International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | Jun 2022 www.irjet.net p-ISSN: 2395-0072

Hate Speech Identification Using Machine Learning

Nikhilraj Gadekar[1], Mario Pinto [2]

[1]Student, Information Technology Department, Goa College of Engineering, India
[2]Assistant Professor, Information Technology Department, Goa College of Engineering, India

---

Abstract - *With the rise of social media people have the liberty to express themselves, many users misuse this liberty and we can see abuse and hate spread all across social media platforms, be it in the form of comments, blogs, etc. Machine Learning is widely used in various fields of research, we intend to use machine learning to automate the detection of hate speech. In this paper, we have used subjectivity analysis and semantic features to create a lexicon that builds a classifier to identify hate speech.*

Key Words: Hate Speech, Hostile, Subjectivity Analysis, Lexicon, Machine Learning, Cyber-bullying

1. INTRODUCTION

With a population of 1.40 billion people which is also increasing at a very steady pace day by day, India has the second-largest population in the world. India also boasts of having one of the fastest-growing economies in the world which means it becomes a very important playground for social media companies.

The people of India spend about 2.36 hours of their everyday time on social media on average. The country's internet penetration rate stands at 47%, and it has 467 million social media users as of 2022 and that number will only grow with time.

The world is shrinking with the use of the internet but with the positives also come the negatives, and one of the major cons of social media is Hate speech which can also be called abusive writing, cyberbullying, etc. The National Crime Records Bureau figures show a 36% increase in cyberstalking and cyberbullying cases in India post the pandemic. Around 9.2% of 630 adolescents surveyed in the Delhi-National Capital Region had experienced cyberbullying and half of them had not reported it to teachers, guardians, or the social media companies concerned, a recent study by Child Rights and You, a non-governmental organization, found.

The manual process to identify hate speech is slow, tedious, and labor-intensive. Therefore automatic hate speech detection becomes very important.

Despite Hindi being the third most spoken language in the world, and a significant presence of Hindi content on social media platforms we couldn't find much work done to detect hate speech using technology.

Twitter is the 3rd most widely used social media platform in India with 295.44 million active users. Political and social issues discussed on Twitter are heavily polarizing as people are very attached to their political ideology the threads on Twitter discussing these issues seem to have a lot of hate in them.

In this paper the dataset that we have used deals with tweets that are majorly political and social, the dataset has been segregated into two major categories which are hostile and non-hostile, the hostile category has been further divided into Non-hostile, fake, defamation, hate and offensive.

In our research, we propose a model that can successfully classify if the tweets are fake, defamation, hate, offensive and non-hostile. We use a rule-based approach which has three major steps I Subjectivity Analysis II Building hate speech lexicon III Identifying theme-based nouns

The other part of the paper is organized as follows: In sections 2 and 3 related works and methodology have been elaborated. In section 4 dataset details have been discussed and in section 5 the experimental results have been done. The research paper is terminated in section 6.

2 Related Works:

A lot of work on hate speech detection on social media platforms has been done in the past in various languages like English and many other Western languages, but very little work has been done in this area in low-resource languages like Hindi which is the 3rd largest spoken language in the world. The author of the paper [1] has collected 197,566 comments from four social media platforms which are YouTube, Reddit, Wikipedia, and Twitter with 80% of the comments labeled as non-hateful and the remaining labeled as hateful. The author has used classification algorithms like Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machines (SVM), Extreme Gradient Boosting (XGBoost), and Fast Feed Neural Network (FNN), the author has used various Feature Engineering techniques which are Simple features, BOW, TF-IDF, Word2Vec and BERT in combination with these classification algorithms and found that XGBoost with BERT gives the best accuracy. The study in [2] states that the author has collected the dataset from Facebook and in the Bangla language and has labeled the data as Hate or Non-Hate, the authors have used TF-IDF for extracting the features and have used SVM and Naïve Bayes algorithms for classification. The authors achieve an accuracy of 70% and 72% using SVM and Naïve Bayes respectively. The research in [3] presents an approach to detect offense in memes using Natural Language

Processing (NLP) and Deep Learning, the model which the authors have used is majorly divided into three steps 1) Extracting the text from the image. 2) Classify the text as offensive or non-offensive 3) further classifies it into slight offensive, very offensive, and hateful offensive. The model uses simple architecture with a multi-layer dense network structure involving NLP with RNN and LSTM along with word embeddings such as GloVe and FastText. [4] In this paper, the authors have presented a social engineering attack detection method that is based on NLP and machine learning. The proposed method has been evaluated using a semi-synthetic dataset and well-known metrics along with a comparison to a state-of-the-art alternative method. It has been shown that it performs well and outperforms the alternative methodology.

