# Temporal Analysis and Sentiment-Aware Detection of Hate Speech and Cyberbullying


**Group Members**

Het Shah (21BIT183)

Mihir Kosrekar (21BIT198)

Yashrajsinh Dodiya (21BIT199)


**Problem Statement**

The increasing prevalence of online hate speech and cyberbullying demands improved detection systems. Current models struggle with the nuances of language, context, and evolving expressions, leading to inaccurate moderation. This project aims to develop a robust, adaptable system that accurately identifies such harmful content, leveraging sentiment analysis to enhance detection while preserving freedom of expression. Additionally, the project will incorporate temporal analysis to identify targeted hate speech and potential attacks.

**Introduction**

The rapid growth of online platforms has led to a rise in hate speech and cyberbullying, necessitating the development of more sophisticated detection systems. Current models struggle with accurately identifying harmful content due to the evolving nature of language, context, and subtle expressions in online communication. This project aims to build a robust and adaptable detection system that can address these challenges. By leveraging advanced techniques like deep learning, BERT-based models, sentiment analysis, and data augmentation, the system will enhance the detection of hate speech and cyberbullying while preserving freedom of expression. The models achieving high accuracy but facing limitations related to data variety, contextual nuances, and computational costs. This project seeks to overcome these challenges by integrating cutting-edge methods to create a more flexible, context-aware detection model. To achieve this, the project will follow a comprehensive plan of action:

- **Data Collection:** Find Dataset or Scrape Dataset from Reddit, Twitter etc.
- **Pre-processing:** Tokenize and Clean the data and Label the Data
- **EDA :** Visualizing histogram, word clouds and class distribution plots to identify patterns in dataset also analyze word frequency and distribution of hate speech across a dataset.
- **Model Development:** Use basic machine learning models to understand the Dataset like Logistic Regression, Support Vector Machine etc. Then we can use Deep learning models like SVC,Linear SVC,Random Forest, Bagging, Decision Tree etc followed by Testing the Dataset on BERT Natural Language Processing Models. We will also try integrate Hybrid Models like pairing Random Forest to gain higher accuracy.
- **Model Evaluation:** We would use standard evaluation metrics like precision, recall, F1 score, and AUC-ROC.
- **Temporal Analysis:** Do a Temporal Analysis of Hate Speech and Detect Targeted Attacks.
-

## Literature Review

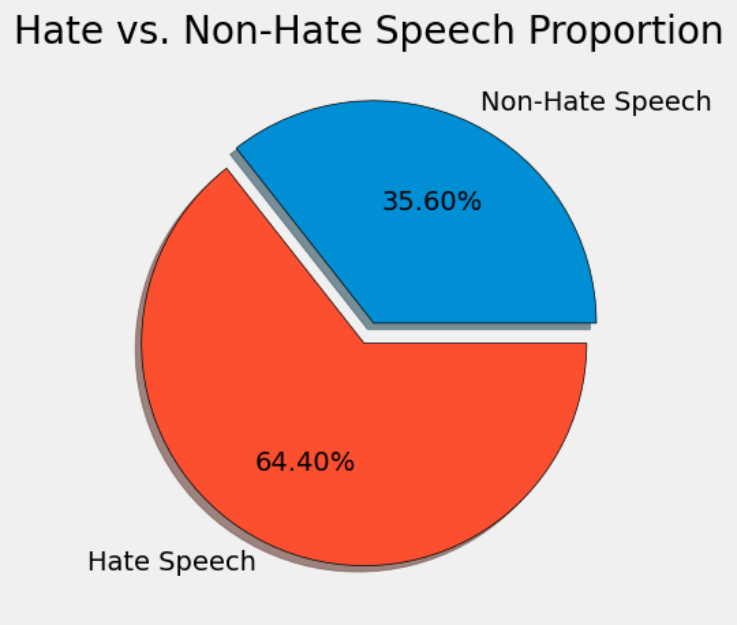
| Sr No | Title | Author | Year | Techniques Used | Findings | Limitations |
|---|---|---|---|---|---|---|
| 1. | Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study | Maral Dadvar and Kai Ecker | 2018 | Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) | Achieved 85% accuracy in reproducing results using deep learning-based models. | Limited data variety, results may not generalize to other social networks |
| 2. | Transfer Learning for Hate Speech Detection in Social Media | Lanqin Yuan, Tianyu Wang | 2023 | BERT, Transfer Learning | Transfer learning improved hate speech detection accuracy to 92% on target datasets. | Dependency on pre-trained models, domain adaptation challenges |
| 3. | Detecting Online Hate Speech: Approaches Using Weak Supervision and Network Embedding Models | Michael Ridenhour, | 2020 | Weak Supervision, Node2Vec (Network Embedding) | Achieved a 76% detection accuracy using weak supervision and network embedding. | Lack of large annotated datasets, scalability concerns |
| 4. | Cyberbullying Detection in Twitter Using Sentiment Analysis | Chong Poh Theng, Nur Fadzilah Othman | 2021 | Sentiment Analysis, Naive Bayes | Sentiment analysis achieved an F1 score of 0.74 for detecting cyberbullying in tweets. | Limited to Twitter, may not work for more nuanced cases |
| 5. | Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model | Hind Saleh, Areej Alhothali | 2023 | BERT, Word2Vec Embedding, LSTM | BERT model reached 89% accuracy in detecting hate speech compared to traditional models. | Requires high computational power, dataset limitations |
| 6. | A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter | Amgad Muneer, Suliman Mohamed Fati | 2020 | Support Vector Machines (SVM), Random Forest, Logistic Regression | SVM performed best with an accuracy of 83%, compared to 78% for Random Forest. | Dataset imbalance, inability to handle contextual nuances |
| 7. | Exploring Data Augmentation for Gender-Based Hate Speech Detection | Muhammad Amien Ibrahim, Samsul Arifin, Eko Setyo Purwanto | 2023 | Data Augmentation, Random Forest | Data augmentation increased detection accuracy from 68% to 75% | May introduce noise into the dataset, generalizability concerns |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | for gender-based hate speech. | |
| 8. | Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods | Aigerim Toktarova, Dariga Syrlybay | 2023 | Random Forest, CNN, LSTM | Hybrid method combining Random Forest and CNN achieved 90% accuracy in detecting hate speech. | High computational costs, dataset limitations |
| 9. | Artificial Intelligence-Enabled Cyberbullying-Free Online Social Networks in Smart Cities | Abdulsamad Al-Marghilani | 2022 | Reinforcement Learning, Sentiment Analysis | AI-based systems can create 90% cyberbullying-free environments in smart cities through proactive detection mechanisms. | Lack of real-world application, implementation challenges in diverse environments |
| 10. | BullyNet: Unmasking Cyberbullies on Social Networks | Aparna Sankaran Srinath, Hannah Johnson | 2021 | Graph Convolutional Networks (GCN), Logistic Regression | GCN model unmasked cyberbullies with an 88% accuracy in social networks. | May not handle more subtle forms of cyberbullying, limited scalability |
| 11. | Towards a Cyberbullying Detection Approach: Fine-tuned Contrastive Self-supervised Learning | Lulwah M. Al-Harigy, Hana A. Al-Nuaim | 2024 | Contrastive Self-supervised Learning, Data Augmentation | Fine-tuned contrastive learning model reached 84% accuracy, with improved generalization across domains. | Self-supervised models require more training, may not generalize well across platforms |
| 12. | Cyberbullying-related Hate Speech Detection Using Shallow-to-Deep Learning | Daniyar Sultan, Aigerim Toktarova | 2023 | Shallow Neural Networks, Deep Learning (CNN) | Shallow-to-deep learning approach increased accuracy to 87% for hate speech detection. | Computational costs and model complexity could limit real-time applications |
| 13. | Cyber-Bullying Detection in Hinglish Languages Using Machine Learning | Karan Shah | 2022 | Ada Boost Classifier,K-NN,SGD Clasifier,Random Forest Classifier. | Random Forest classifier provided the highest accuracy (97.1%) and F1-score (97.2%) with TF-IDF. | The dataset size for Hinglish language (around 3000 rows) |

**Data Analysis**

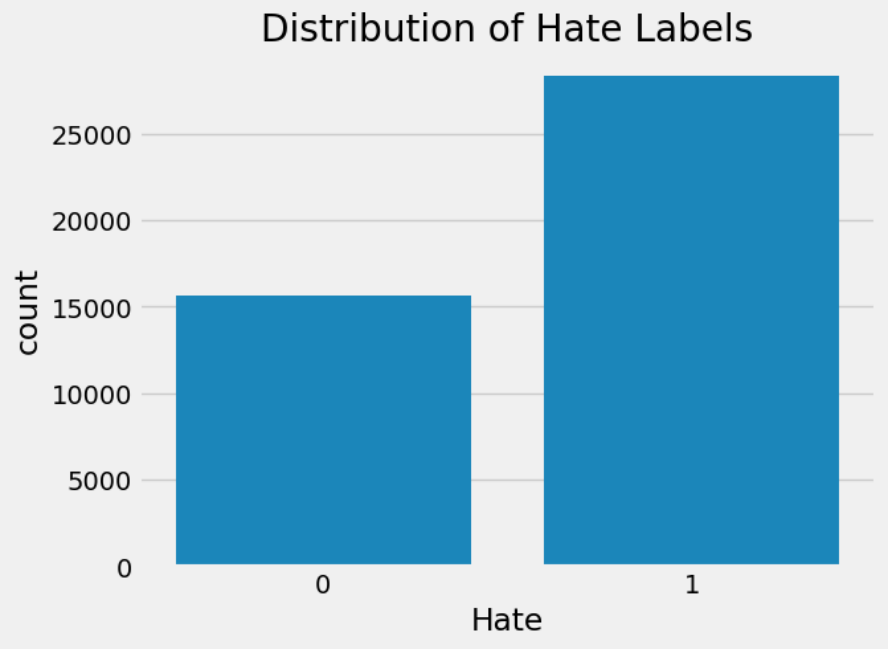
Our Dataset contains data from various datasets which contains data of Youtube comments, Tweets on twitter, Reddit Posts/comments etc.

