IJE TRANSACTIONS A: Basics Vol. 37 No. 07, (July 2024) 1284-1295

International Journal of Engineering

Journal Homepage: www.ije.ir

Investigating Hostile Post Detection in Gujarati: A Machine Learning Approach

B. J. Rameshbhai*[a], K. Rana[b]

[a] *Department of Computer Engineering, Gujarat Technological University, Ahmedabad, Gujarat, India*
[b] *Department of Computer Engineering, Sarvajanik College of Engineering and Technology, Gujarat Technological University, Ahmedabad, Gujarat, India*

PAPER INFO

Paper history:
Received 03 September 2023
Received in revised form 13 December 2023
Accepted 13 Janurary 2024

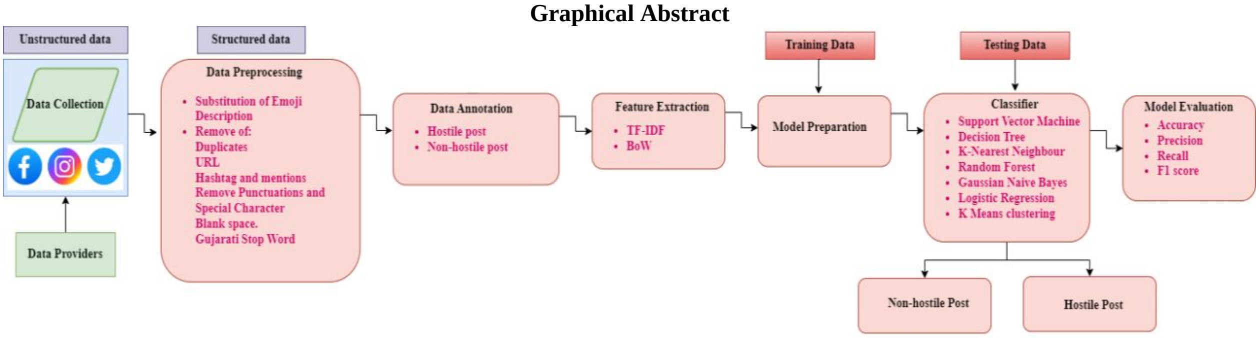
Keywords:
Hostile Text Detection
Machine Learning
Hate Text Detection
Text Classification
Gujarati Text Dataset

ABSTRACT

Hostile post on social media is a crucial issue for individuals, governments and organizations. There is a critical need for an automated system that can investigate and identify hostile posts from large-scale data. In India, Gujarati is the sixth most spoken language. In this work, we have constructed a major hostile post dataset in the Gujarati language. The data are collected from Twitter, Instagram and Facebook. Our dataset consists of 1,51,000 distinct comments having 10,000 manually annotated posts. These posts are labeled into the Hostile and Non-Hostile categories. We have used the dataset in two ways: (i) Original Gujarati Text Data and (ii) English data translated from Gujarati text. We have also checked the performance of pre-processing and without pre-processing data by removing extra symbols and substituting emoji descriptions in the text. We have conducted experiments using machine learning models based on supervised learning such as Support Vector Machine, Decision Tree, Random Forest, Gaussian Naive-Bayes, Logistic Regression, K-Nearest Neighbor and unsupervised learning based model such as k-means clustering. We have evaluated performance of these models for Bag-of-Words and TF-IDF feature extraction methods. It is observed that classification using TF-IDF features is efficient. Among these methods Logistic regression outperforms with an Accuracy of 0.68 and F1-score of 0.67. The purpose of this research is to create a benchmark dataset and provide baseline results for detecting hostile posts in Gujarati Language.

doi: *10.5829/ije.2024.37.07a.08*

Graphical Abstract



1. INTRODUCTION

Gujarati is the sixth most spoken language in India, spoken by 56 million people[1]. During covid19 pandemic the use of social media platforms such as Twitter, Facebook, Instagram, linkedin, Reddit and YouTube has increased drastically (1, 2). Social media platforms are now used for many purposes such as education, politics, entertainment, business, charity work etc. It affects people of all age groups. On social media, people share

* Corresponding Author Institutional Email: *jagrutiboda10@gmail.com* (B. J. Rameshbhai)

[1] https://en.wikipedia.org/wiki/Gujarati language

Please cite this article as: Rameshbhai BJ, Rana K. Investigating Hostile Post Detection in Gujarati: A Machine Learning Approach. International Journal of Engineering, Transactions A: Basics. 2024;37(07):1284-95.

their opinion, suggestions and emotions very openly (3, 4). Many times abusive language, the spreading of a violent message and swear words are also used. Hostile contents affect the mental health of people and promote violence thus spreading negativity among the users. The hostile post refers to abusive language mentioned in the post that targets individuals, organizations, groups of people, communities, religious, races, gender etc. (5). According to the survey, there is an increase in the content of Indian language hostile posts on social media, like Hindi (1, 2, 6-9), Marathi (7, 10-12), Bengali (8), Khasi (9), Punjabi, Gujarati and English language (9). Very less research work has been done in the low-resource Indian languages (5). Hence, we require a system that can automatically detect hostile posts written in these languages (10). Social media have a vast number of people who write their posts in Gujarati language. It is essential to have a system for Gujarati language. The state-of-the-art hostile text detection methods are available for English language. To enhance the research in low resource Indian languages, we have studied various methods which can detect hostile posts in Hindi (1, 2, 6-8, 13, 14), Marathi (7, 11, 12), Bengali (12), Saudi (4), Roman Urdu (15, 16), Tamil (17) and code-mixed language (7, 18-22). Figure 1 shows basic approaches used for hostile post-detection. Mainly these approaches are divided into two major categories: (i) Machine Learning based (23-35) and (ii) Deep Learning based (5, 6, 27-29, 34-38). Machine learning-based approach is categorized into two subparts:(i) Supervised machine learning based and (ii) Unsupervised machine learning based (27, 30-32, 37). Support vector machine (SVM) (39), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), Gaussian Naive Bayes (GNB) and Logistic Regression (LR) algorithms are used for supervised learning (24). k-means clustering algorithm is used for unsupervised learning (31, 32).

To improve accuracy, researchers use deep learning-based approaches to detect hostile post in high-resource languages. The deep learning-based approach is divided into main two categories :(i) Encoder-Decoder based (9) and (ii) Transformer Based (10, 19). The Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long short-term memory (LSTM) and Bidirectional Long short-term memory (BiLSTM) (29) [32] based approaches come under the first category and Bidirectional Encoder Representations from Transformers (BERT), Romanian BERT Model (RoBERT), Multilingual BERT etc. come under the transformer based approach (3, 33, 40). To the best of our knowledge, we are the first in detecting Guajarati hostility post. We prepared a 10,000 manually labeled dataset that is valuable work in Gujarati Language Processing. This research is significant as it provides a baseline result to detect hate text in Gujarati. The proposed approach has the potential to make a meaningful impact on the online community and create a more inclusive and respectful environment. The objective of our paper is to create a benchmark Guajarati hostility detection text dataset, prepare a systematic literature review and detect hostile posts using seven machine learning classifiers SVM, DT, RF, KNN, GNB, LR and K means clustering.