3 Methodology:

The main task that we have is to create a lexicon of sentiment expressions using semantic and subjectivity features relative to hate speech and use these features to make a classifier that will detect hate speech

Our approach consists of three major steps,

A. Subjectivity Analysis
B. Building a lexicon with hate-related words
C. Theme-Based Grammatical patterns

A. Subjectivity Analysis

We consider that the tweets which are only subjective will have hate instead of facts and thus we mark the objective tweets in the dataset as "not hate". To do this task successfully we use SentiWordNet and assign scores for the subjectivity of each word. The key in the lexicon is now matched with our corpus and the score is assigned to every word in our corpus. The total score of a tweet or sentence is then calculated by adding the scores of all the keys in the respective tweet or sentence. A sentence is considered positive if it has an average score of 0.5 or above while a score of -0.5 and below makes the sentence negative. The sentence is considered objective otherwise.

B. Building a lexicon with hate-related words

Now to build the lexicon we first identify opinionated words from the subjective sentences which were identified previously. All the words which are strongly and weakly subjective and have a polarity tag of negative are extracted.

We now include the "hate" verbs that are not a part of the lexicon, we extract all the verbs which have a relation with the hate words from our dataset. We have then used an initial list of seed verbs consisting of manually selected verbs that were not a part of the lexicon, we then use SYSNET to find all the synonyms of these verbs and add them to the list of hate words. The "hate-word" list is now matched with the corpus and all the matched words from the corpus are then added to the final hate lexicon.

C. Theme Based Grammatical Patterns

To have metaphors that represent non-literal meaning rather than the literal meaning, only the hate words cannot be used to directly conclude the meaning of the sentences. We, therefore, use grammatical patterns to represent the subjective expression.

To get this done we divide our corpus into noun phrases and note the most frequently used noun phrases that are used in the tweets marked as hateful.

Having generated all the lexicons that we need we can now use them on our corpus to automatically identify hate speech in our corpus. We now have two options 1) We can run the entire algorithm once and analyze the result 2) We can run each lexicon separately and then integrate the entire system. We choose the second option since this will help us analyze each lexicon in a much easier and better way, we can also gauge the role played by each lexicon in the classification of hate speech.

We use well-defined criteria to segregate the tweets by the hateful content present

- If the tweet has two or more words qualified as strongly negative, it is marked as strongly hateful. If it has one strongly negative word with non-zero hate-verbs or non-zero theme-based nouns, it is marked as strongly hateful. If it has at least one word each from our hate-verbs lexicon and themed nouns, it is likewise marked as strongly hateful.
- If it has one strongly negative word, but no other word from our other lexicons, it is weakly hateful. If it contains one weakly hateful word with a theme-based noun, it is weakly hateful. If it contains neither of the above but contains a hate verb, it is also weakly hateful.
- If the tweet satisfies neither of these criteria, it is judged as non-hateful.

4 Dataset Details:

The dataset that we have used has 8192 tweets, it is divided into Non-hostile, fake, defamation, hate, and offensive, of which 4358 belong to the non-hostile category while the rest 3834 tweets belong to the remaining categories.

We define these categories as

Fake News: A claim or information that is verified to be not true. We have included tweets belonging to clickbait and satire/parody categories as fake news as well.

Hate Speech: A post targeting a specific group of people based on their ethnicity, religious beliefs, geographical belonging, race, etc., with malicious intentions of spreading hate or encouraging violence.

Offensive: A post containing profanity, impolite, rude, or vulgar language to insult a targeted individual or group.

Defamation: A miss-information regarding an individual or group, which is destroying their reputation publicly.

Non-Hostile: A post with no hostility.
The categories are not mutually disjoint, hence some of the tweets have multiple categories instead of just one.

Fig -1: Hostile Word Cloud

Fig -2: Non-Hostile Word Cloud

The figure below shows category-wise overlaps for the hostile dimension of the dataset.