**1.Hate Speech vs Non-Hate Speech Pie Chart**

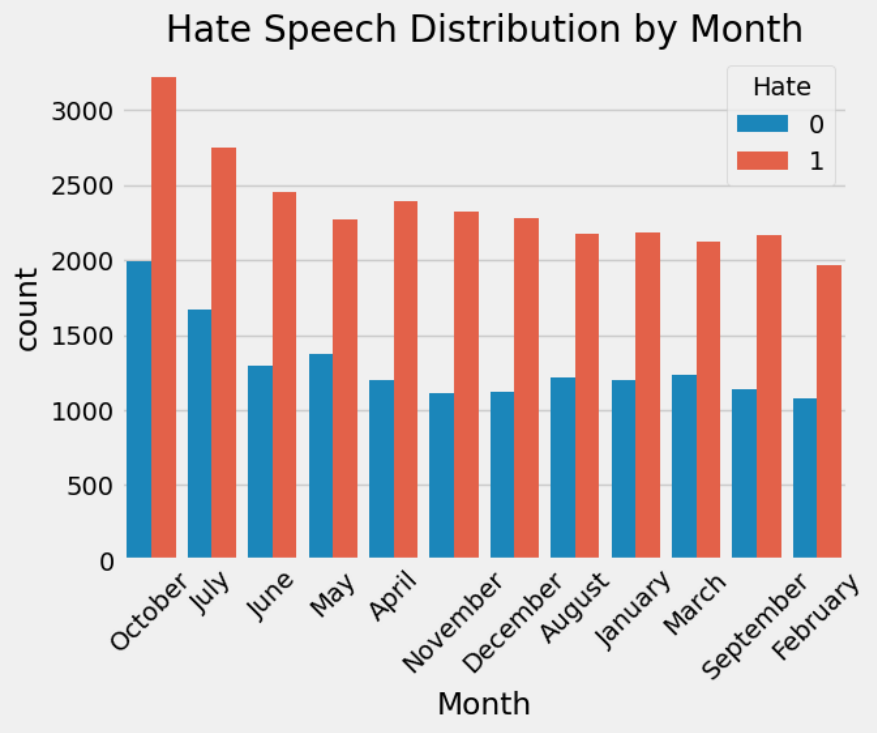



This Pie chart demonstrates the proportion of Hate Speech and Non-Hate Speech in the dataset. In our dataset, 65.08% of total speech is Hate Speech, which is enough to train the model well on Hate Speech data.

**2.Distribution of Hate Speech and Non-Hate Speech Labels**

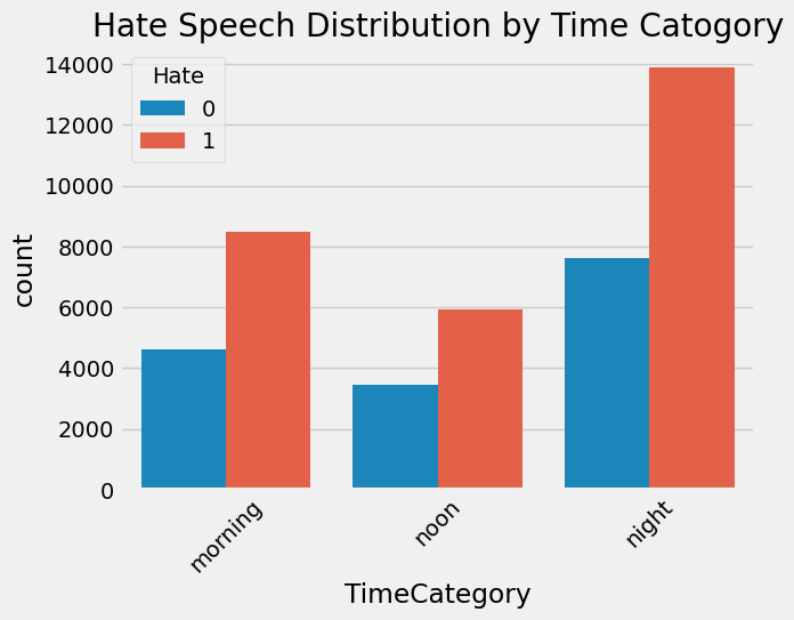


**3.Hate Speech Distributed by Month**



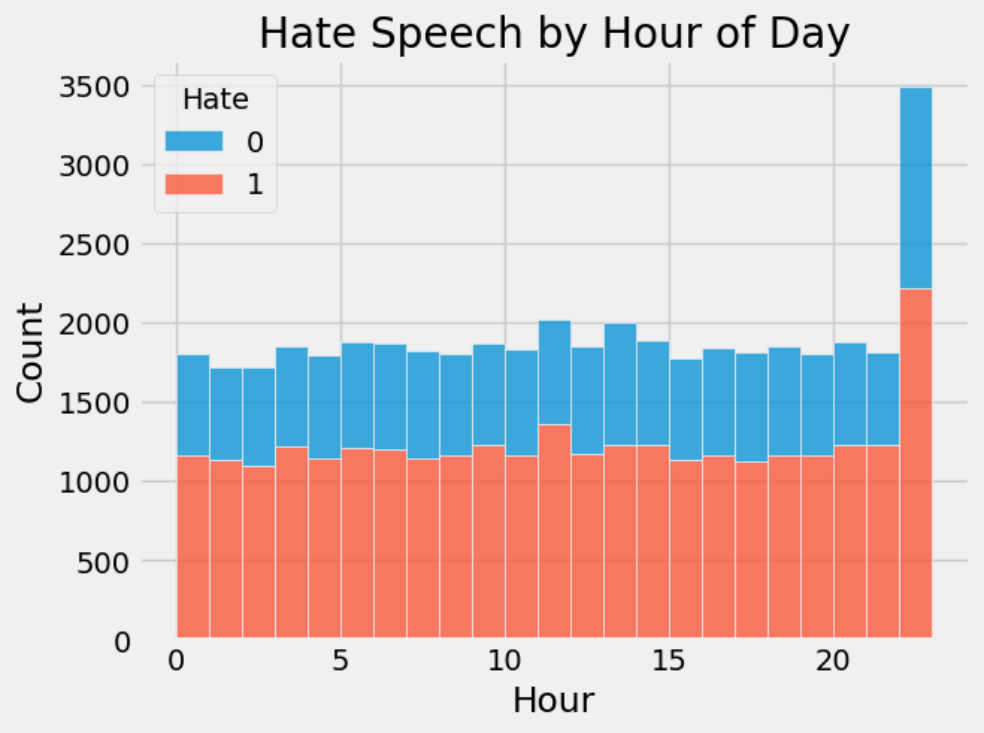

This histogram describes the Hate Speech generation in each month. Frequency of Hate Speech generation is highest in October and lowest in February also frequency of Non-Hate Speech generation is highest in Octpber and lowest in February**.**

**4.Hate Speech Distribution by the Time Category of the Day**



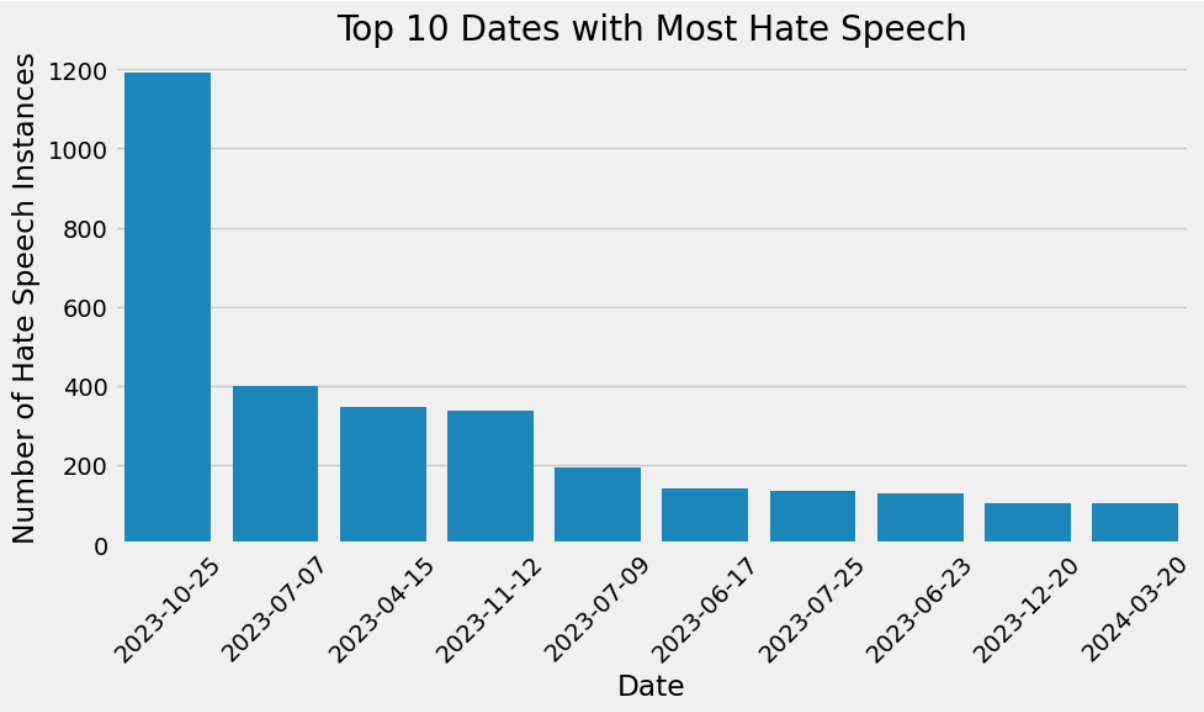

This histogram describes that in night time, Hate Speech generation is the highest and in noon time Hate Speech generation is the lowest