1. 1. Motivation

People normally use their native language frequently for sharing opinions, suggestions and ideas on social media. As per research literature, Gujarati text processing has not been done for a large scale. We aim to detect hostile posts in the Gujarati language. To the best of our knowledge, machine

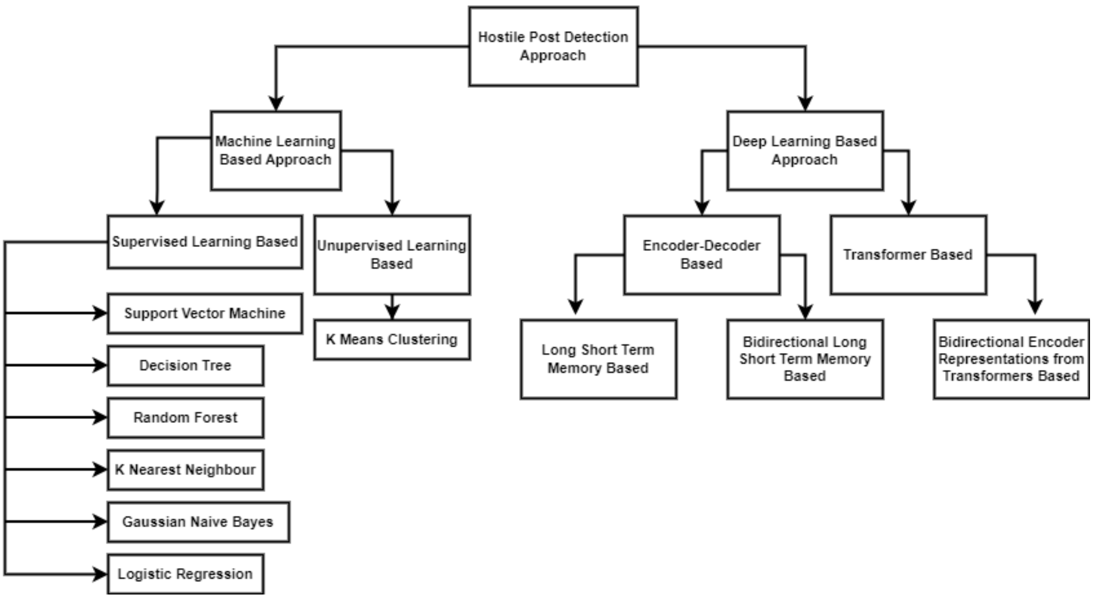


Figure 1. Existing hostile post detection approaches

learning techniques are being used for hostile post detection in the Gujarati language for the first time.

1. 2. Contribution Following are the contributions of this research work.

- There is no standard dataset available for Gujarati Language Processing (GLP). Therefore, we have created a big dataset for Gujarati language processing, consisting of 1,51,000 posts scraped from the social media platform Twitter, Instagram, and Facebook. Total 10,000 posts are manually labeled with two categories: (i) Non-Hostile and (ii) Hostile.
- The dataset is used in two ways:(i) Gujarati Text Data collected from social media (ii)Translation-based approach - the whole dataset is translated into English using Google API.
- Usefulness of emoji's in Gujarati Text Understanding
- Compared the performance of two feature extraction methods: Bag-of-Words (BoW), Term frequency-inverse document frequency (TF-IDF).
- Compared the performance of machine learning models: SVM, DT, RF, KNN, GNB, LR and k-means clustering.
- Listed major challenges that are identified for Gujarati text processing.

The remaining paper is arranged as follows: Review of existing hostile post detection is discussed in section 2. Dataset construction is discussed in section 3. The method that is used for data preprocessing, feature extraction and model development is described in section 4. Sections 5 and 6 discuss the experimental results. The error produced by our model and challenges are described in section 7. Finally, section 8 concludes our work and discusses future direction.

2. RELATED WORK

The usage of social media drastically increases day by day by people of various educational backgrounds and cultures. The inappropriate content on social media spreads negativity and damages the hygiene of the web. Detecting the hostile post for a low-resource Gujarati language is challenging because of the lack of sufficient labelled data and very less contribution from the researchers (5). We studied various approaches used for hostile post detection in different languages. Bhatnagar et al. (1) has developed an automated system that can automatically detect hostile posts in Hindi language. The proposed novel approach is used for multi-label classification. The model can also distinguish hostile posts and offensive speech. The deep learning-based model outperformed Hindi language hostile post-detection. Velankar et al. (7) presented the Marathi Language Dataset L3CubeMahaHate that contains 25000 various samples that are classified into 4 classes. Deep learning-based various models such as CNN, LSTM, BiLSTM and transformer-based BERT models like IndicBERT, mBERT, and A Robustly Optimized BERT Pretraining Approach (RoBERTa) are used on their dataset. The dataset is also evaluated on monolingual Marathi BERT models like MahaBERT, MahaALBERT and MahaRoBERTa. The transformer approach provides a result for 4 class classification. Banerjee et al. (40) explored various transformer approaches such as mBERT, XMR-large, XLMR-based etc. These models were developed for hostility detection in Indo-Aryan and English language. The model performed excellently for English multi-class classification. The model classifies text into hate, offensive and profane categories. They tested the model for English, Hindi, Marathi and code-mixed language. They concluded that XLM-Roberta-large model outperforms. Warjri et al. (41) introduced a fake news detection approach for the Khasi language for the first time. The fake news collected from social media articles and posts. They have manually annotated 116 news and applied three machine learning classifiers RF, DT and LR. Among these classifiers DT provides the highest accuracy. Bhardwaj et al. (14) proposed a novel deep learning based HostileNet architecture. HostileNet used the concept of transformer based approach BERT. The author added hand-crafted features such as lexicon, emoticon, and hashtag embeddings with Hindi BERT to improve the accuracy for hate text detection. They used publicly available dataset CONSTRAINT-2021 dataset for coarse-grained that means binary label such as hostile and non-hostile classification and fine-grained that means multi label such as fake, hate, offensive, defamation classification. Luo et al. [26] have discussed the supervised machine learning-based approach for text classification. They created their own data that consist of different categories of data such as women, sports, literature etc. Supervised Machine Learning based classifiers SVM, KNN, NB and LR are used to classify data. SVM outperforms their data. Finally, they conclude that the classification algorithm accuracy depends on the type and size of the dataset. Felber (25) used a machine learning model for COVID-19 Fake News Detection. He focused on various text features such as n-grams, readability, emotional tone and punctuation. These text features are used for text understanding. SVM with linear Kernel, Random Forest, Logistic regression, Naive Bayes and Multilayer Perceptron are used to identify fake news detection. SVM provides the best performance for fake news detection. Fahad et al. (26) proposed a novel approach for finding bad intended news using machine learning. They aimed to set the best accuracy for detecting fake news. Most of the methods try to work on specific article domains but it is challenging to work for diverse news domains. They are using various textual characteristics and machine learning classifiers on publicly available real world dataset to detect the fake news. The SVM and LR provide best results to detect fake news. Aluru et al. (34) performed a deep survey for hate speech detection in multilingual languages. The 16 different data sets and 9 languages were covered for hate speech detection. The deep learning-based model was developed for multilingual hate speech classification. The experimental result shows that LASER + LR (Language-Agnostic SEntence Representations + Logistic Regression model) based approach outperforms for low resource language. The transformer-based approach BERT is more effective for high-resource language. Akram et al. (32) tried a novel Deep Auto-Encoder Based Linguistics approach for Urdu News headlines clustering. For the first time a clustering model for Urdu News clustering was used. The result analysis exhibited that Urdu news headlines were easily categorized. The deep learning-based text clustering and k-means clustering algorithm was implemented for news headline clustering. The Deep learning-based approach outperformed k-means clustering approach. Deep Literature study shows that the majority of the researchers focused on high resource language. They applied Deep learning and transformer based models on existing dataset to improve the detection accuracy. From recent study, it is clear that no standard dataset s available in Gujarati language and none have set a baseline result for hate text detection in Gujarati language. After discussing all the existing work, we have found that the system can perform the following task for hostile post detection. Figure 2 shows the flow of process for hostile post detection. The detailed explanation of each phase is discussed in the subsequent sections:

