# A Practical Approach to Detect Hateful and Non-hateful Language

Dr. D. Sameera
*Dept. of IT*
*B V Raju Institute of Technology*
Narsapur, Medak(dist),Telangana
sameera.d@bvrit.ac.in

A. Greeshma
*Dept. of IT*
*B V Raju Institute of Technology*
Narsapur, Medak(dist),Telangana
20211a1205@bvrit.ac.in

B. Kanakathara
*Dept. of IT*
*B V Raju Institute of Technology*
Narsapur, Medak(dist),Telangana
20211a1218@bvrit.ac.in

A. Sai Vineel Reddy
*Dept. of IT*
*B V Raju Institute of Technology*
Narsapur, Medak(dist),Telangana
20211a1202@bvrit.ac.in

2024 10th International Conference on Communication and Signal Processing (ICCSP) | 979-8-3503-5306-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCSP60870.2024.10543682

***Abstract*—Hate speech is a big problem in today's online world due to the rise of social media and internet communities. Researchers are coming up with novel approaches to recognize and suppress hate speech using language processing tools. Hate speech has the potential to incite violence and prejudice in the real world as well as create a negative online atmosphere among communities. Owing to the massive amount of data, automated systems employing NLP techniques are essential in recognizing and removing this type of information. We will discuss the advancements in Natural Language processing (NLP) recently for hate speech identification and include anecdotes from our own research projects. We look at the difficulties in detecting hate speech and discuss how cutting-edge NLP methods can help. We also report the outcomes of our experiments using NLP algorithms to identify hate speech.**

***Keywords—Hate speech, NLP (Natural Language Processing) techniques, social media***

## I. INTRODUCTION

In current world, the problem of hate speech has significantly grown in the fast pace world. Social media has become a vast variety of communication systems where one could share their own views and interact with others. Social media platforms have a good and bad. That is though it is used positively in promoting information sharing and communication by connecting different people from different places, it is also becoming adverse in case of bad commenting or hate speech.[1]Online hate speech not only undermines the principles of equality and inclusivity, but it also perpetuates discrimination and violence. Given the pervasive nature of hate speech online, it has become crucial to develop effective means of identifying and combating it.

In this paper, we will explore innovative natural language processing techniques that have shown promise in detecting and filtering out hate speech. By leveraging NLP models, we can analyse large volumes of textual data from online platforms and communities to identify harmful content and take proactive measures to address it. The purpose of this research project is to investigate alternative approaches for creating a prototype that can recognize cyberbullying on social media sites automatically. This study adds to the body of knowledge on online harassment and cyberbullying detection. The number of user comments that contain cyberbullying language is rising. This could contribute to the platform's efforts to restrict such content and encourage the kind of constructive discourse that the community platform was intended to foster. The opportunity to express one's own opinions on an issue has altered due to the abundance of public forums available online.

The rising challenges in hate speech detection call for advanced NLP strategies that can adapt to the evolving nature of online language use. Furthermore, our research will highlight the experimental results of employing NLP models to detect hate speech, shedding light on the effectiveness and potential limitations of these techniques. Through our study, we aim to contribute to the ongoing efforts in combating hate speech and fostering a safer digital environment for all users.

## II. PROBLEM STATEMENT

The challenges faced in hate speech detection using NLP techniques are twofold. First, there is a lack of benchmark datasets and guidelines for data annotation in languages other than English. This makes a significant challenge in detecting hate speech for different linguistic communities. Second, hate speech is a complex and dynamic phenomenon that constantly evolves and adapts to new forms of online communication. Therefore, an approach to detect different languages hated speech is essential to address this problem wholly. And, the different languages used on social media platforms extends another layer of complexity in detecting accurate results for hate speech. Moreover, the contextual meaning of language used by the people further complicates the issue. On top of that, the existing algorithms, NLP tools which are used for English datasets may not be applicable or effective for the other languages. Therefore, there is a need for research and development of hate speech detection models that are specifically tailored to different languages, considering the linguistic nuances and cultural context of each language.

This significant issue was not previously thought to be a research topic due to its lack of severity, but it is now in a dangerous phase. On the cyber platform, this effect cannot be disregarded. To manage this activity, researchers and cybercrime organizations need to give it significant consideration.