**5.Hate Speech by the Hour of the Day**



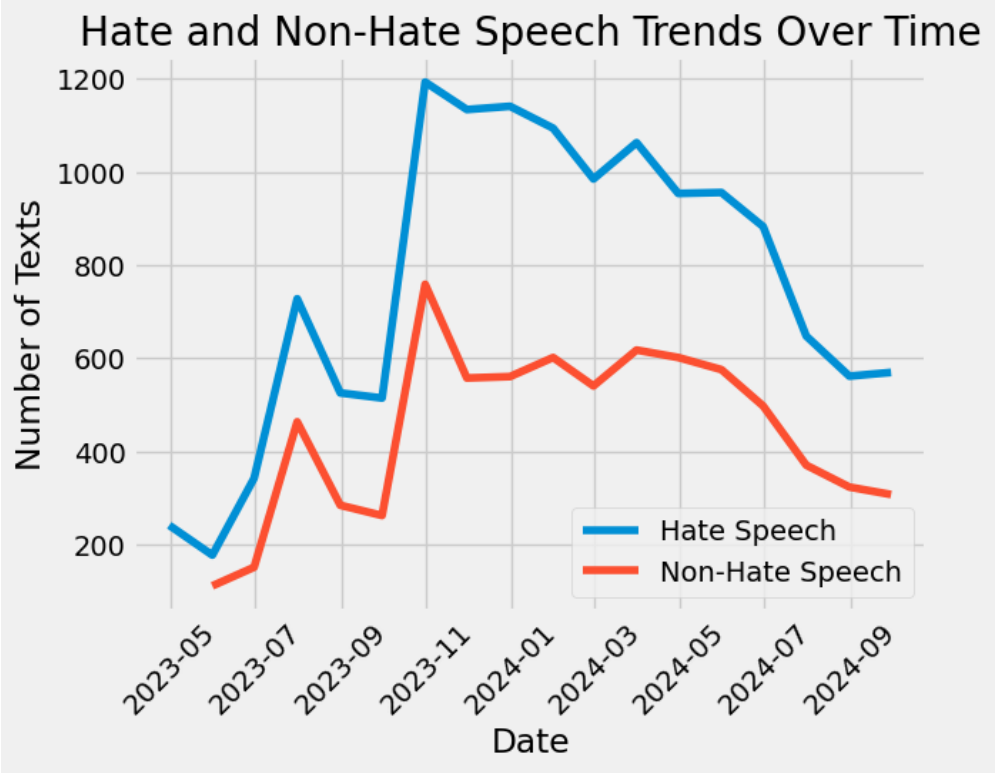

This histogram describes that near 12 am in night, Hate Speech generation takes sudden increase, also in mid-night, around 2-3 am, Hate Speech is generated the lowest.

**6. Dates with the Highest amount of Hate Speech**



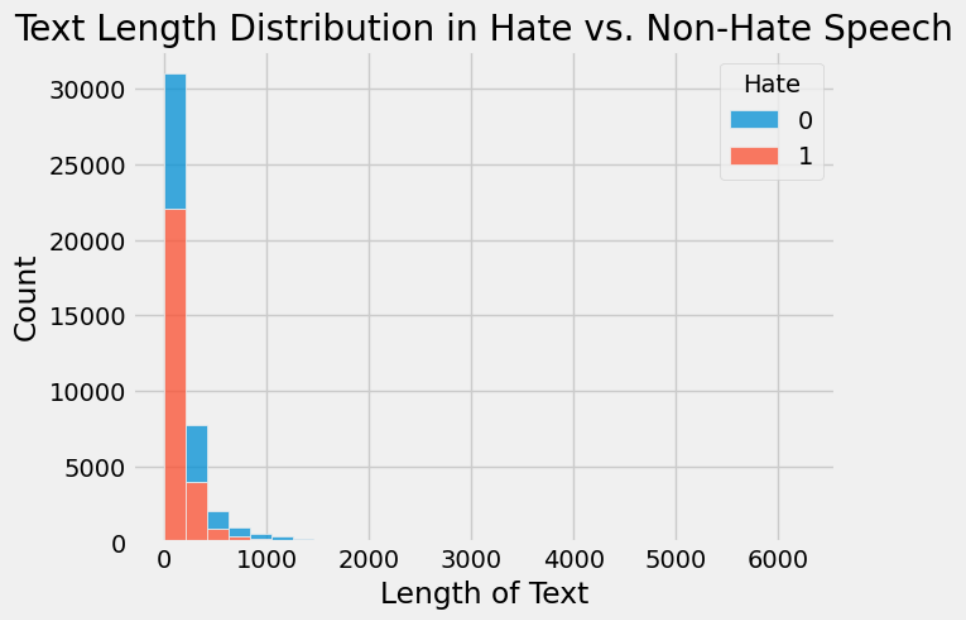

This graph demonstrates the dates on which highest number of Hate Speech is generated, which can help us to understand the event and its reaction on these days.

**7.Hate and Non-Hate Speech Trends Over Time**




This is the time-series analysis of Hate and Non-Hate Speech, which describes the frequency of the speech after specific date intervals. It can help us to understand the trend of speech.

**8.Text Length Distribution**

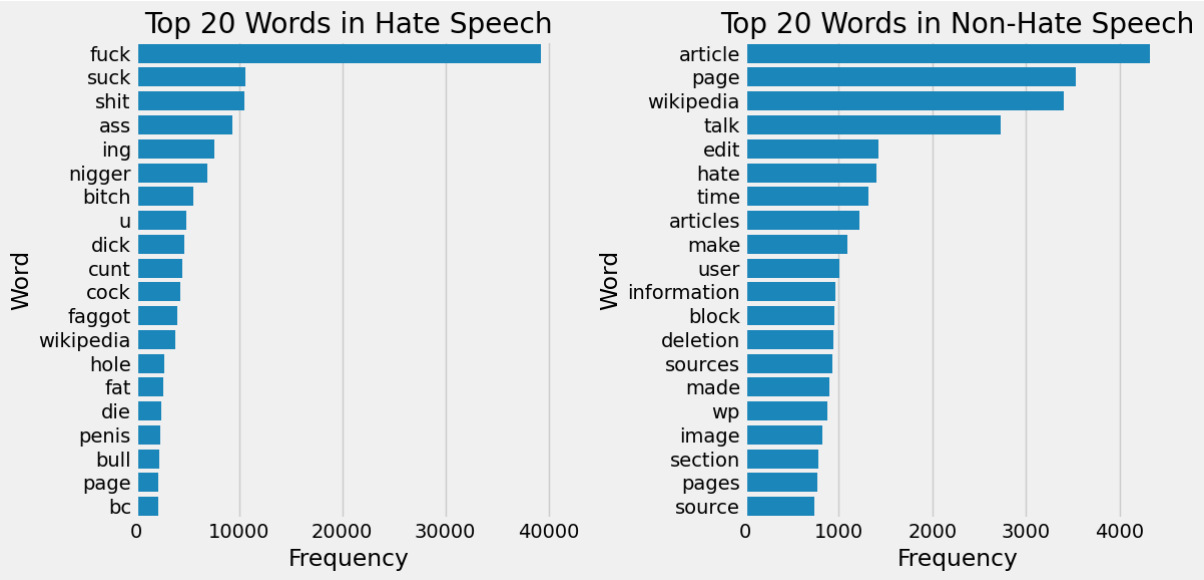



This histogram compares the frequency of Hate Speech and Non-Hate Speech based on the length of the speech. It describes that Hate Speech does not contain high length of speech, while Non-Hate Speech contains high length of speech.

## 9.Word Cloud

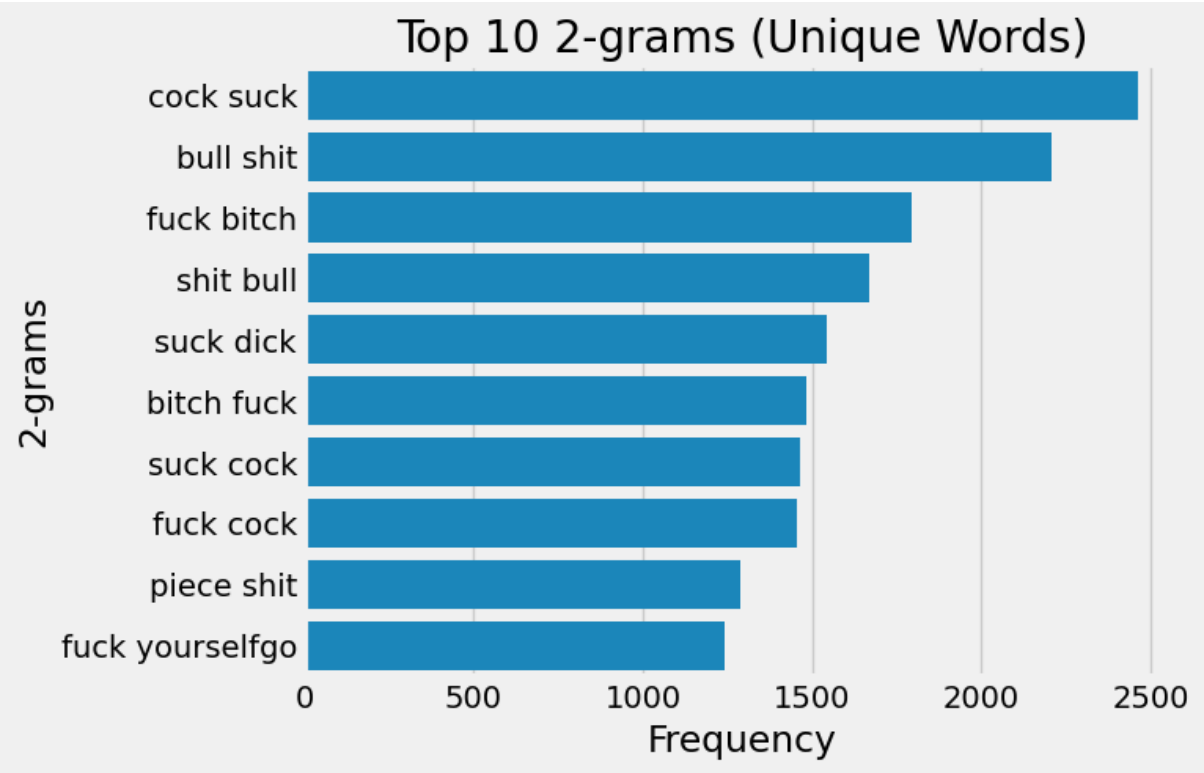

Here are some Hate Speech and Non-Hate Speech words, based on the labels and frequency of the words in our dataset. Here, all the words, in label Hate Speech= 1, are considered as Hate Speech words, and same for Non-Hate Speech word. Therefore, there may be common words in both.