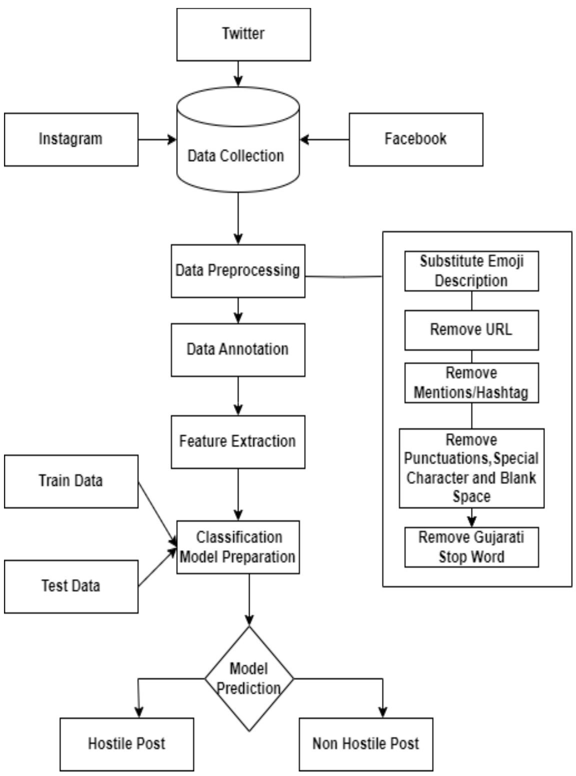


Figure 2. System flow

3. DATASET CONSTRUCTION

We created a new dataset that covers various types of hostile posts such as racist, religious, political, educational and festival in Gujarati text. To obtain the hostile post, we have prepared a list of good and bad keywords in Gujarati with the help of two Gujarati language experts. Both experts have completed their postgraduate studies in Gujarati. These keywords are frequently used by social media users to spread hostile posts. We identified the events that occurred in Gujarat over the last five years between 2017 to 2022. Based on that we have prepared a list of 95 keywords to retrieve comments from social media. Table 1 shows a few examples of terms used for data retrieval. We collected data from the most widely used three social media platforms Twitter, Instagram and Facebook. We have used a web scrapper tool Apify to collect comments using various 95 keywords such as 'Gujarat', 'Patidar', 'Mandi', ' Sports', 'Janta', 'Corona', 'Mataji' and many more. We identified swear words that are frequently used in Gujarati language. Swear words are also used to retrieve hateful comments from social media posts. All the comments are written exclusively in Gujarati.

Twitter Apify library gives a maximum 3300 comments for a particular word search We have collected our data without any bias. Many comments are written in a mixed language such as Gujarati-Hindi and Gujarati-English. These mixed-language comments are eliminated and Gujarati comments are selected manually. Each comment is collected using Python code and extracted

TABLE 1. Examples of terms used for data retrieval

Sr. No.	Search word	Total number of posts
1	પાટીદાર આંદોલન	3300
2	ગરબા	1176
3	દેશભક્તિ	3300
4	નવરાત્રી	3300
5	ગુજરાત સરકાર	507
6	મહાભારત	3300
7	માતાજી	3300
8	મુખ્યમંત્રી	3332
9	રામાયણ	3300
10	કોરોના	1020

data is collected in a CSV file. The sample of collected data is shown in Table 2.

3. 1. Data Annotation The data are manually labelled with basic two categories:

- Hostile Post: It contains hateful text and swear words that targets some gender, religious, organization, government and individual person.
- Non-Hostile Post: It does not contain harmful text which is normally neutral and positive content.

4. PROPOSED METHODOLOGY

4. 1. Pre-Processing While manually checking the data, we noticed that the dataset consists of emoji's, blank spaces, links, special symbols etc. Preprocessing is necessary step to obtain better performance for text classification problem. In natural language processing, first step is to preprocess the data. This step cleans the data and prepares comments as inputs to the classification model. We used appropriate regular expressions to clean the data. We used scikit-learn for preprocess the data.

TABLE 2. Sample of collected data

Sr. No.	Comments
1	@AAPGujarat આતંકવાદી નું સમર્થન કરતી આપ સરકાર
2	પાકિસ્તાન ભણવા ગયેલા કાશ્મીરના યુવાનો આતંકવાદી બની ગયા, એમાંથી૧૭નાં મોત #National News #Kashmir #Pakistan #Terrorist #MidDay News #MidDay Gujarati https://t.co/YheFEs4R4O
3	મારી શાયરી પસંદ કરવા બદલ હું તમારો હ્રદય પૂર્વક આભારી છું.શું હું તમને ટેગ કરી શકું છું ?😊🙈@panktinipanktio
4	રશિયા યુક્રેન સમજૂતી...ફોડો ફટાકડા,વગાડો ઢોલ. https://t.co/vlIKOoUO4a
5	@NobatDaily અલગાવ નહિ આતંકવાદી કહો. સાંતી દુતો થી ફાટતી લાગે છે

TABLE 3. Dataset Description

Dataset Characteristics	Collected Data
Total number of posts	1,51,000
After pre-processing dataset size	1,21,000
Annotated Data	10,000
Hostile Label Data	5000
Non-Hostile Label Data	5000
Minimum number of words in one post	1
Maximum number of words in one post	252
Minimum number of character in one post	2
Maximum number of character in one post	1444
Annotated data with Emoji	1200

- Removal of Duplicate Sentence: The duplicate data are removed from the dataset.
- Uniform Resource Locator (URL): The URL or link is removed using regular expression.
- Mentions and Hashtag: The mentions and hashtags are deleted from the data.
- Punctuations: The Special characters and punctuation marks are removed from the data.
- Blank Space: The extra spaces are also removed from the comment.
- English Hindi and Coded Mixed Sentence: Sometimes user writes comments in Hindi-Gujarati or Gujarati-English mixed language. We have removed such comments manually from the dataset.
- Gujarati Stop Words: We have prepared a list of Gujarati stopwords so that these stopwords can be removed from the dataset.
- Emojis: People normally write comments with emojis. We have replaced emoji's with appropriate description. The flow of emoji description substitution is shown in Figure 3. Table 4 illustrates all the pre-processing tasks with example.

