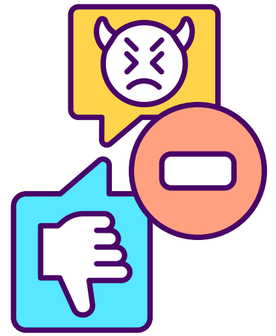
HATE SPEECH ANALYSIS AND PREDICTION



By : Banuprakash Vellingiri

Agenda

- Hate speech
- Problem Statement And Goal
- About Dataset
- Exploratory Data Analysis(EDA)
- Natural Language Processing(NLP)
- Machine Learning
- Model Evaluation
- Sugesstions

What is Hate Speech?

Hate speech refers to any form of communication, whether verbal, written, or expressed through actions, that spreads or incites **hatred, hostility, discrimination, or violence against individuals or groups .** These characteristics can include **race, ethnicity, nationality, religion, gender, sexual orientation, disability, or other identifiable traits**.

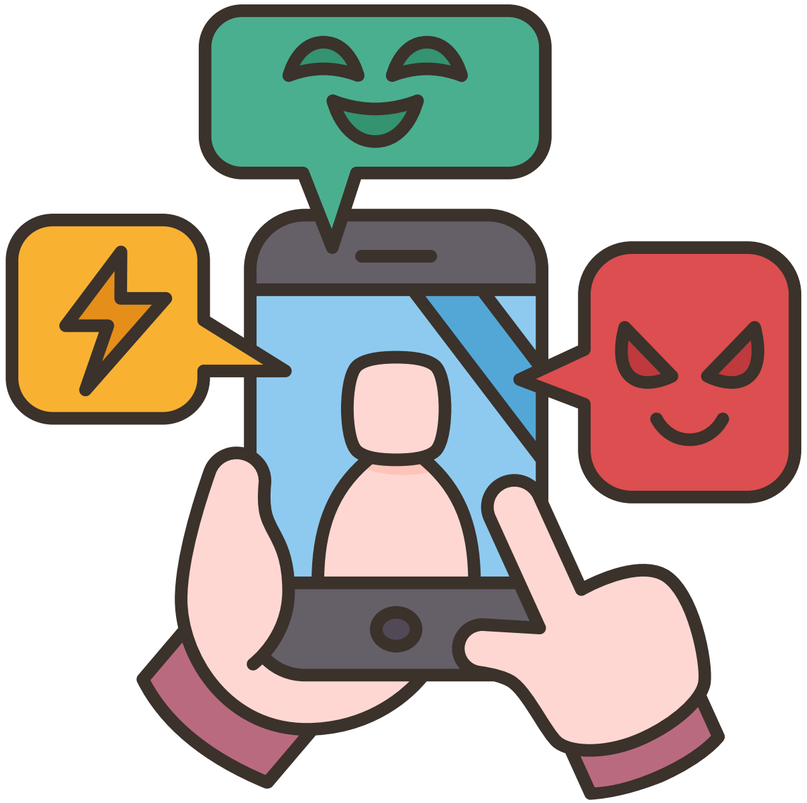
Hate speech in Social Media :

Hate speech in social media can manifest in various forms, including **text, images, memes, videos, and symbols.** It often spreads rapidly and can have profound negative impacts on **individuals' mental well-being, social cohesion, and community relations.**

Why Hate Speech Analysis?

Key Factors :

- Protection of Vulnerable Groups (**Toxic Communities**)
- Maintaining Social Cohesion
- Preventing **Violence and Extremism**
- Promoting Freedom of Expression
- Digital and Online Safety



Problem Statement :

The increasing use of social media has brought about a concerning rise in hate speech. The problem we face is how to **effectively identify and address hate speech** on social media platforms.