## 10.Top 20 Words in Hate Speech and Non-Hate Speech Category



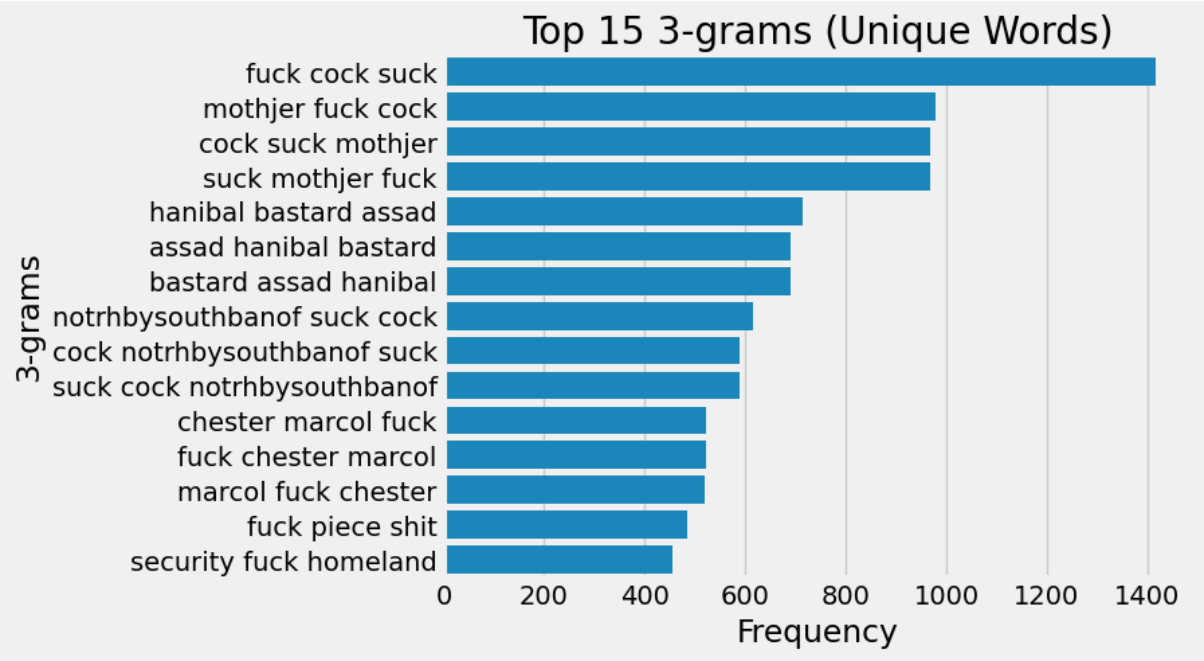

This histogram, describes the top 20 most frequent words in Hate Speech and Non-Hate Speech. This chart can help us understand which words are widely used in Hate Speech.

**11.Top 10 2-gram Words in Hate Speech**



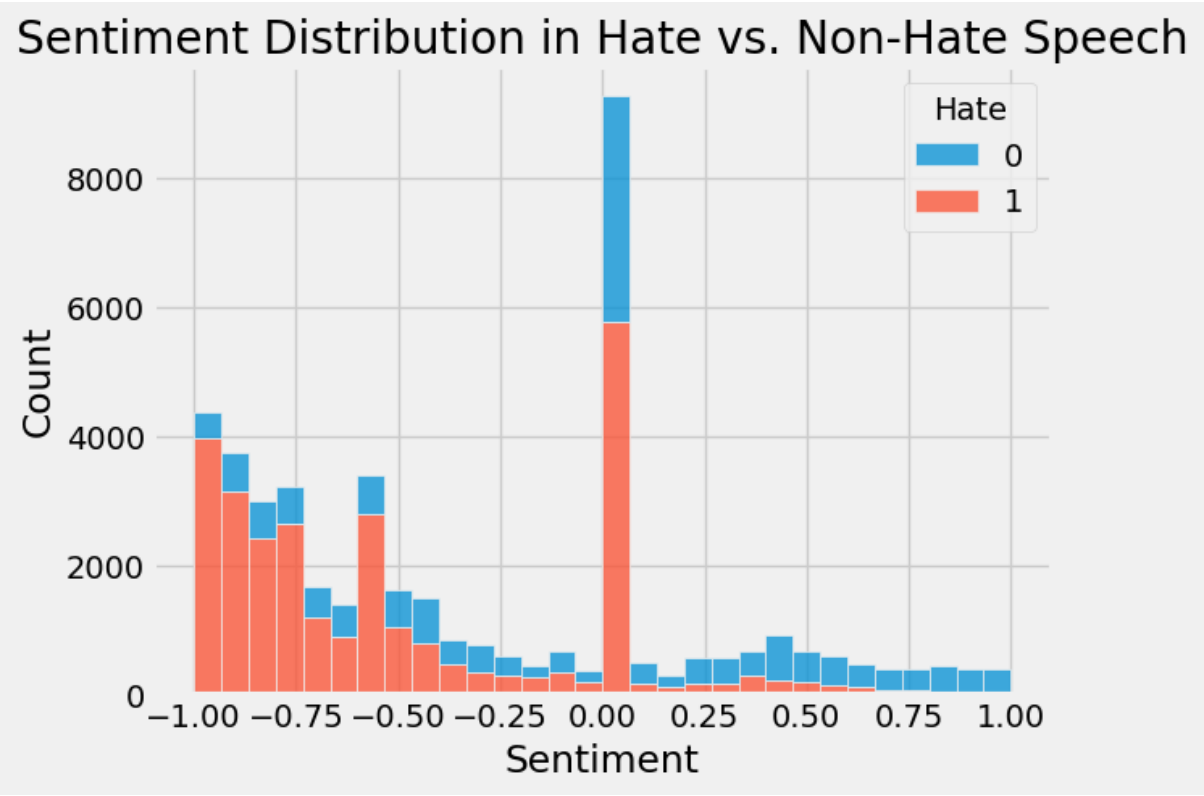

This histogram describes top 10 most frequent 2-grams (2 continuous words) used in the Hate Speech.

**12.Top 15 3-gram Words in Hate Speech**



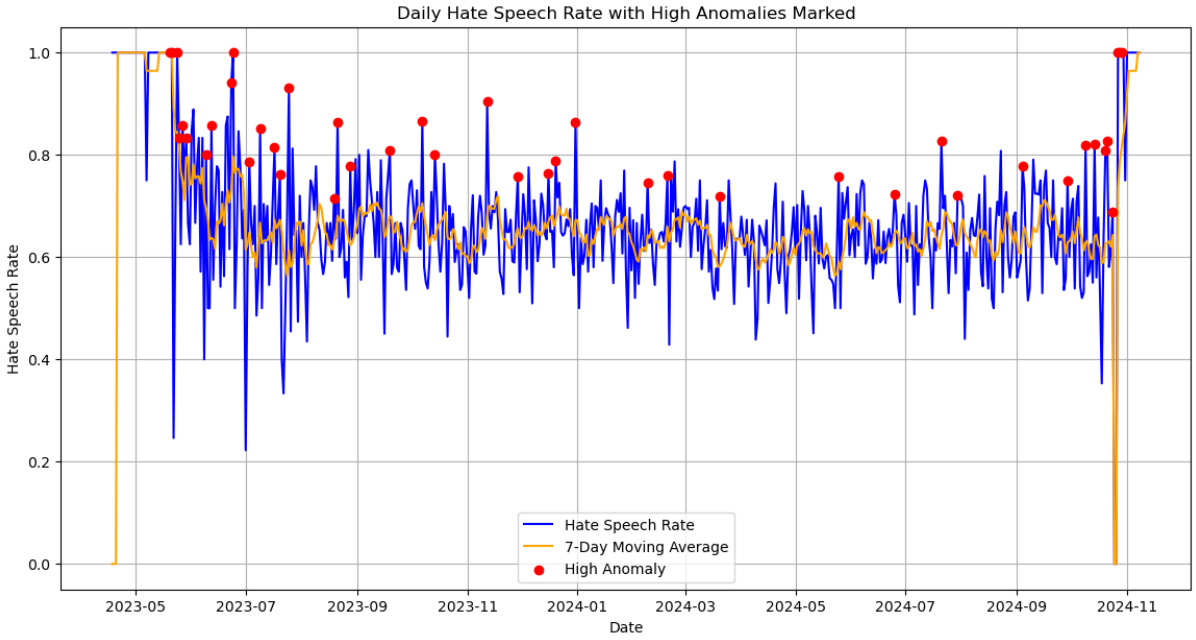

This histogram describes top 15 most frequent 3-grams (3 continuous words) used in the Hate Speech.

**13.Sentiment Distribution in Hate vs Non-Hate Speech**




This histogram describes the sentiment analysis, based on the sentiment polarity score. Here Hate Speech is mostly present in negative sentiment region.

**14. Targeted Hate Speech Detection**




.To detect anomalies in the hate speech rate, the code calculates a 7-day moving average and standard deviation, and flags days with z-scores above 1.5 as significant spikes.

