# HATE SPEECH DETECTION

**ABSTRACT**

Hate speech represents a pervasive and detrimental form of online discourse, often manifested through an array of slurs, from hateful tweets to defamatory posts. As such speech proliferates, it connects people globally and poses significant social, psychological, and occasionally physical threats to targeted individuals and communities. Current computational linguistic approaches for tackling this phenomenon rely on labelled social media datasets for training. For unifying efforts, our study advances in the critical need for a comprehensive meta-collection, advocating for an extensive dataset to help counteract this problem effectively (1). We scrutinized over 60 datasets, selectively integrating those pertinent into Meta Hate. This paper offers a detailed examination of existing collections, highlighting their strengths and limitations. Our findings contribute to a deeper understanding of the existing datasets, paving the way for training more robust and adaptable models. (2) These enhanced models are essential for effectively combating the dynamic and complex nature of hate speech in the digital realm.

**INTRODUCTION**

 Hate speech refers to any speech or expression that attacks an individual or group on the basis of their identity, such as race, religion, gender, sexual

orientation, and ethnicity. It can have serious negative effects on individuals and society as a whole, including increased discrimination, violence, and

social exclusion. (3) With the rise of social media and online communication, hate speech has become more prevalent and widespread, making it a problem.

One approach to addressing hate speech is through the use of automated detection systems. These systems can automatically identify hate speech in text,

allowing for quick and efficient moderation of online content. There have been numerous studies and research on hate speech detection using ML.

In this paper, we propose a hate speech detection system that utilizes a decision tree algorithm. Decision trees are a simple and effective machine learning

algorithm that can handle large datasets and have been used successfully in various classification tasks. (4) Our proposed system uses a dataset of labelled

hate speech and non-hate speech text to train the decision tree model. We pre-process the input text by removing stop words and stemming the words,

extract relevant features using the TF-IDF method, and then use the decision tree algorithm to classify the input text as hate speech or non-hate speech.

The experimental results show that our proposed system achieves high accuracy and outperforms other existing hate speech detection systems..

**HATE SPEECH DETECTION DATASET**

[**https://drive.google.com/drive/folders/12GRDPmJ8KKZYb6rpVp88AlvIeF\_eEC9K**](https://drive.google.com/drive/folders/12GRDPmJ8KKZYb6rpVp88AlvIeF_eEC9K)

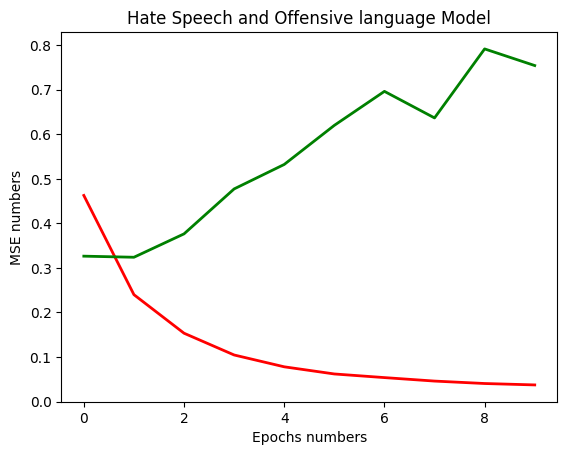
Twitter, Facebook, Reddit, Stormfront, Gab, Whisper, Wikipedia, Civil Comments, YouTube, and Bit Chute. The strategies employed for the dataset creation include: utilizing lexicons, keywords, hashtags, and phrase structures, and (5) randomly sampling from sites with a likelihood of containing hate content. In terms of conceptualization, the majority of works adopt a binary strategy. However, a vast of them take a multiclass approach, distinguishing between abusive, hate, offensive, or normal speech, among other terms.