4. 2. Feature Extraction Using Bag-of-Words and TF-IDF Model The pre-processed sentences are converted into numeric data using the feature extraction method. Out of many feature extraction methods, we have used BoW and TF-IDF (28). During Literature study, we identified that these two feature extraction techniques are widely used to extract the feature and convert text data into numeric data.

- Bag-of-Words (36): Bag-of-words (BoW) method is used to extract the text feature to prepare a model. Bag-of-Words consist of: (i) A measure of the presence of words in the document (ii) A Vocabulary of words from the document.
- TF-IDF (27, 28, 37): One of the most effective feature extraction methods is TF-IDF. TF-IDF stands for Term Frequency – Inverse Document Frequency. TF-IDF method converts words into numeric data based on the frequency of words in the document and the importance of the word. TF is Term frequency that refers to the total number of times a given term appears in the document against the total number of words in the document. The IDF is the inverse document frequency that measures how much information a word provides. It measures the weight of a given word in the entire document.

4. 3. Model Preparation We have implemented machine learning-based techniques for hostile post detection on our dataset. We have used supervised learning based and unsupervised learning based approaches.

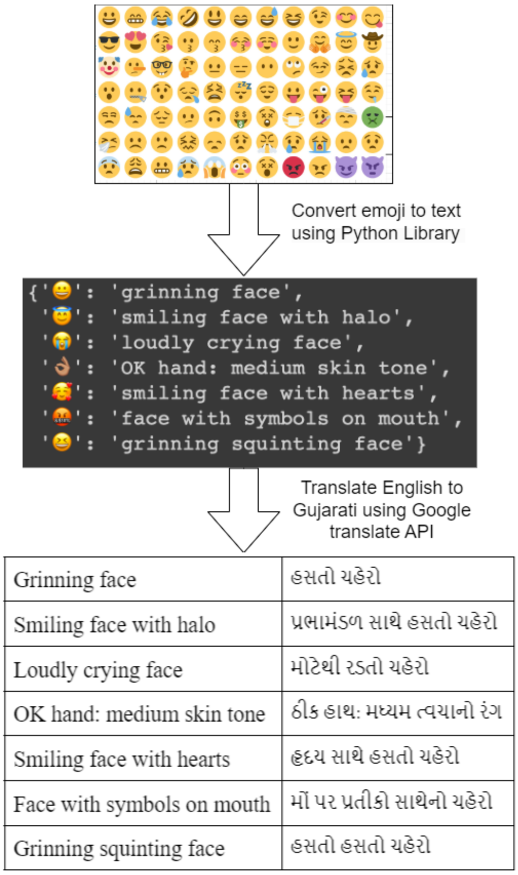


Figure 3. Flow of emoji description substitution

TABLE 4. Pre-processing task

Preprocessing step	Before preprocessing	After preprocessing
Substitute emoji description	મોંઘવારી, બેરોજગારી, શિક્ષણનાં ખાનગીકરણથી બચીએ, હવે પરિવર્તન લાવીએ. 😀☺	મોંઘવારી, બેરોજગારી, શિક્ષણનાં ખાનગીકરણથી બચીએ, હવે પરિવર્તન લાવીએ. હસતો ચહેરો ખુશખુશાલ, લાલી ચહેરો
Remove URL	ભારત માતા કી જય' બોલનાર દેશપ્રેમી અને ન બોલનાર દેશદ્રોહી સાબિત થતાં નથી - https://t.co/bw1A62wF6q	ભારત માતા કી જય' બોલનાર દેશપ્રેમી અને ન બોલનાર દેશદ્રોહી સાબિત થતાં નથી
Remove hashtag	પાટીદાર આંદોલન ને આશ્રય લઇને ગુજરાત રાજનીતિ માં ભૂસકો મારવો એ લાલચી @HardikPatel_. રેલી માં તો આને કહ્યું તું કે તે રાજનીતિ માં નહિ આવે. #Hardik #Fakepatidarleader	પાટીદાર આંદોલન ને આશ્રય લઇને ગુજરાત રાજનીતિ માં ભૂસકો મારવો એ લાલચી @HardikPatel_. રેલી માં તો આને કહ્યું તું કે તે રાજનીતિ માં નહિ આવે.
Remove mentions	આવિ ગયા લુખાઓ ધમકી ઓ આપવા આખી ફોજ ને છુટી મુકી દીધી છે . ભાઇ લોગ હવેતો માનસો ને કે આપીયાવ કરતા ચમચા સારા હતા કમસે કમ સંસ્કાર હતા . @AAPGujarat @VISHAL_DAVE_ @Gopal_Italia	આવિ ગયા લુખાઓ ધમકી ઓ આપવા આખી ફોજ ને છુટી મુકી દીધી છે . ભાઇ લોગ હવેતો માનસો ને કે આપીયાવ કરતા ચમચા સારા હતા કમસે કમ સંસ્કાર હતા
Remove blank space	કર્ણાટકના બે પૂર્વ મુખ્યમંત્રી સહિત 64 લોકોને જાનથી મારી નાખવાની ધમકી મળી	કર્ણાટકના બે પૂર્વ મુખ્યમંત્રી સહિત 64 લોકોને જાનથી મારી નાખવાની ધમકી મળી
Remove punctuations	રાજસ્થાનના પ્રતાપગઢ જિલ્લામાં બે કોમ વચ્ચે અથડામણ ફાટી નીકળી, ત્રણનાં મરણ, ૬ને ઈજા: &કર્ફ્યૂ લાગુ કરી દેવાયો!........:;*	રાજસ્થાનના પ્રતાપગઢ જિલ્લામાં બે કોમ વચ્ચે અથડામણ ફાટી નીકળી ત્રણનાં મરણ, ૬ને ઈજા કર્ફ્યૂ લાગુ કરી દેવાયો
Remove code mixed text	ધમકી આપવાની ફરિયાદ મામલે સુનાવણી कानून से ऊपर कोई नही !!	ધમકી આપવાની ફરિયાદ મામલે સુનાવણી
Remove stop word	ગજબ દેશ છે મારો, ખેતીપ્રધાન હોવા છતાંય દાળ આયાત કરી રહ્યો છે અને જીવદયા પ્રેમી હોવા છતાંય માંસની નિકાસ કરે છે	ગજબ દેશ છે મારો, ખેતીપ્રધાન હોવા છતાંય દાળ આયાત કરી રહ્યો છે જીવદયા પ્રેમી હોવા છતાંય માંસની નિકાસ કરે

- Supervised learning based approach (23): It requires labeled data while preparing the model. During the training, the model learns features and understands the text. It then classifies using classification algorithms such as Support vector Machine, Decision Tree, Random forest, K Nearest Neighbour, Gaussian Naive Bayes, Logistic Regression.
- Unsupervised learning based Approach (30, 32): For the data without labels, we use an unsupervised learning-based approach. It learns from input text features and groups them with a similar pattern. We have used the k-means clustering algorithm for creating clusters with similar features. We have used 70:30 ratio for training and testing data as shown in Table 5.