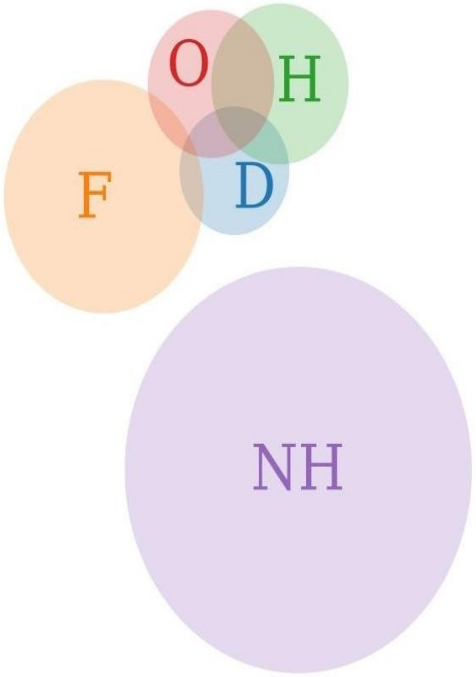

Fig -3: Venn diagram of Multi-Label Hindi hostility dataset. Notations: [F- Fake], [O- Offensive], [H- Hate], [D- Defamation], [NH- Non-hostile].

The table below shows a few annotated examples of the dataset. It can be seen from the table that the 1st and the last tweets belong to multiple categories.

#	Text	Labels
1.	देख लो हिन्दुस्ता के लोगो आज तिरंगा जला रहे है स##ले कल लोगो को जलाऐंगे.. सुधर जाओ और मोदी जी का साथ दो.. नही तो बहुत रोना पडेगा...। Dekh lo Hindustan k logo aaj tiranga jala rhe h kl s##le logo ko jalaenge.. sudhar jao aur modi ji ka sath do.. nahi to bahut rona pdega Listen Indians, they are burning Indian flag today, ras##ls will burn people tomorrow.. Beware and support Modi ji now or else you will cry later	Fake, Hate, Offensive
2.	गृह मंत्री अमित शाह बोन कैंसर से जूझ रहे हैं Greh Mantri Amit Shah bone cancer se jhuj rhe hai Home minister Amit Shah is suffering from bone cancer	Fake
3.	डॉक्टर कफ़ील खान को हाईकोर्ट से मिली ज़मानत Dr. Kafeel Khan ko high court se mili zamanat Dr. Kafeel Khan gets bail from high court	Non-hostile
4.	चीन ने 500cr में कांग्रेसी कुक्ते, सिब्बल टुकड़े गैंग के प्रशांत भूषण को खरीदा ताकि टिकटॉक बैन के खिलाफ सुप्रीम कोर्ट में केस लडें Cheen ne 500 cr me congresi ku##e, sibbal tukde gang k prashant Bhushan ko khareeda taaki tik-tok ban k khilaaf course se case lade China bought congress's dog: Prashant Bhushan, member of Sibbal Tukde gang, for 500 Crores to fight against Tik-Tok ban.	Defamation, Offensive

Table -1: A few annotated examples from the dataset.

	Hostile					Non-Hostile
	Fake	Hate	offense	Defame	Total	
Train	1144	792	742	564	2678	3050
Validation	160	103	110	77	376	435
Test	334	237	219	169	780	873
Overall	1638	1132	1071	810	3834	4358

Table -2: Dataset statistics and Label distribution. Fake, hate, defame, and offense reflect the number of respective posts including multi-label cases. _ denotes total hostile posts.

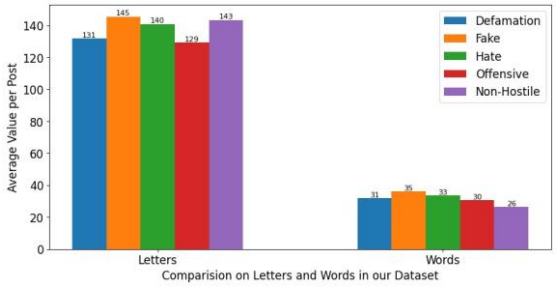

Fig -4: Average number of characters and words per post.