**DATA COLLECTION AND PRE-PROCESSING**

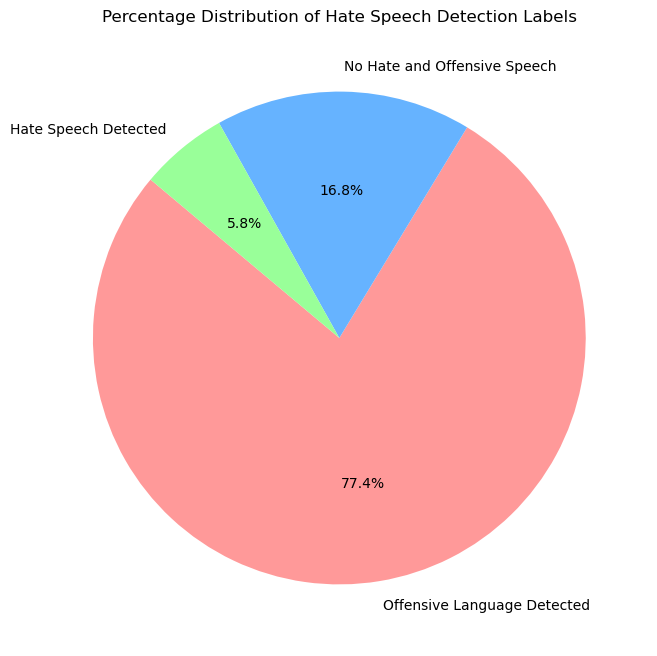
**About this file**

The file train.csv in the dataset is used for training machine learning models to detect hate speech and offensive language on Twitter

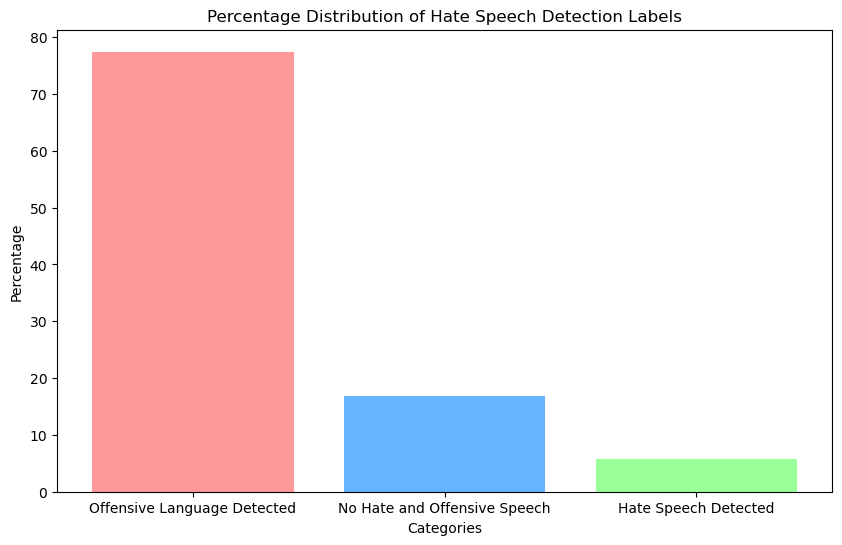
* count: The total number of annotations for each tweet. (Integer)
* count: The total number of annotations for each tweet. (Integer)
* hate speech count: The number of annotations classifying a tweet as hate speech. (Integer)
* hate speech count: The number of annotations classifying a tweet as hate speech. (Integer)
* offensive language count: The number of annotations classifying a tweet as offensive language. (Integer)
* neither count: The number of annotations classifying a tweet as neither hate speech nor offensive language. (Integer)



**PIE CHART**

****

**BAR CHART**



The Civil Comments dataset is a large-scale collection of online comments originally sourced from the Civil Comments platform, a community-moderated platform used for public discussions on news articles. (6) The dataset was created to support research in detecting toxic or harmful language, and it contains over 2 million comments annotated for various attributes, including levels of toxicity, personal attacks, insults, and identity-based discrimination. These comments were collected with additional metadata, such as demographic attributes (e.g., references to race, gender, and religion) to enable nuanced analyses of biases in language models and toxicity detection systems. The dataset is widely used in machine learning research for training and evaluating models that aim to promote respectful online discourse and improve content moderation.