## Machine Learning

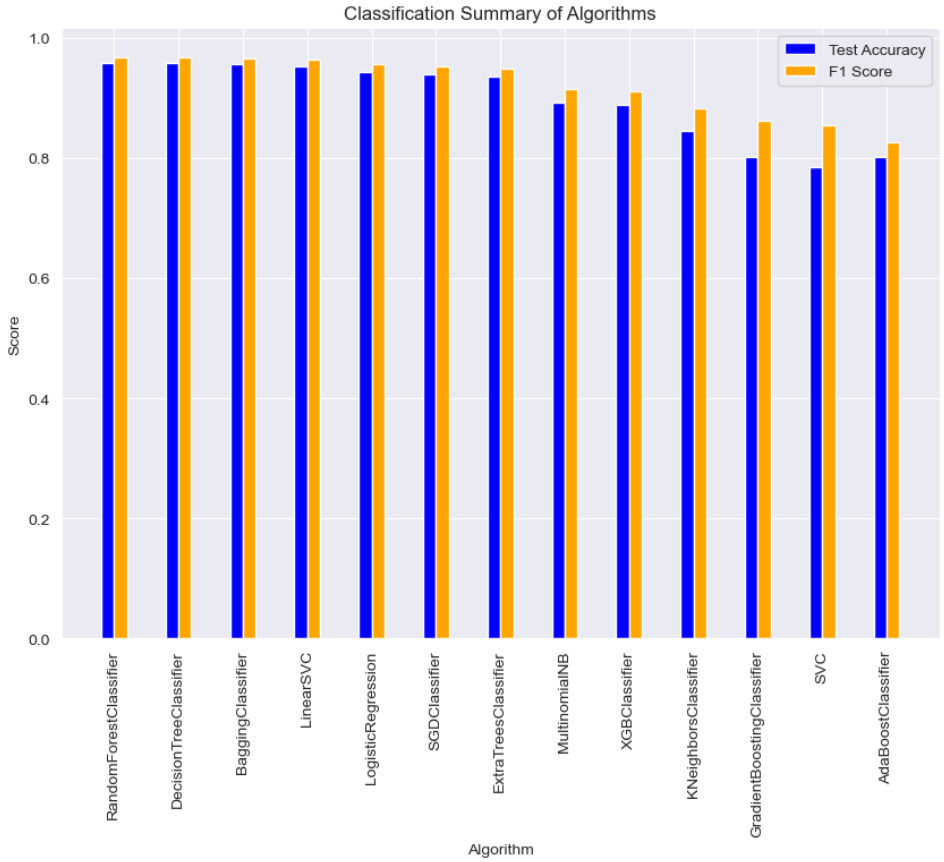
We, with this project implement a machine learning pipeline for text classification, aimed at identifying potentially offensive or bullying content. The process begins with data cleaning and vectorization, where raw text data is transformed into numerical format using a CountVectorizer. Following this, the data is split into training and testing sets, ensuring reliable evaluation. A variety of classifiers, including Random Forest and Decision Tree, are trained, each assessed on metrics like accuracy, F1 score, precision, and recall. The models are saved for future use, enabling efficient reuse for new predictions.

The project includes thorough evaluation and visualization to compare model performance. Performance metrics and confusion matrices provide detailed insight into each model's strengths and weaknesses, while ROC curves and bar plots offer visual clarity. Training and prediction times are also compared, helping to assess computational efficiency. Additionally, a feature for testing new text inputs in real-time allows for practical application and continuous model assessment. This approach provides a scalable solution for text classification, with potential applications in content moderation and online safety.

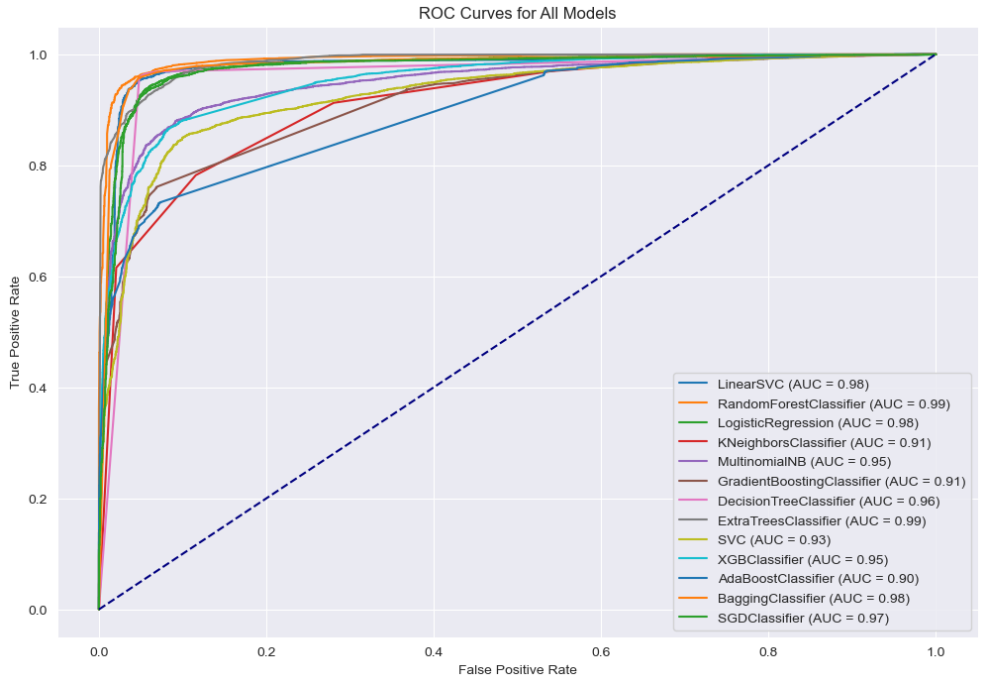
### 1.Table of Machine Learning Model’s Performance

| Unnamed: 0 | Algorithm | Accuracy: Test | Precision: Test | Recall: Test | F1 Score: Test | Prediction Time(in sec) | Training Time(in sec) |
|---|---|---|---|---|---|---|---|
| 1 | BERT | 0.9891 | 0.9891 | 0.9753 | 0.9891 | 54.96 | 7,200 |
| 2 | DecisionTreeClassifier | 0.9576 | 0.9713 | 0.9627 | 0.967 | 0.0199 | 8.5431 |
| 3 | RandomForestClassifier | 0.9565 | 0.9749 | 0.9574 | 0.966 | 2.303 | 113.7677 |
| 4 | BaggingClassifier | 0.9541 | 0.9677 | 0.961 | 0.9644 | 0.2501 | 155.1668 |
| 5 | LinearSVC | 0.952 | 0.9677 | 0.9577 | 0.9627 | 0.0023 | 2.5674 |
| 6 | LogisticRegression | 0.9432 | 0.9569 | 0.9551 | 0.956 | 0 | 1.3079 |
| 7 | SGDClassifier | 0.9409 | 0.9588 | 0.9494 | 0.9541 | 0 | 0.1887 |
| 8 | ExtraTreesClassifier | 0.9363 | 0.9597 | 0.9409 | 0.9502 | 4.7071 | 125.121 |
| 9 | MultinomialNB | 0.8911 | 0.9219 | 0.9082 | 0.915 | 0.0052 | 0.0101 |
| 10 | XGBClassifier | 0.886 | 0.9424 | 0.8772 | 0.9086 | 0.0938 | 5.9699 |
| 11 | KNeighborsClassifier | 0.8426 | 0.8564 | 0.9088 | 0.8818 | 12.913 | 0.0172 |
| 12 | GradientBoostingClassifie | 0.802 | 0.7824 | 0.9607 | 0.8624 | 0.0276 | 20.7935 |
| 13 | SVC | 0.7846 | 0.7602 | 0.9737 | 0.8538 | 70.9657 | 550.6772 |
| 14 | AdaBoostClassifier | 0.8014 | 0.9478 | 0.7328 | 0.8266 | 0.2504 | 6.8414 |

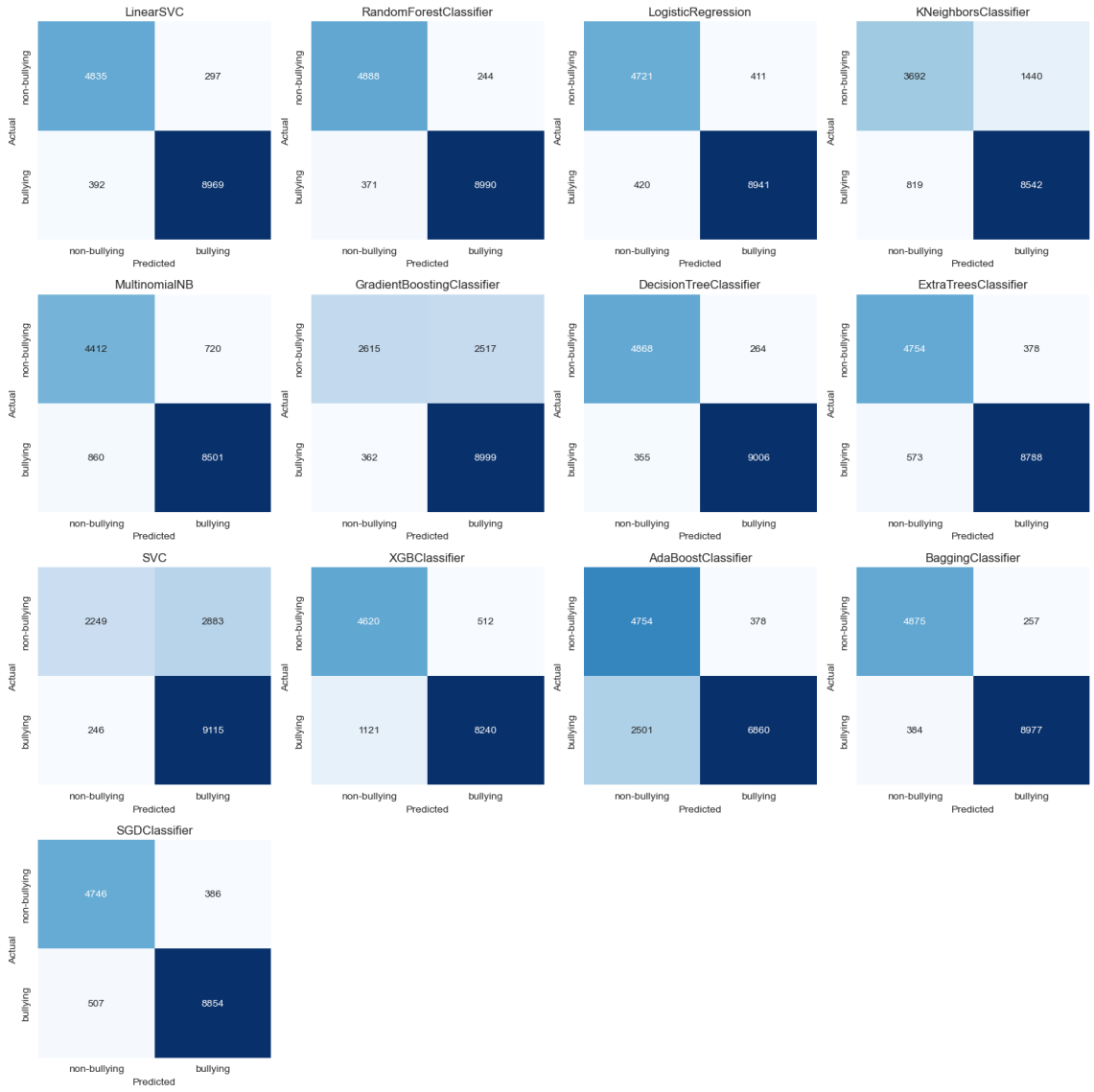
## 2.Classification Summary of Machine Learning Algorithms