## III. LITERATURE REVIEW

The literature review section will provide an overview of existing research on hate speech detection using natural language processing techniques. It will explore studies from various papers that have used natural language processing techniques to tackle the problem on hate speech. The previous papers used different methodologies to detect hate speech, including machine learning algorithms, deep learning models, and keyword-based approaches. Some papers have focused on the creation of annotated datasets for hate speech detection, while others have explored the use of word embeddings and language models to capture the main meaning of hate speech.

The primary goal of the paper [1] was to present an automated system that can identify hate speech on social media sites. Although there have been advancements in the identification of hate speech online, automated systems for identifying abusive language and cyberbullying have not yet been implemented. Thus, to automatically distinguish between hateful and non-hateful material, they devised a model that makes use of NLP and machine learning [10].

Some authors have used techniques such as Support Vector Machine algorithm in detecting hateful and nonhateful speech using NLP and Machine Learning algorithms. These algorithms have shown promising results in identifying hate speech accurately but are limited to English datasets [2][9]. Another approach discussed here is the use of decision tree algorithm from Sklearn package. The study depicts that the multiple decision trees give more efficient results for detecting hate speech. However, it is important to note that hate speech detection is not solely limited to the English language [2].

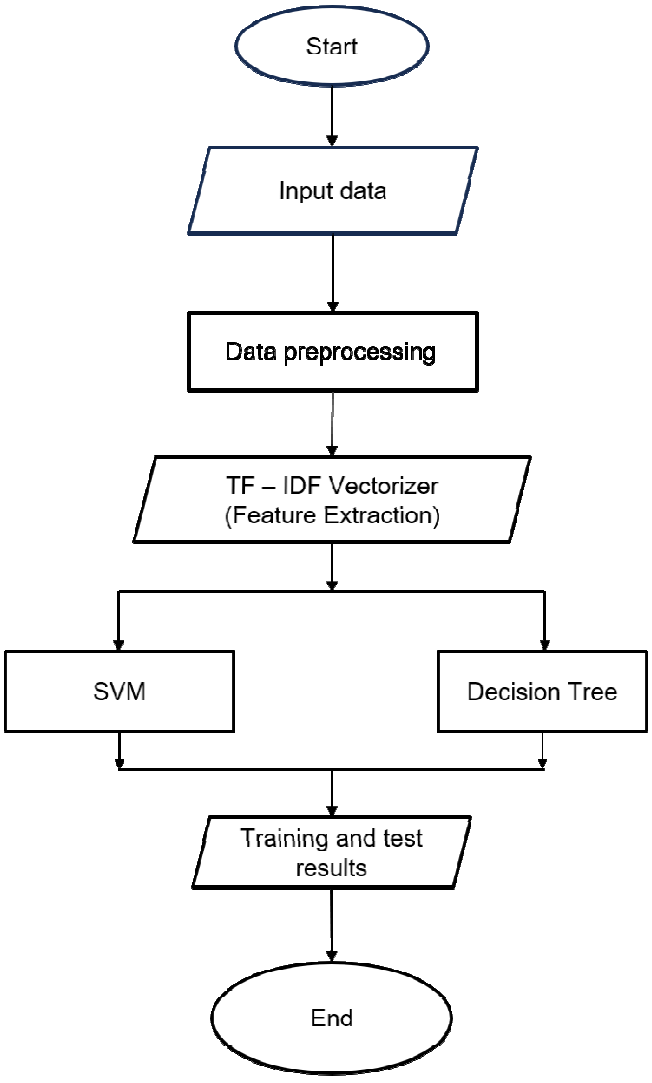
Some other previous researches shown that using deep learning models like bi-LSTM or LSTM also gives a good performance metrics for the detection. The evolution gives new good healthy algorithms to model the machine in a better way. The paper shows the effectiveness of using deep learning models and NLP combined [3][4][5].

Some papers also provide a survey and discuss the valuable resources providing how neural networks are well performed in hatespeech detection [6][10]. However there are some other authors who used different techniques such as Random Forest and Gradient Boosting Machine methods which makes some effective results other than svm and decision tree [8][11].

Many previous research papers gave approaches and solutions from simple machine learning algorithms to complex deep learning algorithms and neural networks. These methods have given drastic and significant improvements in detection of hate speech accuracy and performance. But the problem with the present approaches is that they have developed and depicted results on English datasets neglecting the other languages.

