# Report on Hate Speech Detection in Tweets

## ***[SENTIMENTAL ANALYSIS]***

## Detection of Hate Speech in Tweets

## [S.Vanathi]

## September 15, 2024

### ABSTRACT

This report presents a study on the detection of hate speech in tweets, specifically focusing on identifying racist and sexist sentiments. Using a labelled dataset of 31,962 tweets, The aim is to classify tweets into two categories, those that contain hate speech and those that do not. The findings contribute to the understanding of hate speech dynamics on social media platforms.

### 1.INTRODUCTION

Hate speech has become a significant issue on social media, urging the need for automated detection systems. This project focuses on identifying tweets containing racist or sexist sentiments to reduce the spread of hate speech. By classifying tweets as either containing hate speech (label '1') or not (label '0'), the aim is to develop a model that can predict labels for unseen tweets.

### 2.DATA

The project utilizes two datasets:

* **train.csv**: A labelled dataset comprising 31,962 tweets. Each entry includes Tweet ID, Label, Tweet text.
* **test\_tweets.csv**: This file contains tweets for which predictions are to be made. It includes Tweet ID, Tweet text.

### 3.METHODALAGY

The project started with loading the training and test datasets. Dropped the tweet ID columns as it does not contribute for our classification process. Performed the following preprocessing steps to prepare the tweets for analysis,

* **Data Cleaning**: Cleaned the tweets by removing URLs, special characters, and stop words.
* **Rare Word Removal**: Words that appeared infrequently were removed, as they may not significantly contribute to the prediction.
* **Stemming and Lemmatization**: Both techniques were applied to reduce words to their root forms, to increase the model's ability to generalize.
* **Tokenization**: Tokenization of text was performed to convert it into a vector format.
* **Feature Generation**: Features were generated using the Term Frequency-Inverse Document Frequency (TF-IDF) technique.

### 4.TRAINING THE MODELS

After preprocessing, the data was split using a Train-Test split technique to prepare for model training. Classification models like Logistic Regression, Support Vector Machine and Random Forest were trained.

* Evaluated each model using metrics such as accuracy, precision, recall, and F1-score. The results of each model is given below,
* **Logistic Regressions:** Accuracy: 0.94, Precision: 0.95, Recall: 0.94 and F1-Score: 0.94
* **Support Vector Machine (SVM)**: Accuracy: 0.96, Precision: 0.96, Recall: 0.96and F1-Score: 0.95
* **Random Forest**: Accuracy: 0.96, Precision: 0.95, Recall: 0.96 and F1-Score: 0.96
* Cross-validation was performed to ensure robustness, the following are the cross-validated F1-scores of Logistic Regression with 0.93 ± 0.00, SVM with 0.95 ± 0.00 and Random Forest with 0.95 ± 0.00
* To optimize the performance of the models, Hyperparameter tuning techniques like Grid Search and Random Search were performed. The results after tuning are as follows,
* **Logistic Regressions:** Accuracy: 0.94, Precision: 0.95, Recall: 0.94 and F1-Score: 0.95
* **Support Vector Machine (SVM)**: Accuracy: 0.96, Precision: 0.96, Recall: 0.96and F1-Score: 0.95
* **Random Forest**: Accuracy: 0.96, Precision: 0.95, Recall: 0.96 and F1-Score: 0.96

Based on these results, the Random Forest model slightly performs well, so it was used to predict the labels on the test dataset.

### 5.RESULT

The analysis of the test dataset is given below,

* **Number of Hate Speech Tweets (Label '1')**: 348
* **Number of Non-Hate Speech Tweets (Label '0')**: 6045

This indicates that out of the total test dataset, 348 tweets were identified as containing hate speech, which includes those with racist or sexist sentiments.

### 6.ANALYSIS

The results enhances the effectiveness of machine learning techniques in detecting hate speech within social media content. The Random Forest model, in particular, demonstrated high performance across various metrics, showing its suitability for this classification task. However, the relatively low number of hate speech instances in the dataset (348 out of 6393) suggests that the occurance of such tweets might be low, yet their impact can be significant.

### 7.CONCLUSION

In conclusion, this project successfully developed a model to detect hate speech in tweets, specifically focusing on racist and sexist sentiments. Through cautious preprocessing and the application of various machine learning algorithms, we identified the Random Forest model as the most effective in accurately classifying tweets. The analysis showed that 348 tweets contained hate speech, highlighting the necessity for ongoing monitoring and intervention strategies on social media platforms. Future work could explore incorporating additional features, or employing deep learning techniques to further enhance detection capabilities.