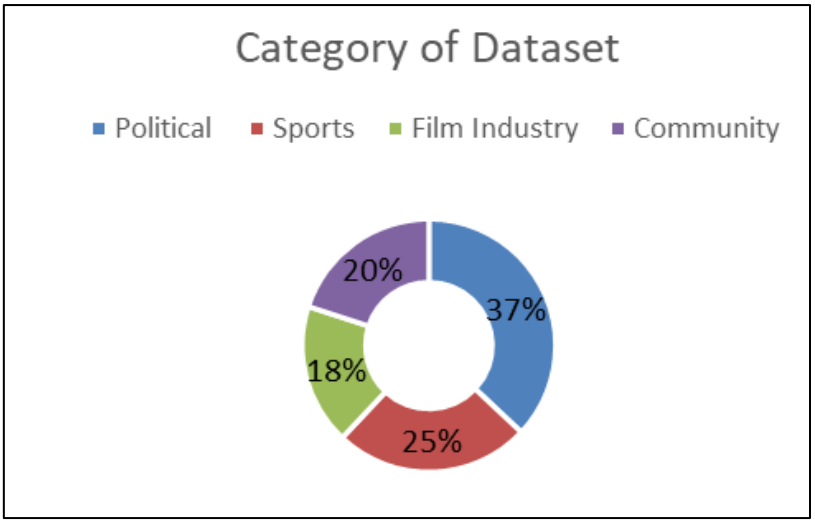

Fig. 2. Categories of dataset

3.2. Text Preparation and Annotation

Text pre-processing improves categorization outcomes, according to a few studies and reviews. We used several pre-processing techniques to remove unclear and noisy data from the corpus. The retrieved data was initially unread due to the Gujarati language and was stored in Unicode. The data was then transformed into readable form using a python tool. The data needed to be cleaned even if it had many new characters in it. We developed an algorithm that eliminates words like "! #," "$," "*," and "1234" from sentences as well as URLs like "http://www.imrobo.com," special characters, emojis, and symbols. In addition, we normalize words with increasing length. A text input is divided into units called tokens using the tokenization process. Additionally, we utilize the stem to trim down the word's affixes and suffixes and exclude the stop word from the tweet. The second stage is annotating the Gujarati corpus after data collection. At present, the amount of annotated data consists of ten thousand tweets. The annotation process includes a multi-step process, and after a fundamental step, it was carried out by Twenty-five annotators manually who were the people of the different age groups. The 28% of people are graduate age range 41- 49 years. The 32% of people are postgraduate age range 29 – 36 years and 40% of people were all college students and language experts with the age range of 19 to 24 years [34].