## 3. ROC Curves for All Models

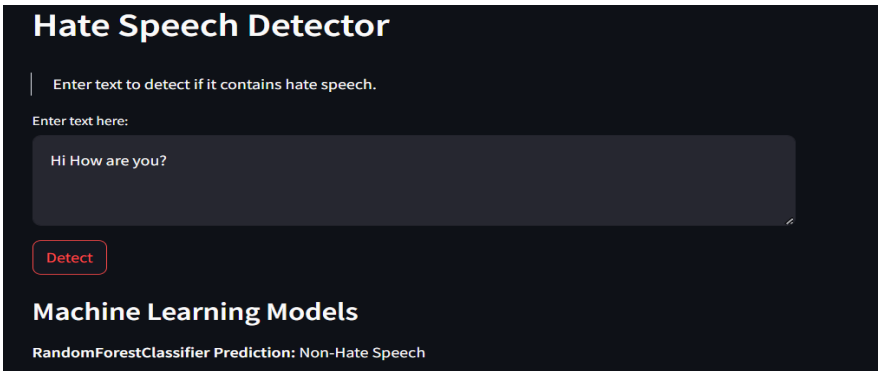


## 4.Confusion Matrix

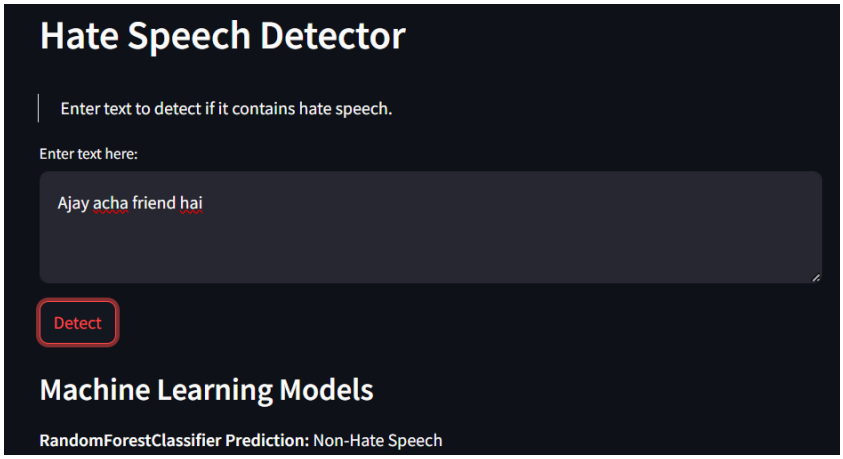



## 5)Predictions

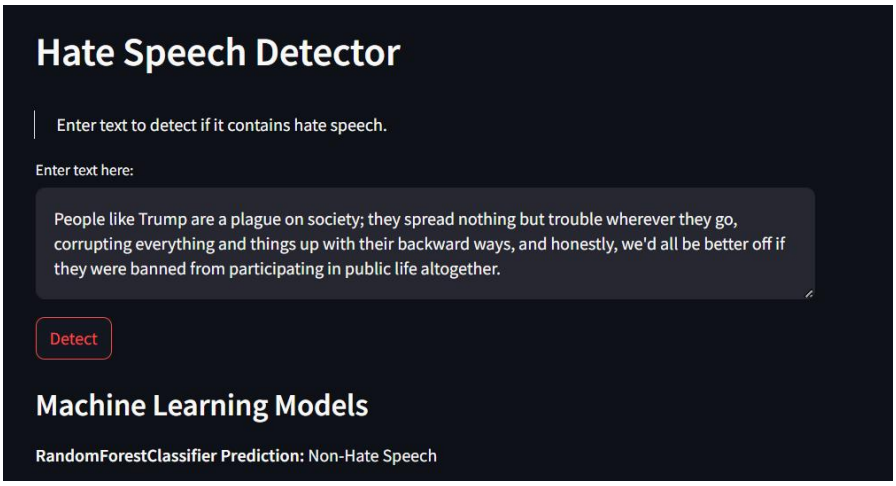
Text: Hi How are you?



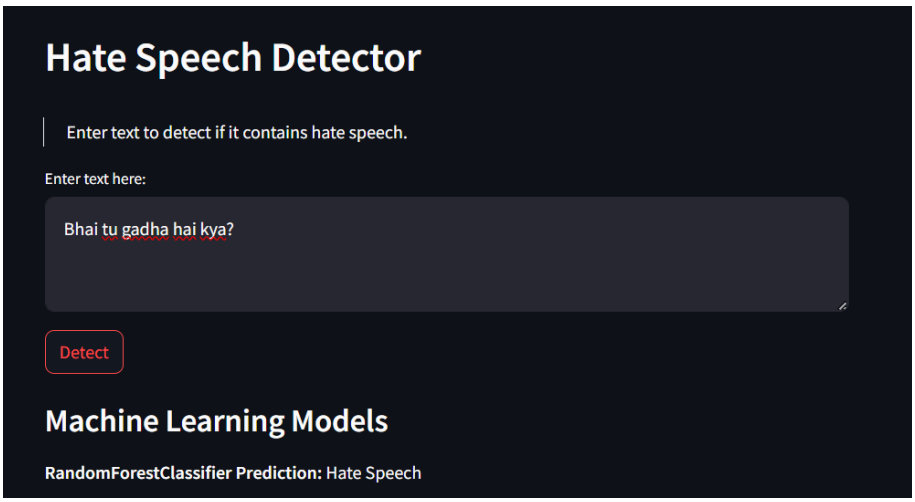
Text:Ajay acha friend hai




Text: People like Trump are a plague on society; they spread nothing but trouble wherever they go, corrupting everything  with their backward ways, and honestly, we'd all be better off if they were banned from participating in public life altogether.




Text: Bhai tu gadha hai kya?



**BERT Model**

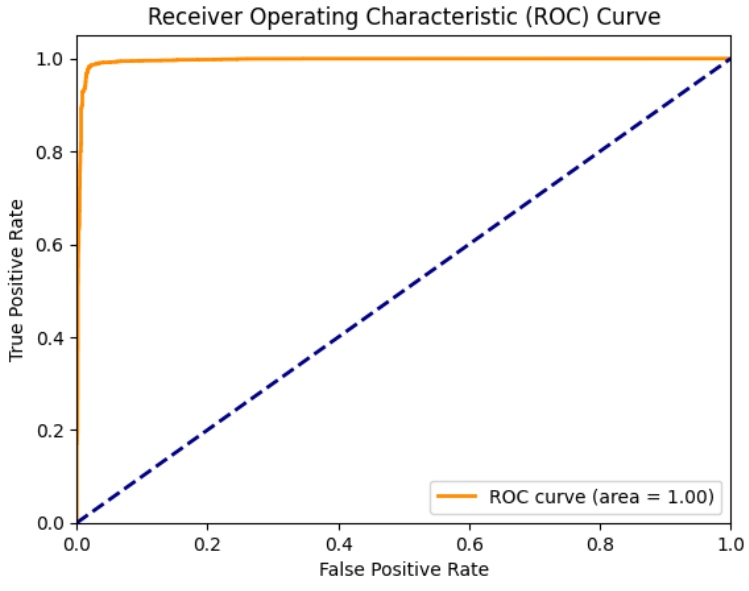
The BERT model (`bert-base-uncased`) was fine-tuned to classify text as either hate speech or non-hate speech. The process begins by loading and preprocessing the data, where text samples are tokenized and converted into numerical tensors using the `BertTokenizer`. These tensors are then split into training and validation sets, with attention masks added to retain focus on essential words in each sentence. The model is trained for five epochs, during which the optimizer minimizes the classification loss, allowing BERT to learn meaningful patterns in the text.

After training, the model's performance is evaluated using accuracy, precision, recall, and F1 score on the validation set. A confusion matrix provides insight into the model's correct and incorrect classifications, while an ROC curve visualizes its effectiveness in distinguishing between classes. Additionally, the model is saved and later used for real-time testing, allowing it to predict labels for new text inputs. The BERT model's high accuracy and robust evaluation metrics make it a powerful tool for detecting hate speech, offering reliable predictions suitable for moderation applications.