5. EXPERIMENTS

There are various machine learning algorithms used for text classification. We have performed various experiments to evaluate the machine learning model for Gujarati hostile post-detection. Amongst different

Machine learning algorithms, selection of a particular algorithm for our work was decided based on experimentation. To understand the behavior of different machine learning algorithms we have experimented using Support Vector Machine, Decision Tree, Random forest, K Nearest Neighbour, Gaussian Naive Bayes, Logistic Regression and k-means clustering algorithms. Posts are classified into hostile or non-hostile classes. In this section we describe the dataset statistics used in our experiments, experiment setup, machine learning methods used for training the model, hyperparameters values and evaluation metrics.

5. 1. Dataset Splitting The dataset contains a total of 10,000 comments to be classified as hostile or non-hostile. We have 5000 hostile and 5000 non hostile posts. We have used 70%-30% for training (7000 instances) and testing (3000 instances). For the supervised machine learning approach, we have used labeled data having text posts and the corresponding label.

5. 2. Experimental Results

5. 2. 1. Implementation Details We implemented a machine learning model using python library scikit-learn, pandas and numpy. We executed the experiments on Google Collaboratory, which provides a free Jupyter notebook environment. Table 6 shows the hyperparameter value of machine learning classifier that we have used for model tuning.

TABLE 5. Statistics of dataset splitting

Dataset	Number of Data
Training Data	7000
Testing Data	3000
Total Data	10000

TABLE 6. Hyperparameter value of classifier

Machine learning classifier	Hyperparameter
SVM	C: 1.0 Kernel: ’rbf’ Degree: 3 Gamma: ’scale’ Coef0: 0 Shrinking: True Probability: False
Decision Tree	Criterion: ‘gini’ Splitter: ‘best’ Max depth: None Min samples split: 2 Min samples leaf: 1 Max features: ‘auto’ Random state: None
KNN	C: 1.0 Kernel: ‘rbf’ Degree: 3 Gamma: ‘Scale’ Coef0:0 Shrinking: True Probability: False
Random Forest	N Estimators:100 Criterion: ‘gini’ Max depth: None Min samples split: 2 Max features:’auto’ Bootstrap: True
Gauusian Naive Bayes	Priors: None Var_smoothing: 1e-9
Logistic Regression	Penalty:’l2’ C:1.0 Solver:’lbfgs’ Max_iter:100 Multi_class:’auto’ Random_state:None Fit_intercept:True
K-means clustering	n_clusters: 2 n_init: 10 max_iter:100 tol: 1e-4 random_state:None algorithm:auto

5. 2. 2. Model Evaluation The performance of the model is evaluated using Accuracy, Macro Precision, Recall and Macro F1-score.

- Accuracy: Accuracy is the ratio between correct prediction and Total Number of given samples
- Macro Precision: Precision is the ratio between the correctly identified positives samples (true positives) and all identified positives samples. We used macro precision that provides arithmetic mean of all the precision values for the both hostile and Non-hostile classes.
- Recall: Recall is the ratio between the true positives and what was actually labeled.
- Macro F1-score: The F1 score is calculated as the arithmetic mean of precision and recall. It is used to find the average rate. The macro-averaged F1 score is the mean of all the individual class F1 scores.

In this section, we discuss the result of supervised and unsupervised machine learning models. The Two datasets used are: (i) Gujarati Text data and (ii) Translated English data. The seven supervised and unsupervised machine learning-based classifiers are used. These classifier results are evaluated using Accuracy, Macro Precision, Recall and Macro F1-score which is shown in Table 7. We have experimented with different scenarios considering Gujarati language and translated into English language datasets. We have also considered cases without Emoji and with Emoji's

TABLE 7. Machine learning based classifier result

Sr.no	Algorithm used	Dataset	Emoji description	Feature extraction	Accuracy	Macro Precision	Recall	Macro F1-score
1	SVM	Gujarati	No	BoW	0.65	0.65	0.67	0.66
2	Decision Tree	Gujarati	No	BoW	0.65	0.66	0.63	0.65
3	KNN	Gujarati	No	BoW	0.59	0.59	0.63	0.61
4	Random Forest	Gujarati	No	BoW	0.65	0.65	0.68	0.66
5	Gaussian Naive-Bayes	Gujarati	No	BoW	0.65	0.65	0.67	0.66
6	Logistic Regression	Gujarati	No	BoW	0.66	0.66	0.67	0.67
7	K-means clustering	Gujarati	No	BoW	0.51	0.51	0.52	0.52
1	SVM	Gujarati	No	TF-IDF	0.66	0.66	0.67	0.67
2	Decision Tree	Gujarati	No	TF-IDF	0.64	0.65	0.65	0.65
3	KNN	Gujarati	No	TF-IDF	0.58	0.58	0.66	0.62
4	Random Forest	Gujarati	No	TF-IDF	0.65	0.65	0.70	0.67
5	Gaussian Naive-Bayes	Gujarati	No	TF-IDF	0.63	0.69	0.47	0.56
6	Logistic Regression	Gujarati	No	TF-IDF	0.68	0.67	0.67	0.67
7	K-means clustering	Gujarati	No	TF-IDF	0.52	0.52	0.53	0.53
1	SVM	Gujarati	Yes	BoW	0.65	0.65	0.67	0.67
2	Decision Tree	Gujarati	Yes	BoW	0.66	0.67	0.64	0.66
3	KNN	Gujarati	Yes	BoW	0.59	0.59	0.63	0.61
4	Random Forest	Gujarati	Yes	BoW	0.65	0.65	0.68	0.66
5	Gaussian Naive-Bayes	Gujarati	Yes	BoW	0.65	0.65	0.67	0.66
6	Logistic Regression	Gujarati	Yes	BoW	0.66	0.66	0.67	0.67
7	K-means clustering	Gujarati	Yes	BoW	0.51	0.51	0.52	0.52
1	SVM	Gujarati	Yes	TF-IDF	0.66	0.66	0.67	0.67
2	Decision Tree	Gujarati	Yes	TF-IDF	0.65	0.66	0.65	0.65
3	KNN	Gujarati	Yes	TF-IDF	0.58	0.58	0.66	0.62
4	Random Forest	Gujarati	Yes	TF-IDF	0.65	0.65	0.70	0.67
5	Gaussian Naive-Bayes	Gujarati	Yes	TF-IDF	0.63	0.69	0.47	0.56
6	**Logistic Regression**	**Gujarati**	**Yes**	**TF-IDF**	**0.68**	**0.67**	**0.67**	**0.67**
7	K-means clustering	Gujarati	Yes	TF-IDF	0.52	0.52	0.53	0.53
1	SVM	English	No	BoW	0.64	0.64	0.68	0.66
2	Decision Tree	English	No	BoW	0.62	0.62	0.62	0.62
3	KNN	English	No	BoW	0.57	0.56	0.70	0.62
4	Random Forest	English	No	BoW	0.64	0.67	0.67	0.65
5	Gaussian Naive-Bayes	English	No	BoW	0.65	0.65	0.67	0.66
6	Logistic Regression	English	No	BoW	0.66	0.66	0.67	0.67
7	K-means clustering	English	No	BoW	0.51	0.51	0.52	0.52
1	SVM	English	No	TF-IDF	0.66	0.65	0.66	0.66
2	Decision Tree	English	No	TF-IDF	0.63	0.63	0.64	0.63
3	KNN	English	No	TF-IDF	0.55	0.53	0.94	0.68
4	Random Forest	English	No	TF-IDF	0.66	0.66	0.69	0.67
5	Gaussian Naive-Bayes	English	No	TF-IDF	0.65	0.65	0.65	0.66
6	Logistic Regression	English	No	TF-IDF	0.66	0.66	0.66	0.67
7	K-means clustering	English	No	TF-IDF	0.52	0.52	0.52	0.53
1	SVM	English	Yes	BoW	0.65	0.65	0.66	0.66
2	Decision Tree	English	Yes	BoW	0.65	0.65	0.67	0.66
3	KNN	English	Yes	BoW	0.58	0.58	0.64	0.61
4	Random Forest	English	Yes	BoW	0.65	0.65	0.68	0.67
5	Gaussian Naive-Bayes	English	Yes	BoW	0.65	0.65	0.67	0.66
6	Logistic Regression	English	Yes	BoW	0.66	0.66	0.67	0.67
7	K-means clustering	English	Yes	BoW	0.51	0.51	0.52	0.52