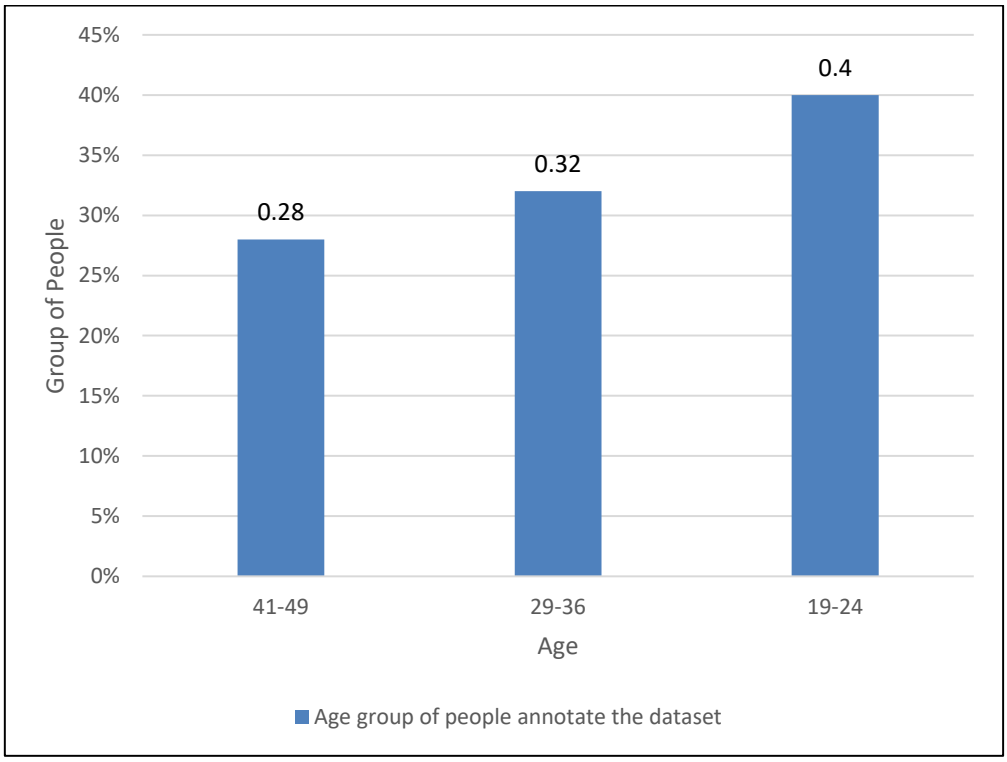

Fig. 3. The age wise distribution of the annotator

There are two factor which defines the hate and non-hate tweets, The first factor considered as a target means the tweet should address, or simply refer to, one of the minority groups previously known as hate speech targets or the person considered for its membership in that category. The second is action, or more explicitly pronounced abusive force, in which it is capable of spreading, inciting, promoting, or justifying violence against a target. To check inter agreement between the annotator we used the Fleiss' kappa κ inter-annotator agreement metric and achieved the 0.87% of agreement which denotes almost perfect agreement as per the kappa's technique [7]. The final dataset contains 67.3% of hate and 32.7% of none hate tweets out of ten thousand tweets. The fig. 4 represents the screenshot of collected dataset along with the annotation process.

Tagging	Data
None Hate	બાંગ્લાદેશનો શાનદાર વાપસી, ભારત 314 રન પર અટકી ગયું #INDvBAN # CWC19
Hate	કાંગ્રેસ નો એક ટાંગો @અખિલેશ યાદવના ત્યાં ગીરો મૂક્યો.
Hate	લાનત છે તારા જેવા પર... #**પાકિસ્તાન** સાથે રમવા એશિયા કપ ભારત ની બહાર રમવા જાય છે એના પર બોલને ભ**!
Hate	શુ **નાલાયક** નફ્ફટ મીડિયા વાળાજવાબનથીઆપીશકતા .લુખ્ખાઓ ખાલી અમારા ધર્મ ને બદનામ જ કરી શકો છો તમે .હિમ્મત હોય તો લઘુમતી વિશે આવી કોઇ પોસ્ટ મૂકીજેવો.
None Hate	**નેહા કક્કર** નાં તો સાવ કાકરા કાઢી નાખ્યાં🤣
None Hate	હું #**કાશ્મીર ફાઇલ્સ** જોઉં છું ત્યારે હું પલંગ પર ચોંટી રહીશ. તે એવી મૂવી નથી કે જેની હું એક મિનિટ ચૂકી જવા માંગુ છું.
Hate	લીમડાનું સરનામું કડવું છે, થર્ડ અમ્પાયર ભ** છે.
Hate	બેવકૂનોથી ભરેલો .. આઈસીસી આઈસીસી #મીડિયા કમ્સ @આઇસીસી #ક્રિકેટવર્લ્ડકપ
None Hate	પ્રણયત્રિકોણ મૂકે છે બધાને 😕મૂંઝવણમાં! ફિલ્મ #KutchExpress માં છે તેનાથી પણ વધુ વિકટ પરિસ્થિતિ
Hate	કુદરતી ધરોહર ના રક્ષક છે એવા #વન_રક્ષક અને #વન_પાલ ના કર્મચારીઓ નાના બાળકો સાથે #ગાંધીનગર માં આંદોલન કરી રહીયા છે અને બહેરી **મૂંગી** સરકારના પેટનું પાણી નથી હાલતું આનાથી મોટી કરૂણતા શું હોઈ ? નાના ભૂલકાઓનો તો અવાજ સાંભળો સરકાર