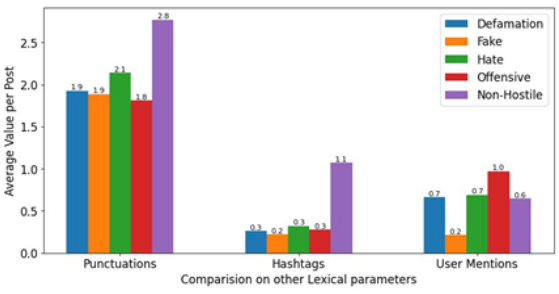

Fig -5 Average Punctuations (— , : ? " ; !), Hashtags, and User Mentions per post.

5 Experimental Results:

Instead of using criteria in the corpus, we classify the tweets as Strongly Hateful, Weakly Hateful, or Not hate. We have considered all the tweets classified as non-hostile by the algorithm as No Hate, All the tweets marked as fake or defamatory are marked as Weakly Hateful, The tweets that are marked hate, Offensive, or a combination of one or more tags correspond to Strongly Hateful. The following parameters are used to evaluate the performance of the classifier.

Precision: Precision is the ratio between the True Positives and all the Positives (Addition of True positives and False positives).

$$Precision = \frac{True\ Positive\ (Tp)}{True\ Poritive\ (TP) + False\ Positive\ (FP)}$$

Recall: The recall is the measure of our model correctly identifying True Positives.

$$Recall = \frac{True\ Positive\ (Tp)}{True\ Poritive\ (TP) + False\ Negative\ (FN)}$$

F1-Score: F1 is a function of Precision and Recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Table -2: Results Obtained without Subjectivity Analysis:

Feature Set	Precision (%)	Recall (%)	F1-Score (%)
Semantic (weakly/strongly negative)	39.33	77.99	52.29
Semantic + hate verbs	39.35	78.03	52.32
Semantic + hate verbs + nouns	42.54	84.35	56.55

Table -3: Results Obtained without Subjectivity Analysis:

Feature Set	Precision (%)	Recall (%)	F1-Score (%)
Semantic (weakly/strongly negative)	45.37	84.98	59.16
Semantic + hate verbs	45.49	85.02	59.27
Semantic + hate verbs + nouns	48.58	91.94	63.54

6 CONCLUSION

The detection of hate speech in languages such as Hindi with a wide user base is getting increasingly relevant in this digital age. Our model uses a rule-based approach to look for lexical patterns that allow us to detect and quantify hate speech to a reasonable precision. An important function here was played by the analysis of subjectivity in our algorithm, accounting for which made a significant improvement in the accuracy of hate-speech detection.

REFERENCES

[1] Salminen, Joni & Hopf, Maximilian & Chowdhury, Shammur & Jung, Soon-Gyo & Almerekhi, Hind & Jansen, Jim. (2020). Developing an online hate classifier for multiple social media platforms. 10. 1. 10.1186/s13673-019-0205-6.

[2] S. Ahammed, M. Rahman, M. H. Niloy and S. M. M. H. Chowdhury, "Implementation of Machine Learning to Detect Hate Speech in Bangla Language," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), 2019, pp. 317-320, doi: 10.1109/SMART46866.2019.9117214.

[3] R. K. Giri, S. C. Gupta and U. K. Gupta, "An approach to detect offence in Memes using Natural Language Processing(NLP) and Deep learning," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-5, doi: 10.1109/ICCCI50826.2021.9402406.

[4] M. Lansley, S. Kapetanakis and N. Polatidis, "SEADer++ v2: Detecting Social Engineering Attacks using Natural Language Processing and Machine Learning," 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2020, pp. 1-6, doi: 10.1109/INISTA49547.2020.9194623.

[5] S. Chiramel, D. Logofătu and G. Goldenthal, "Detection of social media platform insults using Natural language processing and comparative study of machine learning algorithms," 2020 24th International Conference on System Theory, Control and Computing (ICSTCC), 2020, pp. 98-101, doi: 10.1109/ICSTCC50638.2020.9259730

[6] S. Mussiraliyeva, M. Bolatbek, B. Omarov, Z. Medetbek, G. Baispay and R. Ospanov, "On Detecting Online Radicalization and Extremism Using Natural Language Processing," 2020 21st International Arab Conference on Information Technology (ACIT), 2020, pp. 1-5, doi: 10.1109/ACIT50332.2020.9300086