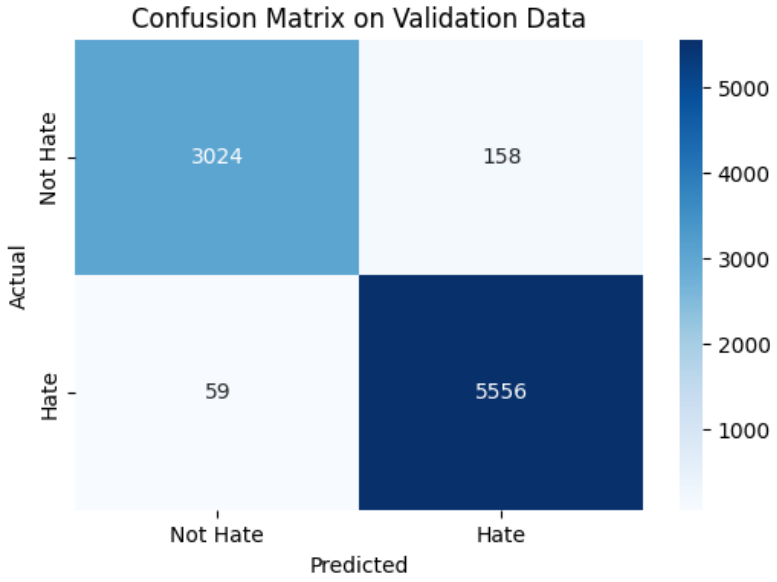
**1)The model was trained on 5 epochs:**

```
100%|██████████| 1100/1100 [10:49<00:00,  1.69it/s]
Epoch 1, Average training loss: 0.252968562299555
Epoch 1, Validation accuracy: 0.9232
100%|██████████| 1100/1100 [10:50<00:00,  1.69it/s]
Epoch 2, Average training loss: 0.15450984242033552
Epoch 2, Validation accuracy: 0.9504
100%|██████████| 1100/1100 [10:50<00:00,  1.69it/s]
Epoch 3, Average training loss: 0.08479525165251371
Epoch 3, Validation accuracy: 0.9666
100%|██████████| 1100/1100 [10:50<00:00,  1.69it/s]
Epoch 4, Average training loss: 0.04574592518778941
Epoch 4, Validation accuracy: 0.9756
100%|██████████| 1100/1100 [10:50<00:00,  1.69it/s]
Epoch 5, Average training loss: 0.02904804276582912
Epoch 5, Validation accuracy: 0.9753
```

**2)Model ROC Curve**

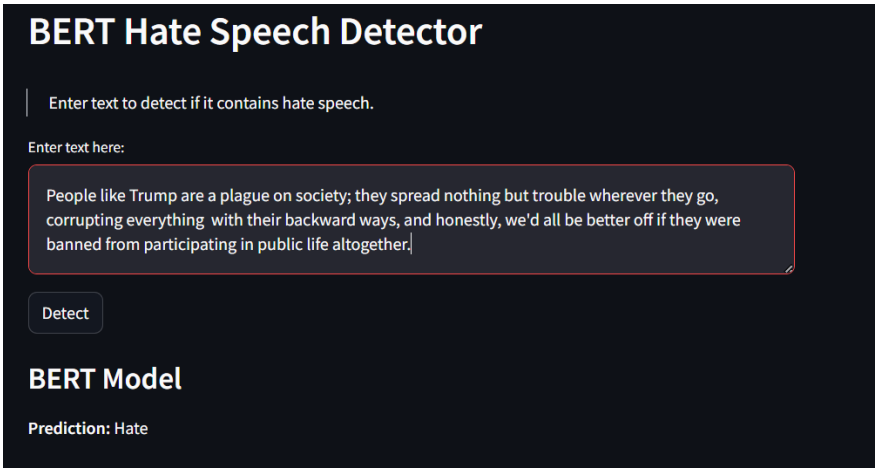


### 3)Model Confusion Matrix:

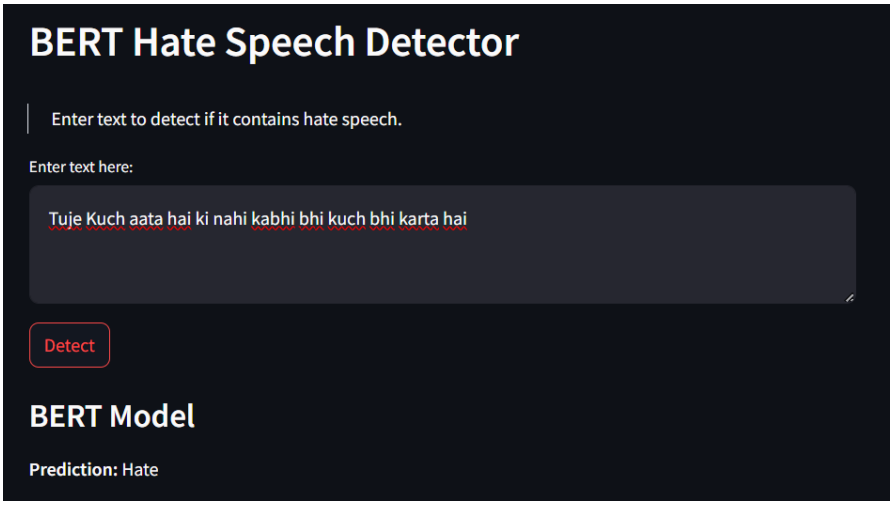



### 4)Predictions

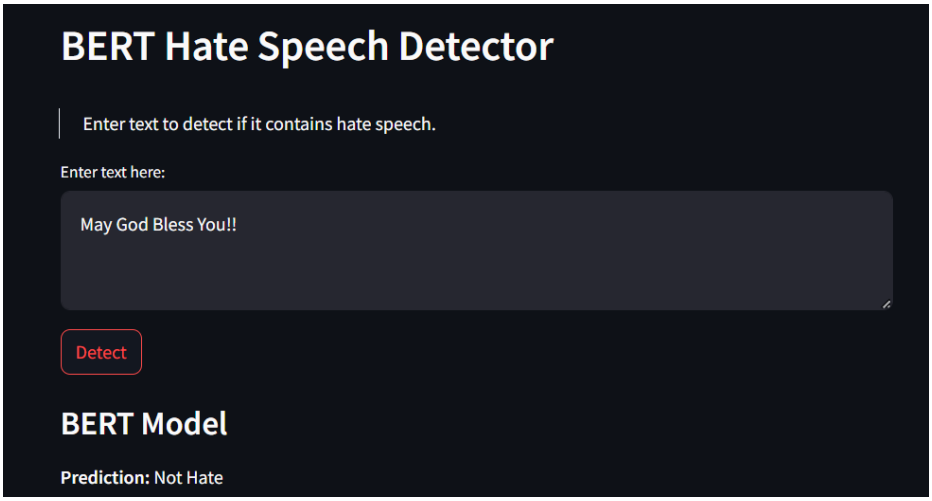
Text: People like Trump are a plague on society; they spread nothing but trouble wherever they go, corrupting everything  with their backward ways, and honestly, we'd all be better off if they were banned from participating in public life altogether.



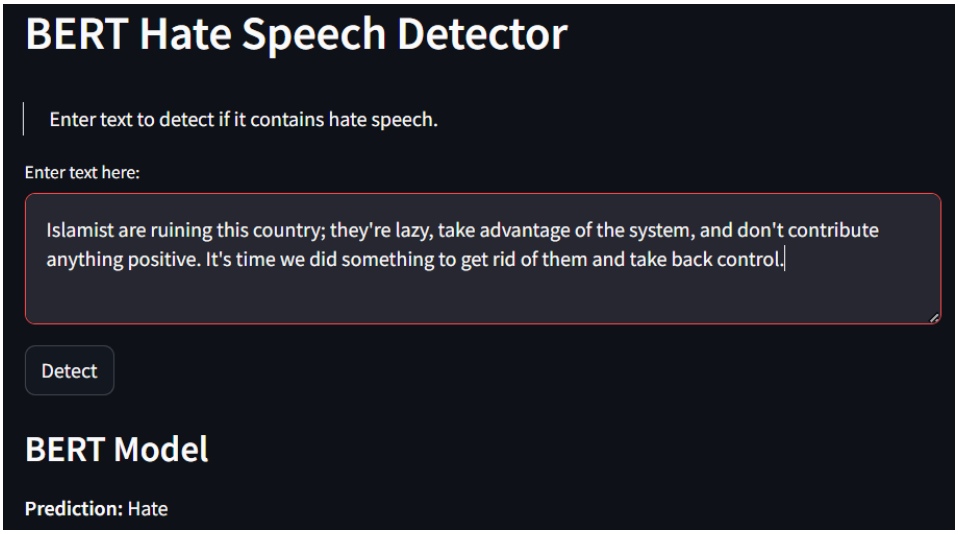
Text: Tuje Kuch aata hai ki nahi kabhi bhi kuch bhi karta hai




Text:May God Bless you!!



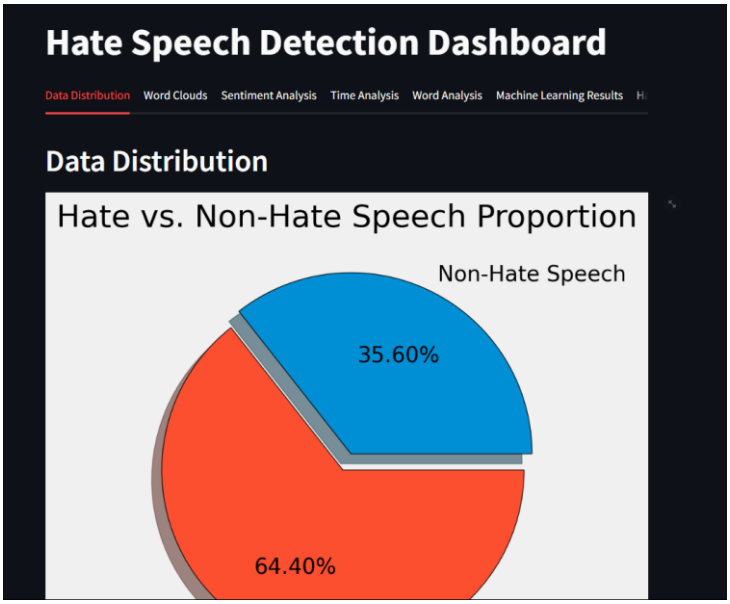

Text: Islamist are ruining this country; they're lazy, take advantage of the system, and don't contribute anything positive. It's time we did something to get rid of them and take back control.