Fig. 4. Collection of datasets with annotation procedure

3.3. Feature Extraction

The ML algorithm requires numerical features to understand classification rules since it cannot provide an effective response from raw data. Therefor, we have used the Bag of word and TFIDF feature extracting technique to extract numerical feature.

Bag of word (BOW): The embedding vectors of the input text are created using the statistical technique. To create an embedding vector for a phrase, we utilize how frequently the term appears in a document. For the complete set of documents, a matrix is generated with columns denoting each word and rows denoting each document. The values of a term's occurrence frequency in a document are contained in the cells. To train the conventional machine learning methods, feature representation was utilized.

Term Frequency–Inverse Document Frequency (TFIDF): This method converts a document into a vector format. The relevance of a word is inversely correlated with its frequency throughout the document and directly correlated with its frequency within the document.

$$TF\text{-}IDF = TF * IDF$$

$$TF\text{-}IDF = TF * \log(N/DF)$$

Where, Word frequency in a document is represented by TF. N represents total number of the documents in the corpus. DF refers to the total number of documents containing a term within the corpus.

3.4. Machine Learning Classifier

To evaluate the classification model, we kept the 80-20(i.e., 80% Train dataset and 20% for Test Dataset) ratio. According to the "No free launch Theorem" [10], no one classifier can deliver the best results across a variety of datasets. In order to attain a better result, we compare the various classifiers and feature vectors before passing the feature vectors to the classifiers for categorization of hate and non-hate. The machine learning classifiers like Naive Bayes [10], Support Vector Machine (SVM) [17], Decision tree [23], Logistic Regression [23] were picked based on binary classification problem, as well as how well they performed in earlier studies.

3.4.1. Support-vector machines (SVM):

The SVM is the popular machine learning algorithm and used to applied for text classification. Finding a hyperplane that optimizes the minimal distance of the class is the basic idea of SVMs. Support vectors are the instances that define the hyperplane. In binary classification, the instances are split into two distinct classes by a hyperplane created by the support vectors. The experiments by Dadvar et al. [24], Salminen et al. [27], Xu et al. [25], and Nobata et al. [26], are at least the ones that have used SVM for online hate detection with positive outcomes. SVM has a lower cost of computation than deep learning models and gives easier ability for interpretation [28,35]. These make SVM an appropriate addition to our investigations.

3.4.2. Naïve Bayes (NB):

The Nave Bayes (NB) classifier is another standard method that is frequently used as an initial stage in machine learning models. The approach uses a simple probabilistic technique based on the Bayes theorem, conditional independence, and total prob-ability theories. By counting the frequencies and combinations of data in the provided dataset, it determines sets of probability. Despite the fact that conditional independence is rarely true in real-world data, the approach performs well in a variety of supervised classification applications, including text analysis [29].

3.4.3. Logistic regression (LR):

Logistic regression is a linear model, which means it provides interpretable results. You can easily understand the relationship between the input features and the predicted probabilities of hate speech.it can also help identify which features (words or phrases) are the most important in making classification decisions. This feature se-lection can provide insights into the language and content associated with hate speech, aiding in further analysis and content moderation. At least Xiang et al. [30], Burnap and Williams [31], and Salminen et al. [27] have utilized LR for online hate detection.