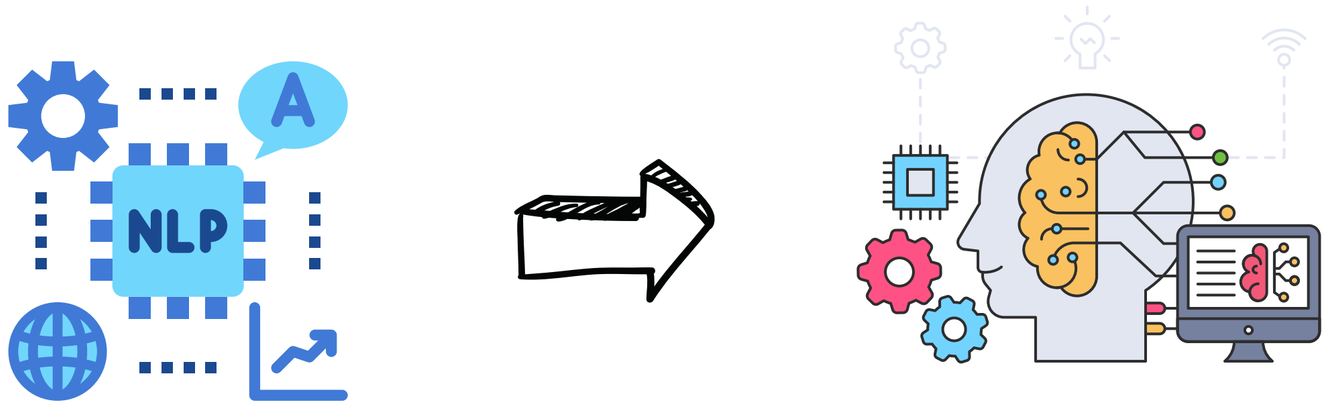
Goal :

1. Analysis :

The project aims to enhance our understanding of hate speech dynamics, **identify patterns and trends over time, profile users prone to engaging inhate speechs** and ultimately develop effective strategies to detect and mitigate hate speech online.

2. Predictive Model

The goal is to create a model that can distinguish between harmful speech and acceptable speech, ensuring a safer online environment for users. This involves leveraging technologies like **Natural Language Processing (NLP)** and **Machine Learning** to analyze text and identify harmful language patterns.



Project Benefits

- **Content Removal** : Platforms can automatically flag or remove content identified as hate speech based on the detection algorithms. This can help in preventing the spread of harmful content and maintaining community guidelines.

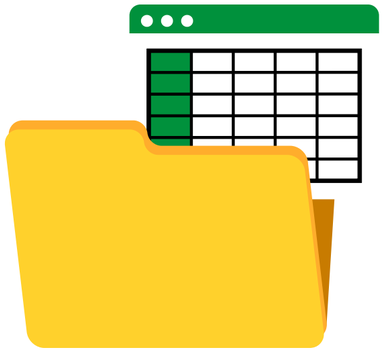
- **User Warning and Education :** Users who engage in or are exposed to hate speech can be provided with warnings about the nature of the content and educational resources about the impact and consequences of hate speech.

- **User Suspension or Ban :** Depending on the severity and frequency of hate speech occurrences, users responsible for generating or promoting hate speech may face temporary suspension or permanent ban from the platform.

- **Reporting Mechanisms :** Platforms can encourage users to report instances of hate speech, enabling the moderation team to review flagged content promptly and take appropriate actions.

About Dataset :

- The Dataset taken for this project is provided by ‘Zendsplacements‘ from guvi.com.
- The dataset consists of context posted by peoples in different subforums which are labeled with “hate” and “noHate” speech based on their sentiments.
- There are around 6 columns and 10944 rows of data



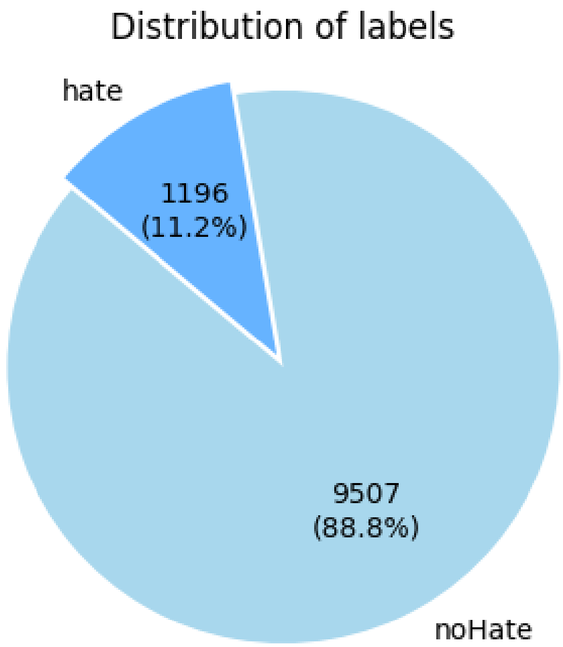
Column Description :

1. **file_id :** Unique identifier for each file containing textual context.
2. **context :** textual context posted by the user.
3. **user_id :** Identifier for the user posting the context.
4. **subforum_id :** Identifier for the subforum where the context was posted.
5. **num_contexts :** Number of contexts associated with each file.
6. **label :** Label indicating whether the context contains hate speech or not.