## IV. PROPOSED SYSTEM

To achieve this, we will develop a model that combines machine learning algorithms and language analysis techniques to detect the hate speech comments. The NLP methodology preprocesses the dataset by tokenising text, removing stop words and do performs stemming to normalise the text. The model will extract the relevant features from the normalised text using techniques such as TF-IDF and Bag of words. The extracted features are used to train the model using different classification algorithms such as Support Vector Machine, Decision Tree, etc. Evaluation metrics such as accuracy, precision, recall, and F1 score will be used to evaluate the performance of the model. The evaluated results will be used to determine the effectiveness of algorithm and the accuracy of hate speech detection for Telugu English code-mixed language can be improved.




Flow diagram

*Techniques used:*

## Support Vector Machine

Support Vector Machine (SVM) is known to be a popular supervised machine learning algorithm which is effectively used for classification and regression tasks. The advantages of SVM are it can handle linear and non-linear data. There are many other uses where SVM is advantageous, it works effectively in high dimensional spaces and, they are used in kernel functions for non-linear problems. In our proposed system, SVM is used as a classifier known as Support Vector Classifier. SVC does the classification in 2 ways, binary and multiple based on the data. Here, SVC converts the data into hateful and non-hateful content i.e. binary classification is performed.[9]

## DECISION TREE

Decision Tree Algorithm is also a key algorithm in machine learning algorithm used for classification. It works like a tree model consisting of observations and conclusions. A decision tree with nodes and branches is created based on the dataset given. The decision tree is a non-parametric and greedy algorithm, it doesn't need many assumptions, it handles numerical as well as categorical data. It is easy to understand as it is a tree like structure. In our proposed system, decision tree classifier is used to decide if the comment is based on hate speech or not.[2]

## TF-IDF

Term Frequency Inverse Document Frequency is a vectorizer used in NLP which counts word frequency and gives the measure of importance to each word in the document. It is an NLP technique which is carried out after text cleaning process. Through this the machine extracts features from the normalised data. In our paper, TF-IDF is used for feature extraction and it represents the words in document.[2]

**Methodology:**

The process of detecting hate speech using NLP techniques involves several steps.

*A. Data Acquisition and preprocessing :*

To collect the data, we have gone through all the social media where code-mixed Telugu English comments are used. On the social media platforms like X (old name Twitter), Instagram, Facebook, etc. many use these kinds of linguistic comments. We have collected a dataset which consisting of 4000 code-mixed comments which are annotated with hate speech or non-hate speech labels. These datasets were found by various other sources over the internet such as Kaggle websites etc. After collecting the dataset, data is analysed and pre-processed to prepare data for training process.

| | S.No | Comments | Label |
|---|---|---|---|
| 0 | HATE_1001 | Thappu chesina vaallaku vanike kaadu inka anni... | hate |
| 1 | HATE_1002 | Dhusta chaathuryam! Meeru ilantivi enni chesi... | hate |
| 2 | HATE_1003 | Vetakaram super. Govt ki siggu seram radu. End... | hate |
| 3 | HATE_1004 | Only rajakiyam ga vadukovatanike ee dharidrapu... | hate |
| 4 | HATE_1005 | Katam hogaya narayana pedda bokada college | hate |

**Data collection**

*B. Text cleaning:*

Data is pre-processed to remove duplicates or irrelevant entries. NLP toolkit called as NLTK is imported to preprocess the words to perform cleaning tasks. Cleaning text involves converting the whole text into lowercase, removal of unnecessary special characters and punctuation marks from the data, removing irrelevant data such as stop words and URLs from comments, and performing tokenization to split the text into individual words.

*C. Feature extraction:*

The next step is featuring extraction. In this step, cleaned data is converted into new representations. This can be achieved through techniques such as TF-IDF to represent the importance of words in a document, word embeddings to capture the semantic meaning of words. Cleaned text data is converted into numerical format using word to vector format. A new column is added to the dataset for these representations.

*D. Classification using different algorithms:*

Different algorithms such as Support Vector Machine, Decision Tree are used for classification of hate speech detection. Support Vector Machine uses SVC Support Vector Classifier where as Decision Tree uses Decision Tree Classifier.

*E. Model training and evaluation:*

After feature extraction, the data is split into training and testing sets. From the dataset, two-thirds of the data is used for training the model, and one-third is used for testing the model. During training, machine learns from the labelled data and analyses the text to classify the testing set. In evaluation phase, evaluation metrics is performed and precision, recall and F1 score can be calculated to assess the model.