1	SVM	English	Yes	TF-IDF	0.66	0.66	0.67	0.66
2	Decision Tree	English	Yes	TF-IDF	0.64	0.64	0.66	0.65
3	KNN	English	Yes	TF-IDF	0.59	0.57	0.70	0.63
4	Random Forest	English	Yes	TF-IDF	0.66	0.66	0.69	0.67
5	Gaussian Naive-Bayes	English	Yes	TF-IDF	0.65	0.65	0.67	0.66
6	**Logistic Regression**	**English**	**Yes**	**TF-IDF**	**0.66**	**0.66**	**0.67**	**0.67**
7	K-means clustering	English	Yes	TF-IDF	0.52	0.52	0.53	0.53

description to analyze impact of Emoji. Both methods of feature extraction viz. BoW and TF-IDF with Gujarati dataset having Emoji description as well as without Emoji are evaluated in the Figure 4.

The same evaluation is also carried out for English dataset which is shown in Figure 5. Accuracy performance of the machine learning classifier ranges from 0 to 1. Supervised Learning based classifiers provide good results as compared to unsupervised learning. We have identified that the TF-IDF feature extraction method extracts good features as compared to Bag-of-Words method. Out of all machine learning methods which we have evaluated, Logistic Regression gives better results for the hostile post detection. We have also analyzed that Gujarati data gives good result as compared to the translated English data. The reason could be that the translated English data loses its meaning during conversion because some words are not correctly translated into English. In addition, we have also tried substitution of emoji's appropriate description to understand the sentence in a better way. While analyzing the result we observed that an approach using emoji substitution improves the accuracy. However, the improvement is marginal because only 12% of sentences in our dataset contain emoji.

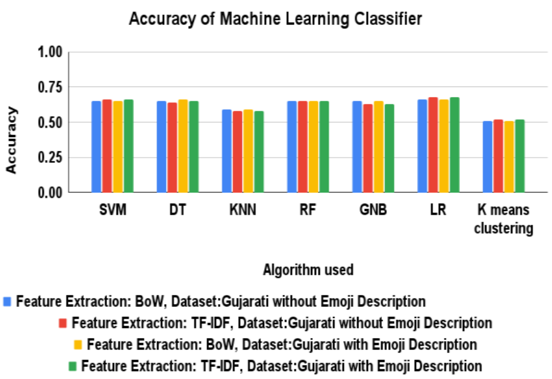


Figure 4. Evaluation of ML approaches for Gujarati data

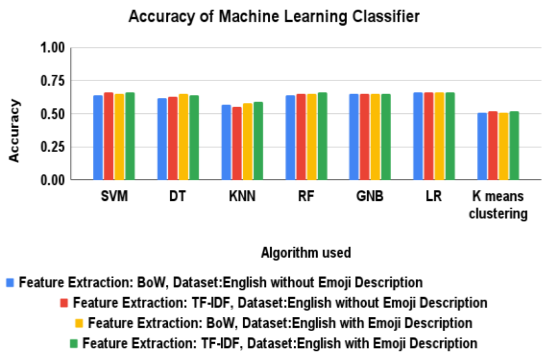


Figure 5. Evaluation of ML approaches for English data

6. RESULTS AND DISCUSSION

We have developed the model for a large dataset and we could achieve the accuracy of 68%. Logistic regression performs well when we use it for two class classification. We are also working on two classes: hostile and non-hostile. Since we are getting comparatively good results for the Logistic regression method (36-39). We have also compared the result of Bag-of-Words and TF-IDF feature extraction methods. We conclude that TF-IDF is more efficient as compared to Bag-of-Words method because BoW provides the frequency of words in a document whereas TF-IDF provides additional data such as how a word is important in the document. Translated English data could not perform well due to inefficient Gujarati text translation. The result of both data shows that there is no wide difference in accuracy. The Gujarati text data provide baseline result 68% and Translated data provide baseline result 66% using TF-IDF feature extraction method.

7. ERROR ANALYSIS

In this section, we have thoroughly analyzed the errors produced by our models to understand limitations and challenges of hostile post detection in Gujarati language. We have described all the challenges we have faced. We have illustrated some of the misclassified input sentences and probable reason for misclassification in Table 8. The result does not get optimal due to many challenges. The challenges are mentioned in the next section:

7. 1. Challenges in Hostile Post Detection in Gujarati Language

- Pre-Processing: There is no standard library

TABLE 8. Error Analysis

Sr. No.	Misclassified Input sentence	Reason for Misclassification
1	આતંકવાદી કયા ધર્મ ના હતા	From one sentence we are unable to predict the actual context of the sentence. It required the full content to understand the sentence's meaning.
2	આતંકવાદી ઓ સે મેરા પુરાના નાતા હે	The sentence contains Hindi words but they are written in Gujarati language. These types of sentences are impossible to understand.
3	આ ફોટો જોઈને કટ્ટર હિન્દૂ વાદ ના જંડા લઈ ને ફરતા હોય એમને વિચાર કરવો સત્તા માટે જ બધા ખેલ હોય છે જેમાં તમે કુદયા કરો સો	Few comments are based on some images or video, so it is difficult to interpret without seeing the associated image.
4	Wow પેટ્રોલ -ડીઝલમાં ભાવમાં ઘટાડો પેટ્રોલમાં 9 અને ડીઝલમાં 6 રૂપિયાનો ઘટાડો ગેસના સિલિન્ડર પર પણ 200રૂપિયાની રાહત જાવ જલસા કરો.	The Gujarati and English mixed sentences are completely not removed from the dataset.

available that perfectly performs the text preprocessing steps in the Gujarati language. Therefore, we have to make regular expressions to preprocess the data.