**CIVIL COMMENT DATASET**

[**https://drive.google.com/drive/folders/12GRDPmJ8KKZYb6rpVp88AlvIeF\_eEC9K**](https://drive.google.com/drive/folders/12GRDPmJ8KKZYb6rpVp88AlvIeF_eEC9K)

**About this file**

### The original competition data uses a toxicity score ranging from 0 to 1. I've simplified this score to either 0 or 1 by thresholding the (7) value: scores > 0.7 are assigned "1", scores < 0.3 are assigned "0", and comments with scores between 0.3 and 0.7 are dropped from the dataset. Additionally, to reduce runtime, I have reduced the size of the dataset with

### **Text Pre-processing**

* **Lowercasing**: Convert all text to lowercase to maintain uniformity, as "Hate" and "hate" should be treated as the same word.
* **Tokenization**: Split the text into individual tokens (words or sub-words).
* **Stemming and Lemmatization**: Reduce words to their root form to handle morphological variations.

### 

### **Feature Extraction**

* **Count Vectorization**: This method converts text into numerical data by representing the frequency of each word in the tweets.
* **TF-IDF**: TF-IDF (Term Frequency-Inverse Document Frequency) weights the importance of words based on their frequency in the document and across the entire dataset.
* **Evaluation Metrics**: Accuracy: Measures overall performance. Precision: Assesses performance for each category. Confusion Matrix: Visualizes true positives, false positives, true negatives, and false negatives.

### **Model Selection and Training**

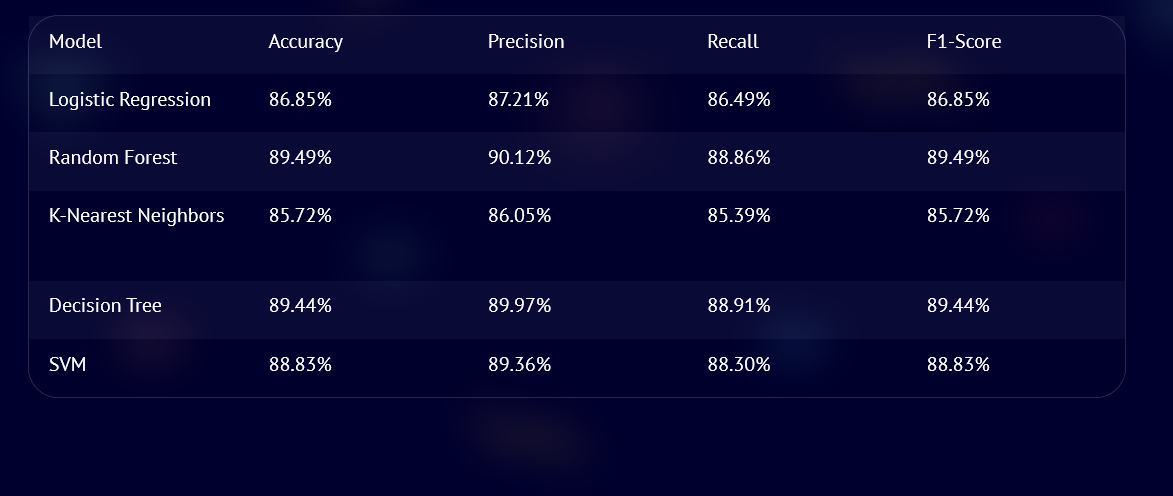
* **Logistic Regression**: This model predicts the probability of a tweet being hate speech.
* **Random Forest**: An ensemble method that combines multiple decision trees to improve accuracy and robustness.
* **K-Nearest Neighbors**: Classifies tweets based on their similarity to labeled examples.
* **Decision Tree:** Provides an interpretable model for understanding feature importance