3.4.4. Decision tree(DT):

The Decision tree classifier is often applied due to their interpretability, simplicity, and ability to handle both numerical and categorical data. it can readily identify which features (words or phrases) are the most important for distinguishing hate speech from non-hate speech. This feature selection can offer insights into the linguistic characteristics associated with hate speech, aiding in content moderation and further analysis. Sometimes, hate speech detection datasets often suffer from class imbalance, where the majority of samples belong to the non-hate speech class. The decision trees can be adapted to handle imbalanced datasets by adjusting class weights or using techniques like SMOTE (Synthetic Minority Over-sampling Technique) [23].

3.5. Classifier Evaluation

In this stage, we evaluate the performance of the classifier, which will predict the result of unlabeled test dataset i.e., “hate” and “non-hate”. The performance of model is evaluated by calculating the Precision, Recall, and F1 measure [2,32]. The following equations are evaluated for classification performance.

Precision. The ratio of relevant examples among the retrieved examples or it con-sider as positive predictive value.

$$\text{Precision} = \frac{TruePositive}{(TruePositive+FalsePositive)} \quad (1)$$

Recall. The percentage of relevant examples out of all relevant examples that have been retrieved.

$$\text{Recall} = \frac{TruePositive}{(TruePositive+FalseNegetive)} \quad (2)$$

F1-score. The integrated average of recall and precision.

$$\text{F1} = \frac{2*Precision*Recall}{(Precision+Recall)} \quad (3)$$

Accuracy. The true prediction among the total number of predictions.

$$\text{Accuracy} = \frac{(TruePositive+TrueNegative)}{(TruePositive+TrueNegative+FalsePositive+FalseNegative)} \quad (4)$$

4. Experiment and Result

This section outlines the experimental parameters for implementing the twelve thousand gujarati hate speech tweet dataset into machine learning models. Where we evaluated two categories hate and none hate labelled 1 and 0. Only documents that had been annotated by 25 distinct age groups of annotators were included in the experiment. We used the Fleiss' kappa to test the inter-agreement between annotator with python tool and obtained an accuracy of 0.87%, which is almost perfect by kappa's standard [7]. Following this procedure, we obtained the data from the entire corpus at 67.3% hate and 32.7% non-hate mentioned in table 2.

Table 2. Category based Dataset

Numerical Representation	Class	Total Instance
0	hate	6930
1	non-hate	3070
	Total	10000

To get the best outcome for hate speech identification, the four distinct classifiers SVM, Naive Bays, Logistic Regression, and Decision Tree were used. The feature extraction techniques TF-IDF and BOW were utilized in this instance to extract the key features from the dataset because the machine learning classifier deals with mathematical features. The feature extraction is done after the preprocessing task mention in table 3.

Table 3. Data Cleaning with preprocessing technique

Preprocessing Techniques	Raw data	Clean Data
Removal of URLs, hashtags, user mentions, other characters and noise	@ મનકગુપ્તા ધોની આજથી રડવાનું શરૂ કરશે. @MSDhonifansclub #INDvBAN https://t.co/A9nlIbWMWo	મનકગુપ્તા ધોની આજથી રડવાનું શરૂ કરશે.
Removal of Emoticons	@ આઈસીસી @ નિસાન આ વિશ્વ કપ 19 પર પ્રતિબંધ હોવો જોઈએ!! 😡😡	@ આઈસીસી @ નિસાન આ વિશ્વ કપ 19 પર પ્રતિબંધ હોવો જોઈએ!!
Removal of punctuations	ધોનીના ગ્લોવ્સ પર આઈ.સી.સી. વિ.આઈ.સી.સી. વર્લ્ડ કપનું આયોજન કરે છ?	ધોનીના ગ્લોવ્સ પર આઇ સી સી વિ આઇ સી સી વર્લ્ડ કપનું આયોજન કરે છે
Removal of number	મહિલા ટી20 ક્રિકેટ : ભારતે બાંગ્લાદેશને 8 રનથી હરાવ્યું, 2-0થી જીતી સીરિઝ	મહિલા ટી ક્રિકેટ : ભારતે બાંગ્લાદેશને રનથી હરાવ્યું, - થી જીતી સીરિઝ
Removal of Stop-words	પુત્ર આદર જૈન ના તારા સુતરીયા સાથે ના સંબંધ ને માતા રીમા જૈને આપી મંજૂરી.	પુત્ર આદર જૈન તારા સુતરીયા સાથે સંબંધ માતા રીમા જૈને મંજૂરી.
Tokenizing	પઠાણ ફિલ્મમાં સલમાન ખાન છે	પઠાણ ,ફિલ્મમાં, સલમાન, ખાન, છે