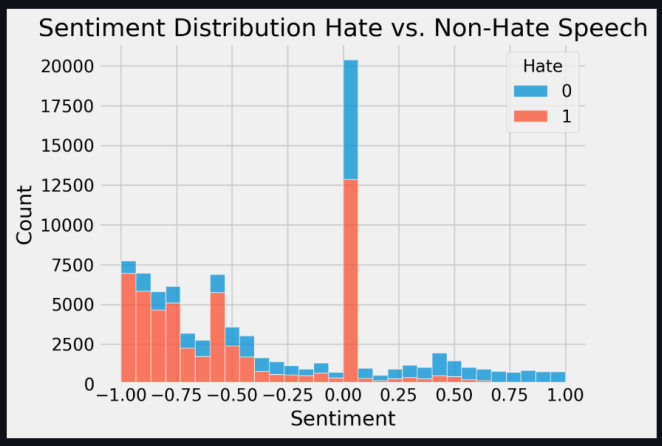
**Frontend Development**

This project's frontend, developed using Streamlit, provides an interactive dashboard for hate speech detection. The application presents multiple visualizations that analyze text data, including a distribution of hate versus non-hate speech, word clouds for commonly used terms, and sentiment analysis to highlight sentiment differences in hate speech. Temporal and frequency-based analyses reveal patterns, such as months and hours with peak hate speech occurrences, along with common bigrams and trigrams, adding insights into language trends within hate speech. These visualizations equip users with an in-depth understanding of text patterns, helping to identify possible spikes or trends in hateful content.

Additionally, the dashboard enables real-time hate speech detection through two machine learning models: a traditional model (e.g., Decision Tree) and a fine-tuned BERT model. Users can input custom text to test, with the app displaying predictions and probabilities. A separate tab presents model performance metrics, confusion matrices, and ROC curves, offering transparency into model accuracy and robustness. This frontend creates a user-friendly environment for hate speech analysis, supporting both exploratory data analysis and predictive testing with machine learning models.




## Sentiment Distribution

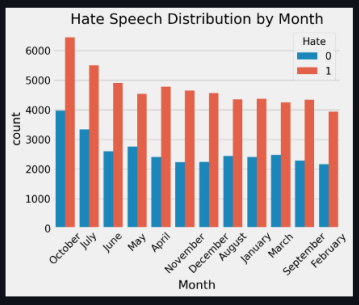



This histogram shows the distribution of sentiment scores. We observe that the sentiment scores are more negative for hate speech compared to non-hate speech.

# Hate Speech Detection Dashboard

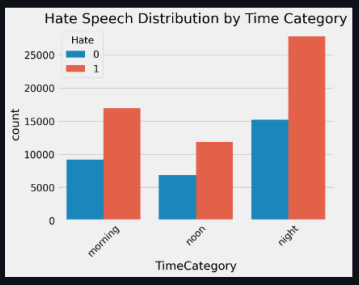
Data Distribution | Word Clouds | Sentiment Analysis | Time Analysis | Word Analysis | Machine Learning Results | H

## Monthly Distribution



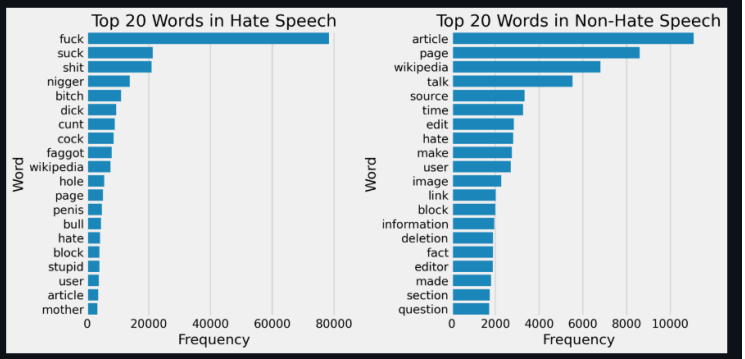

Hate speech is more prevalent in the months of October and July.

## Time Category Distribution

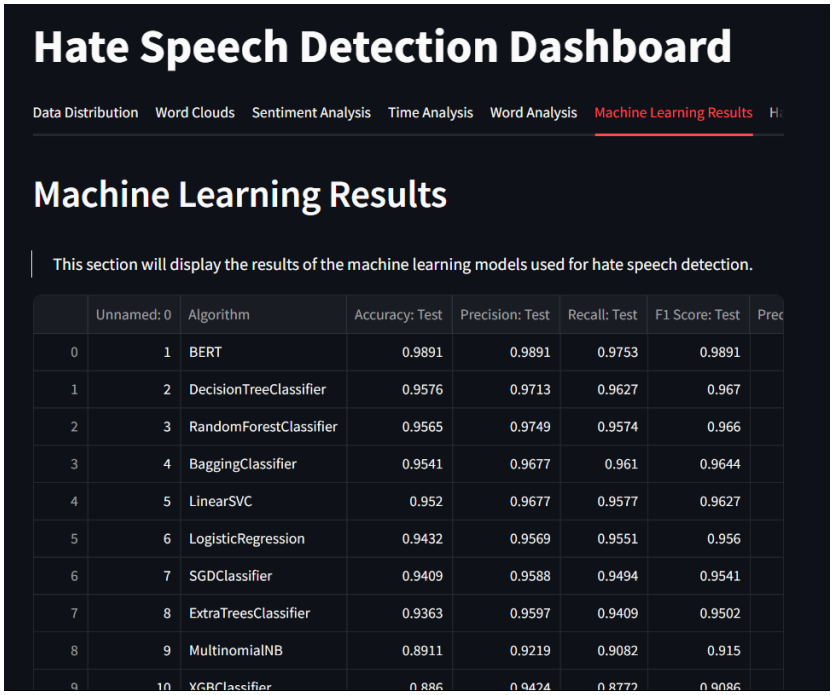



Hate speech is more prevalent in the night time.

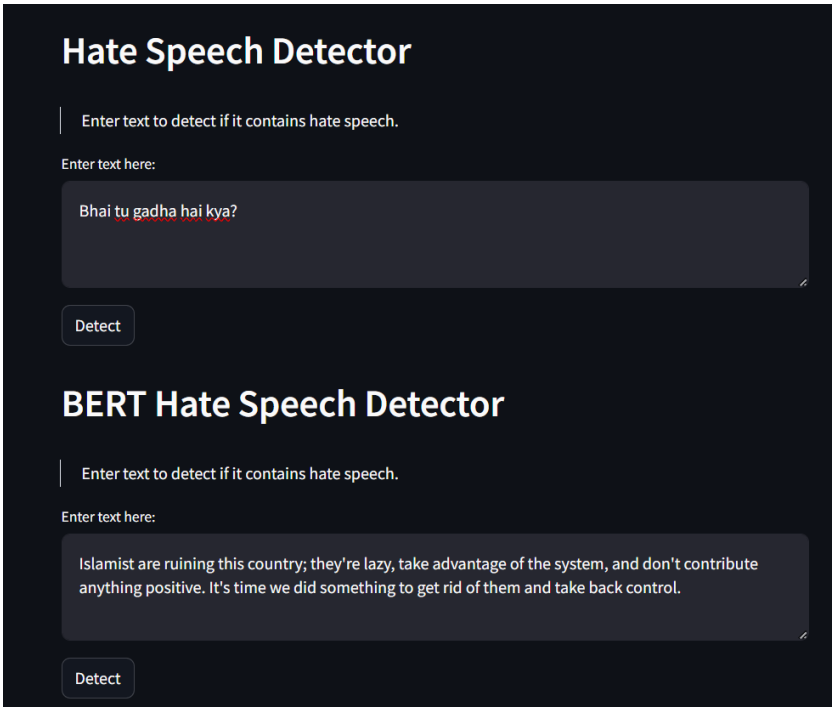
## Word Analysis




The top 20 words in hate speech and non-hate speech are shown above. We observe that the most frequent words in hate speech are more negative and offensive.




| | Unnamed: 0 | Algorithm | Accuracy: Test | Precision: Test | Recall: Test | F1 Score: Test | Prec |
|---|---|---|---|---|---|---|---|
| 0 | 1 | BERT | 0.9891 | 0.9891 | 0.9753 | 0.9891 | |
| 1 | 2 | DecisionTreeClassifier | 0.9576 | 0.9713 | 0.9627 | 0.967 | |
| 2 | 3 | RandomForestClassifier | 0.9565 | 0.9749 | 0.9574 | 0.966 | |
| 3 | 4 | BaggingClassifier | 0.9541 | 0.9677 | 0.961 | 0.9644 | |
| 4 | 5 | LinearSVC | 0.952 | 0.9677 | 0.9577 | 0.9627 | |
| 5 | 6 | LogisticRegression | 0.9432 | 0.9569 | 0.9551 | 0.956 | |
| 6 | 7 | SGDClassifier | 0.9409 | 0.9588 | 0.9494 | 0.9541 | |
| 7 | 8 | ExtraTreesClassifier | 0.9363 | 0.9597 | 0.9409 | 0.9502 | |
| 8 | 9 | MultinomialNB | 0.8911 | 0.9219 | 0.9082 | 0.915 | |
| 9 | 10 | XGBClassifier | 0.886 | 0.9424 | 0.8772 | 0.9086 | |




## Conclusion

In conclusion, this project demonstrates an effective approach to hate speech detection by leveraging both traditional machine learning models and a fine-tuned BERT model. The combination of a robust backend with comprehensive frontend visualizations allows for not only the identification of hate speech but also a deeper analysis of patterns, sentiments, and language characteristics associated with such content. By examining metrics such as accuracy, F1 score, and ROC curves, the project confirms the BERT model's suitability for handling nuanced text classification tasks, offering high accuracy and efficiency in real-time predictions.