## V. RESULTS AND DISCUSSION

The results of Hate speech detection are assessed using evaluation metrics and accuracy, precision, and recall, F1 score are calculated. The accuracy achieved shows that the proposed system gives better results for the problem discussed in the paper.

For Decision Tree Classifier:

Training accuracy: 0.991044776119403
Testing accuracy: 0.6636363636363637

For SVM Classifier:

Training accuracy: 0.9753731343283583
Testing accuracy: 0.7143939393939394

| Model | Class | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree Classifier | 0 | 0.644345 | 0.678683 | 0.661069 |
| Decision Tree Classifier | 1 | 0.683642 | 0.649560 | 0.666165 |
| Support Vector Machine Classifier | 0 | 0.677551 | 0.780564 | 0.725419 |
| Support Vector Machine Classifier | 1 | 0.760684 | 0.652493 | 0.702447 |

Class 0 represents non-hate class, Class 1 represents hate class

## VI. CONCLUSION AND FUTURE WORK

In conclusion, the process of hate speech detection using NLP techniques involves data collection, cleaning, and preprocessing to remove duplicates and irrelevant entries. The data is vectorised and training, testing phases are done. Different types of classification algorithms are performed and evaluations are done. SVM Algorithm provided more accurate results than Decision tree Algorithm. The proposed system gave accurate results for the performed hate speech detection and classification tasks. Hence the model can be used for identifying Telugu English code-mixed hate speech comments used on social media platforms. In future, we aim to provide more accurate results for different code-mixed datasets and different algorithms may also give more accurate solutions to the solution.

## REFERENCES


[1] H. Kumar Sharma, K. Kshitiz and Shailendra, "NLP and Machine Learning Techniques for Detecting Insulting Comments on Social Networking Platforms," 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, France, 2018, pp. 265-272, doi: 10.1109/ICACCE.2018.8441728.

[2] Khan, A., Yousaf, J., Muhammad, T., & Ismail, M., "Hate Speech Detection using Machine Learning and N-Gram Techniques," 2023 JATISI (Jurnal Teknik Informatika dan Sistem Informasi), Vol. 10(1).

[3] Patil, P., Raul, S., Raut, D., & Nagarhalli, T., "Hate Speech Detection using Deep Learning and Text Analysis." In 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 322-330). IEEE.

[4] Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das. "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network." In *IEEE Access*, vol. 8, pp. 204951-204962, 2020, doi: 10.1109/ACCESS.2020.3037073.

[5] S. Khan et al., "HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional Gated Recurrent Unit with Capsule Network," in IEEE Access, vol. 10, pp. 7881-7894, 2022, doi: 10.1109/ACCESS.2022.3143799.

[6] Dr. D.Sameera, "Detection of abnormality in CCTV Footage: Computer Vision" at European journal of molecular & clinical medicine (EJMCM),2020, Volume 7, Issue 4, Pages 1148-1154.

[7] Seble, H., "Hate Speech Detection Using Machine Learning: A Survey", *A J S E*, vol. 17, no. 1, pp. 88–109, Sep. 2023.

[8] K. Nugroho et al., "Improving Random Forest Method to Detect Hatespeech and Offensive Word," 2019 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2019, pp. 514-518, doi: 10.1109/ICOIACT46704.2019.8938451.

[9] Sanz, Hector & Valim, Clarissa & Vegas, Esteban & Oller, Josep & Reverter, Ferran. (2018). SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. BMC Bioinformatics. 19. 10.1186/s12859-018-2451-4.

[10] Endrit Fetahi, Mentor Hamiti, Arsim Susuri, Visar Shehu, and Adrian Besimi, "Automatic Hate Speech Detection using Natural Language Processing: A state-of-the-art literature review.",2023,doi: 10.1109/meco58584.2023.10155070

[11] M. H. Abdurrahman, B. Irawan and C. Setianingsih, "A Review of Light Gradient Boosting Machine Method for Hate Speech Classification on Twitter," 2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), Kuala Lumpur, Malaysia, 2020, pp. 1-6, doi: 10.1109/ICECIE50279.2020.9309565.

[12] Dr. Divanu Sameera, Dr. Durga Prasad Kavadi, "A Novel Approach for Text Classification using Recurrent Neural Networks," 2020 International Journal of Advanced Science and Technology, Vol. 29, No. 5, (2020). 9371-9386.