In common classification tasks, evaluation metrics including precision, recall, and F-score are used. Due to extremely unbalanced datasets, a high accuracy value does not always equate to strong performance on other

assessment measures. For this reason, this study analyses specific metrics for each class (hate and none hate). In that situation, F1 is a more trustworthy metric, and useful hypotheses may also be obtained from Precision and Recall.

The table 4 represents the comparison result of evaluation metrics of hate category with machine learning classifier and feature extraction technique. Based on the comparison, the naive bayes classifier with BOW achieved best F1 - score of 91% which is the best among all classifier. Even the SVM Classifier with BOW is also achieved relatively close performance with 89% of F1- score. Where the Logistic Regression with TF-IDF and Decision tree with BOW technique perform low with 87%, 84% of F1-score respectively.

Table 4. The performance results of hate category based on the precision, recall, and F1 measure

Classifier/Feature Extraction Model	Bigram (TF- IDF)			BOW		
	P	R	F1-Score	P	R	F1-Score
SVM	0.87	0.85	0.86	0.89	0.88	0.89
LR	0.76	0.99	0.86	0.80	0.96	0.87
NB	0.80	0.98	**0.87**	0.90	0.91	**0.91**
DT	0.84	0.87	0.85	0.82	0.86	0.84

Based on the comparison of two feature extraction technique the fig 5 shows the BOW technique performs good as compare as TF-IDF technique.

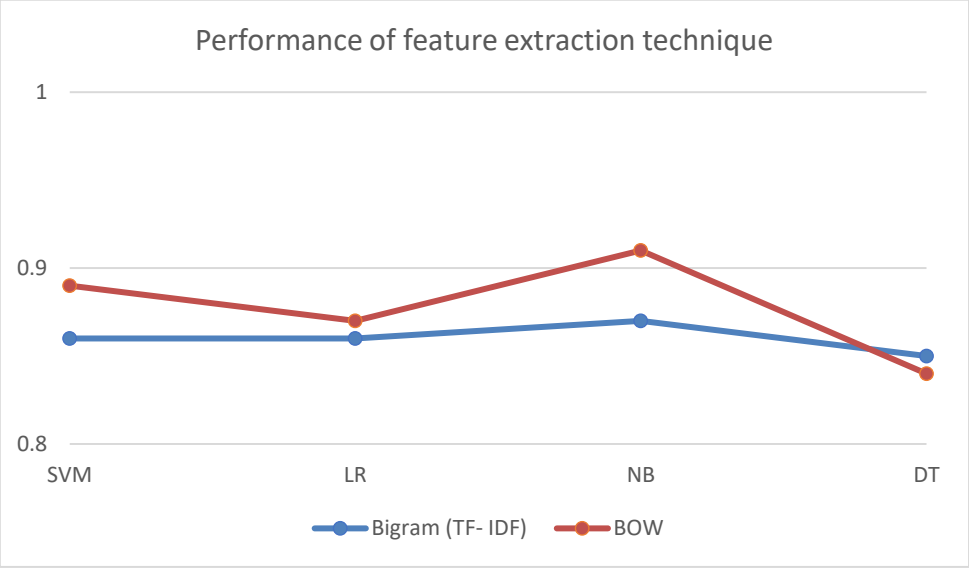

Fig. 5. Evaluation of feature extraction technique for gujarati dataset.