- Data collection: Data collection is also one challenge for us because normally people write comments in code-mixed language. Therefore, it is difficult to collect texts that are purely written in Gujarati text.
- Manually data annotation for testing Data: Machine learning requires labeled data. It is a huge task to manually label the data.
- No standard method is available: High resource languages such as English have huge resources Dataset, wordnet, preprocessing techniques, feature extraction techniques, automatic data annotation techniques and algorithms for text data understanding. For low-resource Indian languages these resources are not available.
- Important data loss: During pre-processing some meaningful data may be lost. While removing mentions, few natural words are removed from the text. For example, "@Gopal_Italia આમ આદમી પાર્ટી માં કોળી સમાજ નુ મહત્વ કેટલું ?" that is a non-hostile post. But the classifier incorrectly classified it as hostile. The reason behind that is that @Gopal_Italia is removed from the sentence during preprocessing. Another example, "21 Days #પરિવર્તન_યાત્રા અંતર્ગત ગુજરાતની તમામ 182 વિધાનસભામાં પહોંચીને જનતાનાં પ્રશ્નો સાંભળીશું." is also not correctly classified because the English word Days and numeric values are removed in preprocessing.
- Code-mixed text: The data having code-mixed sentences loses its meaning after applying pre-processing. For example, the post," રશિયાએ ભારતીય student ને બહાર કાઢવા યુદ્ધ રોક્યું હતું આ પણ Fake news હતા" was a mixed Gujarati and English language sentence. The English words 'student' and 'Fake news' are removed during preprocessing. Therefore, the meaning of the sentence is not correctly identified. Another example, "Aantakvadi ni pream kahani nathi hoti" was a Gujarati post written in English and thus was not correctly classified.

8. CONCLUSION AND FUTURE WORK

Hostile Post detection in the Gujarati language has become a notable problem therefore there is a huge requirement for automation systems for hostile post detection. In this work, we have developed a Gujarati Text dataset that contains more than one lakh posts having hostile posts and non-hostile posts. The data are collected from social media Twitter, Instagram and Facebook in time duration from the year 2017 to 2022. Total 10,000 data are manually labeled with two major categories having equal number of hostile and non-hostile posts. The data contains unwanted symbols, URLs, duplicates, punctuations and emoticons. Therefore, we performed a data cleaning process and removed unwanted data. Emoji's are often used in the short text for the expression. We have replaced emoji's with their description in the post. We evaluated supervised and unsupervised machine learning models: Support vector Machine, Decision Tree, Random forest, K Nearest Neighbour, Gaussian Naive Bayes, Logistic Regression, and k-means clustering algorithms. The result shows that the Supervised learning-based algorithm Logistic Regression outperforms. There are various challenges identified during the implementation. In this paper, we tried to overcome these challenges. However, better ways to address these challenges are still possible. We believe that our dataset will be beneficial in Gujarati Language Processing related studies.

9. DATA AVAILABILITY

This dataset is not publicly available currently due to the thesis defense has not been completed yet. If you have any query, contact the corresponding author.

10. REFERENCES

1. Bhatnagar V, Kumar P, Bhattacharyya P. Investigating hostile post detection in hindi. Neurocomputing. 2022;474:60-81. 10.1016/j.neucom.2021.11.096
2. Bhardwaj M, Akhtar MS, Ekbal A, Das A, Chakraborty T. Hostility detection dataset in Hindi. arXiv preprint arXiv:201103588. 2020. https://doi.org/10.48550/arXiv.2011.03588
3. Dowlagar S, Mamidi R. Hasocone@ fire-hasoc2020: Using bert and multilingual bert models for hate speech detection. arXiv preprint arXiv:210109007. 2021. https://doi.org/10.48550/arXiv.2101.09007
4. Alshaalan R, Al-Khalifa, H.,, editor Hate speech detection in saudi twittersphere: A deep learning approach. In: Proceedings of the Fifth Arabic Natural Language Processing Workshop; 2020: IEEE. 10.1109/I2CT51068.2021.9418073
5. Joshi R, Karnavat R, Jirapure K, Joshi R, editors. Evaluation of deep learning models for hostility detection in hindi text. 2021 6th International conference for convergence in technology (I2CT); 2021: IEEE. 10.1109/I2CT51068.2021.9418073
6. Bhatnagar V, Kumar P, Moghili S, Bhattacharyya P, editors. Divide and conquer: an ensemble approach for hostile post detection in Hindi. 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; 2021: Springer.
7. Velankar A, Patil H, Gore A, Salunke S, Joshi R. Hate and offensive speech detection in hindi and marathi. arXiv preprint arXiv:211012200. 2021. https://doi.org/10.48550/arXiv.2110.12200
8. Phung TM, Cloos J. An exploratory experiment on Hindi, Bengali hate-speech detection and transfer learning using neural networks. arXiv preprint arXiv:220101997. 2022. https://doi.org/10.48550/arXiv.2201.01997
9. Kamble S, Joshi A. Hate speech detection from code-mixed hindi-english tweets using deep learning models. arXiv preprint arXiv:181105145. 2018. https://doi.org/10.48550/arXiv.1811.05145
10. Velankar A, Patil H, Gore A, Salunke S, Joshi R. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. arXiv preprint arXiv:220313778. 2022. https://doi.org/10.48550/arXiv.2203.13778
11. Glazkova A, Kadantsev M, Glazkov M. Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in english and marathi. arXiv preprint arXiv:211012687. 2021. https://doi.org/10.48550/arXiv.2110.12687
12. Chavan T, Patankar S, Kane A, Gokhale O, Joshi R. A twitter bert approach for offensive language detection in marathi. arXiv preprint arXiv:221210039. 2022. https://doi.org/10.48550/arXiv.2212.10039
13. Kamal O, Kumar A, Vaidhya T, editors. Hostility detection in hindi leveraging pre-trained language models. 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; 2021: Springer.
14. Bhardwaj M, Sundriyal M, Bedi M, Akhtar MS, Chakraborty T. HostileNet: Multilabel Hostile Post Detection in Hindi. IEEE Transactions on Computational Social Systems. 2023. 10.1109/TCSS.2023.3244014
15. Khan MM, Shahzad K, Malik MK. Hate speech detection in roman urdu. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP). 2021;20(1):1-19. https://doi.org/10.1145/3414524
16. Haq NU, Ullah M, Khan R, Ahmad A, Almogren A, Hayat B, et al. USAD: an intelligent system for slang and abusive text detection in PERSO-Arabic-scripted Urdu. Complexity. 2020;2020:1-7. https://doi.org/10.1155/2020/6684995
17. Anbukkarasi S, Varadhaganapathy S. Deep learning-based hate speech detection in code-mixed Tamil text. IETE Journal of Research. 2023;69(11):7893-8. https://doi.org/10.1080/03772063.2022.2043786
18. Farooqi ZM, Ghosh S, Shah RR. Leveraging transformers for hate speech detection in conversational code-mixed tweets. arXiv preprint arXiv:211209986. 2021. https://doi.org/10.48550/arXiv.2112.09986
19. Biradar S, Saumya S, editors. Iiitdwd@ tamilnlp-acl2022: Transformer-based approach to classify abusive content in dravidian code-mixed text. Proceedings of the second workshop on speech and language technologies for Dravidian languages; 2022. 10.18653/v1/2022.dravidianlangtech-1.16
20. Mohapatra SK, Prasad S, Bebarta DK, Das TK, Srinivasan K, Hu Y-C. Automatic hate speech detection in english-odia code mixed social media data using machine learning techniques. Applied Sciences. 2021;11(18):8575. https://doi.org/10.3390/app11188575
21. Nayak R, Joshi R. Contextual hate speech detection in code mixed text using transformer based approaches. arXiv preprint arXiv:211009338. 2021. https://doi.org/10.48550/arXiv.2110.09338
22. Sreelakshmi K, Premjith B, Soman K. Detection of hate speech text in Hindi-English code-mixed data. Procedia Computer Science. 2020;171:737-44. https://doi.org/10.1016/j.procs.2020.04.080
23. Luo X. Efficient English text classification using selected machine learning techniques. Alexandria Engineering Journal. 2021;60(3):3401-9. https://doi.org/10.1016/j.aej.2021.02.009
24. Sanoussi MSA, Xiaohua C, Agordzo GK, Guindo ML, Al Omari AM, Issa BM, editors. Detection of hate speech texts using machine learning algorithm. 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC); 2022: IEEE. 10.1109/CCWC54503.2022.9720792
25. Felber T. Constraint 2021: Machine learning models for COVID-19 fake news detection shared task. arXiv preprint arXiv:210103717. 2021. https://doi.org/10.48550/arXiv.2101.03717
26. Fahad N, Goh KM, Hossen MI, Shopnil KS, Mitu IJ, Alif MAH, et al. Stand up against bad intended news: An approach to detect fake news using machine learning. Emerging science journal. 2023;7(4):1247-59. 10.28991/ESJ-2023-07-04-015
27. Defersha N, Tune K. Detection of hate speech text in afan oromo social media using machine learning approach. Indian J Sci Technol. 2021;14(31):2567-78. 10.17485/IJST/v14i31.1019
28. Badjatiya P, Gupta S, Gupta M, Varma V, editors. Deep learning for hate speech detection in tweets. Proceedings of the 26th international conference on World Wide Web companion; 2017.
29. Bangyal WH, Qasim R, Rehman NU, Ahmad Z, Dar H, Rukhsar L, et al. Detection of fake news text classification on COVID-19 using deep learning approaches. Computational and mathematical methods in medicine. 2021;2021:1-14. https://doi.org/10.1155/2021/5514220
30. Aggarwal CC, Zhai C. A survey of text classification algorithms. Mining text data. 2012:163-222. https://doi.org/10.1007/978-1-4614-3223-4_6
31. Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, et al. A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint

arXiv:170702919. 2017. https://doi.org/10.48550/arXiv.1707.02919

32. Akram MW, Salman M, Bashir MF, Salman SMS, Gadekallu TR, Javed AR. A novel deep auto-encoder based linguistics clustering model for social text. Transactions on Asian and Low-Resource Language Information Processing. 2022. https://doi.org/10.1145/3527838
33. Qasim R, Bangyal WH, Alqarni MA, Almazroi AA. A fine-tuned BERT-based transfer learning approach for text classification. Journal of healthcare engineering. 2022;2022. https://doi.org/10.1155/2022/3498123
34. Aluru SS, Mathew B, Saha P, Mukherjee A. Deep learning models for multilingual hate speech detection. arXiv preprint arXiv:200406465. 2020. https://doi.org/10.48550/arXiv.2004.06465
35. Hassan SU, Ahamed J, Ahmad K. Analytics of machine learning-based algorithms for text classification. Sustainable operations and computers. 2022;3:238-48. https://doi.org/10.1016/j.susoc.2022.03.001
36. Indrawan G, Setiawan H, Gunadi A. Multi-class svm classification comparison for health service satisfaction survey data in bahasa. HighTech and Innovation Journal. 2022;3(4):425-42. 10.28991/HIJ-2022-03-04-05
37. Balamurugan V, Vedanarayanan V, Sahaya Anselin Nisha A, Narmadha R, Amirthalakshmi T. Multi-label Text Categorization using Error-correcting Output Coding with Weighted Probability. International Journal of Engineering, Transactions B: Applications. 2022;35(8):1516-23. 10.5829/ije.2022.35.08b.08
38. Dorrani Z. Traffic Scene Analysis and Classification using Deep Learning. International Journal of Engineering, Transactions C: Aspects. 2024;37(3):496-502. 10.5829/IJE.2024.37.03C.06
39. Zare F, Mahmoudi-Nasr P. Feature Engineering Methods in Intrusion Detection System: A Performance Evaluation. International Journal of Engineering, Transactions A: Basics. 2023;36(7):1343-53. 10.5829/ije.2023.36.07a.15
40. Banerjee S, Sarkar M, Agrawal N, Saha P, Das M. Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages. arXiv preprint arXiv:211113974. 2021. https://doi.org/10.48550/arXiv.2111.13974
41. Warjri S, Pakray P, Lyngdoh SA, Maji AK, editors. Fake news detection using social media data for Khasi language. 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC); 2023: IEEE. 10.1109/ISACC56298.2023.10083518

Persian Abstract

چکیده

پست خصمانه در رسانه های اجتماعی یک موضوع حیاتی برای افراد، دولت ها و سازمان ها است. نیاز اساسی به یک سیستم خودکار وجود دارد که بتواند پست‌های متخاصم را از داده‌های مقیاس بزرگ بررسی و شناسایی کند. در هند، گجراتی ششمین زبان پرگویش است. در این کار، ما یک مجموعه داده پست خصمانه اصلی به زبان گجراتی ساخته‌ایم. داده ها از توییتر، اینستاگرام و فیس بوک جمع آوری شده است. مجموعه داده ما شامل ۱۵۱۰۰۰ نظر مجزا است که ۱۰۰۰۰ پست مشروح دستی دارند. این پست ها در دسته های خصمانه و غیر خصمانه برچسب گذاری شده اند. ما از مجموعه داده به دو صورت استفاده کرده‌ایم: (i داده‌های متن اصلی گجراتی) و (ii داده‌های انگلیسی ترجمه شده از متن گجراتی). همچنین عملکرد پیش پردازش و بدون پیش پردازش داده ها را با حذف نمادهای اضافی و جایگزینی توضیحات ایموجی در متن بررسی کرده ایم. ما آزمایش‌هایی را با استفاده از مدل‌های یادگیری ماشین مبتنی بر یادگیری نظارت‌شده مانند ماشین بردار پشتیبان، درخت تصمیم، جنگل تصادفی، گاوسی ساده‌لوح، رگرسیون لجستیک، -Kنزدیک‌ترین همسایه و مدل مبتنی بر یادگیری بدون نظارت مانند خوشه‌بندی k-means انجام داده‌ایم. ما عملکرد این مدل‌ها را برای روش‌های استخراج ویژگی Bag-of-Words و TF-IDF ارزیابی کرده‌ایم. مشاهده می شود که طبقه بندی با استفاده از ویژگی های TF-IDF کارآمد است. در میان این روش ها رگرسیون لجستیک با دقت ۰.۶۸ و F1-score 0.67 عملکرد بهتری دارد. هدف از این تحقیق ایجاد یک مجموعه داده معیار و ارائه نتایج پایه برای تشخیص پست های خصمانه در زبان گجراتی است.