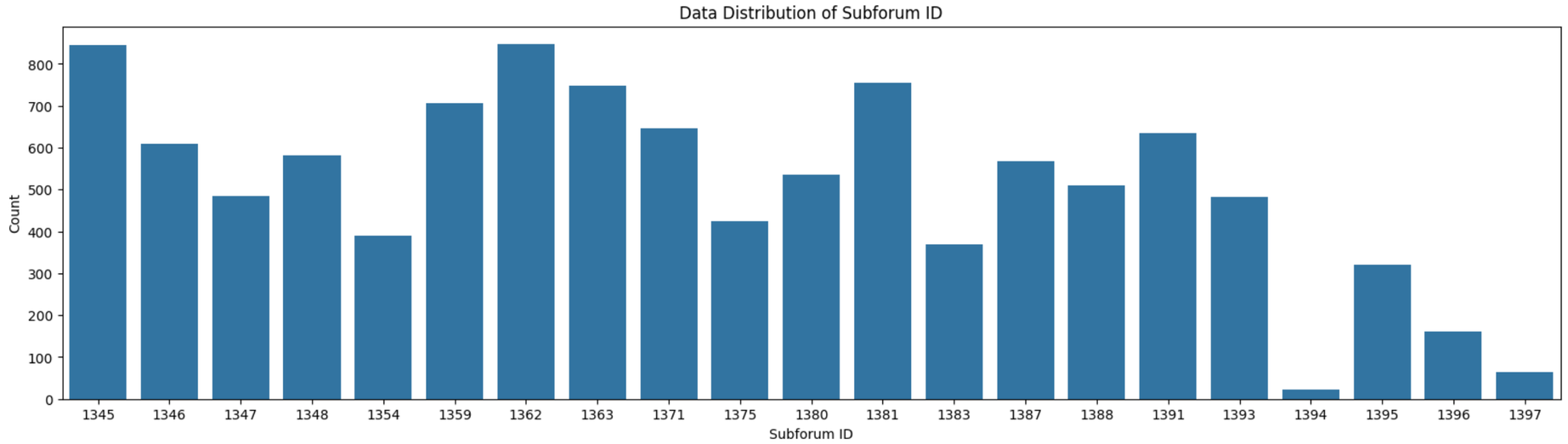
Exploratory Data Analysis (EDA)

Label Distribution

- *The dataset has imbalanced class distribution.*
- *"**noHate**" labeled data is around **88.8%** and "**hate**" labeled data is around **11.2%**.*