Among all machine learning techniques that we examined; the naïve bays classifier has achieved the highest performance with both the feature extraction technique. Therefore, in order to detect hate speech, bag of words feature extraction approach with the naïve bayes classifier has been chosen. So here the experiment with Bag of word technique describes in table 5. For the experiment, we have considered three gujarati sentences as preprocessed. It describes the frequency of word in the document and convert it into vector to perform classification.

Sentence 1: એ નેહા કક્કર છે. (She is Neha Kakkar.)

Sentence 2: **નેહા કક્કર કોન્સર્ટ, અમદાવાદ!** (Neha Kakkar Live Concert, Ahmedabad!)

Sentence 3: **નેહા કક્કર ને આપવા ટકકર આવી ગયા છે પાયલ ટીકટોકર.** (Payal TikToker has come to fight with Neha Kakkar.)

Table 5. Experiment of Bag of word technique

Sentences/ Vocabulary	નેહા	કક્કર	કોન્સર્ટ	અમદાવાદ	ટકકર	પાયલ	ટીકટોકર
Sentence1	1	1	0	0	0	0	0
Sentence2	1	1	1	1	0	0	0
Sentence3	1	1	0	0	1	1	1
Word Frequency	3	3	1	1	1	1	1

The fig. 6 is the result after final implementation of naïve bays classifier. The model automatically predicts the hate speech entered in textbox.

Fig. 6. Hate speech detection model

5. Conclusion

From this experiment, we have shown the comparison of machine learning model with different feature extracting technique to detect hate speech in gujarati dataset. The experiment has been done on the twelve thousand gujarati tweet dataset. which is annotated by twenty-five different people. To check the inter agreement between annotator we implemented the Fleiss' kappa and achieve the 87% of accuracy which is almost perfect. For detecting hate speech, we have compared four different classifiers like SVM, Naïve Bays, Logistic Regression and Decision tree with TF-IDF and Bag of word feature extraction technique. Among all classifier we chose Naïve Bays classifier with Bag of Word technique as it achieved the highest 0.91 F1- score in hate category and achieved the 87.54% of overall accuracy for detection of hate speech.

References

[1] N.A. Ghani, S. Hamid, I.A.T. Hashem, E. Ahmed, Social media big data analytics: A survey, Comput. Hum. Behav. 101 (2019) 417–428.

[2] Subrahmanian V.S., Azaria A., Durst S., Kagan V., Galstyan A., Lerman K., Zhu L., Ferrara E., Flammini A., Menczer F.The DARPA Twitter bot challenge Computer, 49 (6) (2016), pp. 38-46

[3] P.N. Howard, B. Kollanyi, S. Woolley, Bots and automation over Twitter during the US election, Computational Propaganda Project: Working Paper Series, 2016.

[4] Erjavec K., Kovačič M.P. “You don’t understand, this is a new war!” analysis of hate speech in news web sites’ comments Mass Communication and Society, 15 (6) (2012), pp. 899-920, 10.1080/15205436.2011.619679