**Support Vector Machine (SVM):** Handles complex, high-dimensional data effectively for classification.

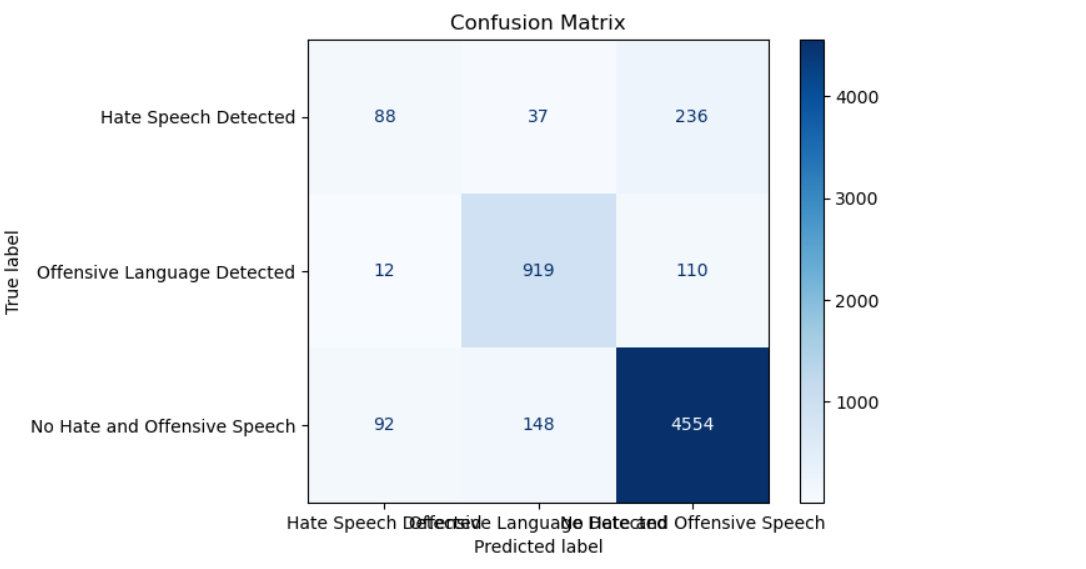
**Model Performance Comparison**

****

**Evaluation Metrics and Results**

****

**Confusion Matrix:** The **confusion matrix** plays a critical role in evaluating the performance of a machine learning model, especially in classification tasks like hate speech detection

****

**Conclusion**

As Hate speech continues to be a societal problem, the need for automatic hate speech detection systems becomes more apparent. We presented the current approaches for this task as well as a new system that achieves reasonable accuracy. We also proposed a new approach that can outperform existing systems at this task, with the added benefit of improved interpretability. Given all the challenges that remain, there is a need for more research on this problem, including both technical and practical matters. (8) However, it is important to note that hate speech detection is a complex and ongoing challenge, as hate speech can be expressed in many different forms and can evolve over time. As such, it is crucial that we continue to develop and refine machine learning models to improve their accuracy and effectiveness in detecting hate speech. Moreover, while machine learning can be a powerful tool for hate speech detection, it should not be relied on as the sole solution to address the issue. Efforts to combat hate speech must also include education, (9) dialogue, and community engagement to promote understanding and respect among individuals from diverse backgrounds. Ultimately, by working together and leveraging the power of technology and human compassion, we can create a more inclusive and welcoming online community for everyone.

**Project Mentor:** Dr. N Jagan Mohan

**Created By:** Soumyadeep Banerjee

**REFERENCES**

1. [**arxiv.org/html/2401.06526v1/#Sx1/**](https://arxiv.org/html/2401.06526v1/#Sx1/)
2. **arxiv.org**
3. **International Journal of Research Publication and Reviews (IJRPR)**
4. **International Journal of Research Publication and Reviews (IJRPR)**
5. **arxiv.org**
6. **kaggle.com**
7. **civil comment dataset Kaggle.com**
8. **International Journal of Research Publication and Reviews**
9. **International Journal of Research Publication and Reviews**