[5] Rosa, J., and Y. Bonilla, Deprovincializing Trump, decolonizing diversity, and unsettling anthropology. American Ethnologist, 2017. 44(2): p. 201-208.
[6] Travis, A., Anti-Muslim hate crime surges after Manchester and London Bridge attacks. The Guardian, 2017.
[7] Kilem L Gwet. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, 2014
[8] Dibyendu Mishra , Syeda zainab akbar ,et al.,Rihanna versus Bollywood: Twitter Influencers and the Indian Farmers' Protest, Microsoft Research, India,2021
[9] Alam, Iftikhar, et al. "Free vs Hate Speech on Social Media: The Indian Perspective." Journal of Information, Communication and Ethics in Society, vol. 14, no. 4, Nov. 2016, pp. 350–63. DOI.org (Crossref), https://doi.org/10.1108/JICES-06-2015-0016.
[10] Joachims, T. Text categorization with support vector machines: Learning with many relevant features. in European conference on machine learning. 1998. Springer
[11] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: International Workshop On Natural Language Processing For Social Media, 2017, pp. 1–10.
[12] Suman Rani et al, International Journal of Computer Science and Mobile Applications, Vol.5 Issue. 10, October- 2017, pg. 83-91
[13] Alsafari, Safa, et al. "Effect of Word Embedding Models on Hate and Offensive Speech Detection." ArXiv:2012.07534 [Cs], Nov. 2020. arXiv.org, http://arxiv.org/abs/2012.07534.
[14] Wang, Bin, et al. "Evaluating Word Embedding Models: Methods and Experimental Results." APSIPA Transactions on Signal and Information Processing, vol. 8, 2019, p. e19. DOI.org (Crossref), https://doi.org/10.1017/ATSIP.2019.12.
[15] Abro, Sindhu, et al. "Automatic Hate Speech Detection Using Machine Learning: A Comparative Study." International Journal of Advanced Computer Science and Applications, vol. 11, no. 8, 2020. DOI.org (Crossref), https://doi.org/10.14569/IJACSA.2020.0110861.
[16] Alatawi, Hind S., et al. "Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding With Deep Learning and BERT." IEEE Access, vol. 9, 2021, pp. 106363–74. DOI.org (Crossref), https://doi.org/10.1109/ACCESS.2021.3100435
[17] Mossie, Zewdie, and Jenq-Haur Wang. "Social Network Hate Speech Detection for Amharic Language." Computer Science & Information Technology, Academy & Industry Research Collaboration Center (AIRCC), 2018, pp. 41–55. DOI.org (Crossref), https://doi.org/10.5121/csit.2018.80604.
[18] Shuhua Liu and Thomas Forss. 2014. Combining N-gram based similarity analysis with sentiment analysis in web content classification. In Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. 530–537
[19] Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media (2016).
[20] Pete Burnap and Matthew L. Williams Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Sci. 5, 1 (2016).
[21] Pete Burnap and Matthew L. Williams Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. Policy Internet 7, 2 (2015), 223–242.
[22] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. 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, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
[23] Omran, E., Al Tararwah, E., & Al Qundus, J. (2023). A comparative analysis of machine learning algorithms for hate speech detection in social media. Online Journal of Communication and Media Technologies, 13(4), e202348.

[24] Dadvar M et al (2013) Improving cyberbullying detection with user context. Eur Conf Inf Retriev 2013:693–696
[25] Xu J-M, et al. Learning from bullying traces in social media. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies; 2012. P. 656–66
[26] Nobata C, et al. Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web, Geneva, Switzerland; 2016. P. 145–53
[27] Salminen J, et al. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In: Proceedings of the international AAAI conference on web and social media (ICWSM 2018), San Francisco; 2018
[28] Karan M, Šnajder J. Cross-domain detection of abusive language online. In: Proceedings of the 2nd workshop on abusive language online (ALW2); 2018. P. 132–137
[29] Wang S, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2; 2012. P. 90–4
[30] Xiang G, et al. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In: Proceedings of the 21st ACM international conference on Information and knowledge management; 2012. P. 1980–4
[31] Burnap P, Williams ML (2016) Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Sci 5(1):11.
[32] M, Hossin, and Sulaiman M.N. “A Review on Evaluation Metrics for Data Classification Evaluations.” International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, Mar. 2015, pp. 01–11.
[33] Vadesara, tanna and joshi (2021). Hate speech detection: a bird’s-eye view, data science and intelligent applications, vol 52. Springer, singapore.
[34] Vadesara, Abhilasha, and Purna Tanna. “Corpus Building for Hate Speech Detection of Gujarati Language.” Soft Computing and Its Engineering Applications, vol. 1788, Springer Nature, 2023, pp. 382–95.
[35] Vadesara, Abhilasha, and Purna Tanna. “Evaluating effectiveness of Feature Extraction technique in Gujarati Hate Speech Detection” SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, vol 14 no 02, Institute of Technology, School of Management Sciences, Lucknow, 2022.