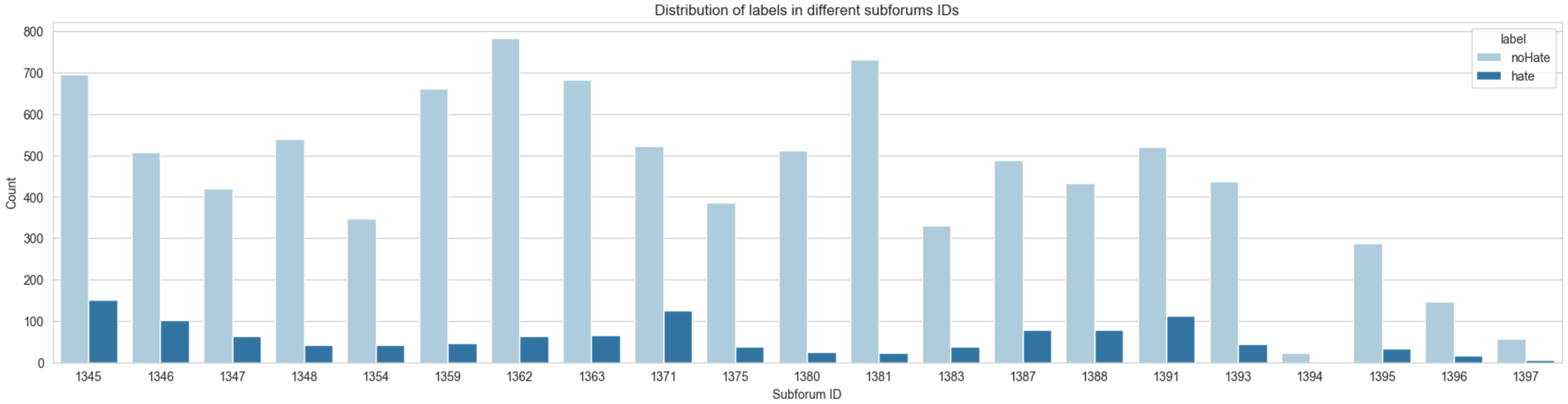


Subforum ID vs Context Distribution



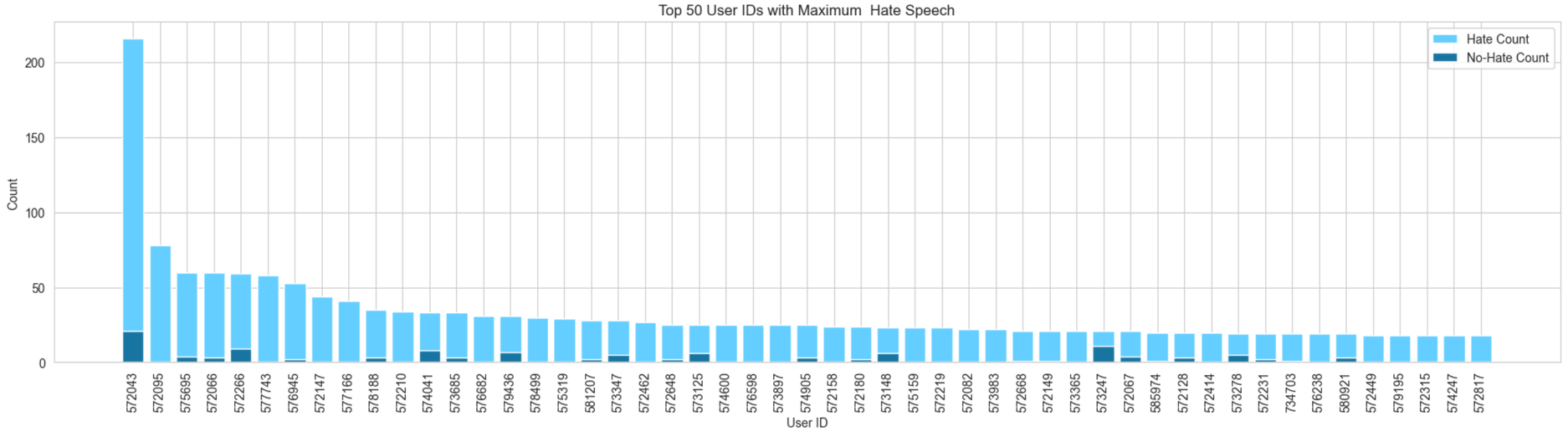
- *There are '**21**' different subforum context participated in this data.*
- *Each of the subforum IDs '**1362**' and '**1345**' contributes approximately **7.9%**of the total content. Least contribution of around **0.2%** is from subforum ID '**1394**'.*

Subforum ID vs Labels Distribution



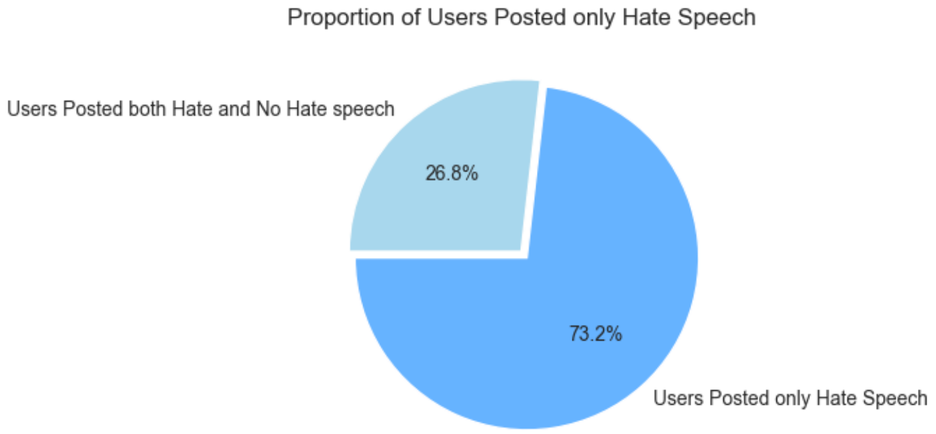
- *Within the total hate speech context, approximately* **12.5%** *originates from subforum ID 1345.*
- *Hate speech context is very less (3.0%) in subforum ID 1381 comparing with its no hate speech.*
- *There is no hate speech context in subforum ID 1394.It is noted that very few number of context is posted in this subforum.*
- *More positive context are posted in subforum ID 1362 among all subforums.* **92.4%** *"nohate" context are posted among total context in this subforum.*
- **Hate and no hate ratio is higher** *in subforum ID* **1371** *by comparing with all.*
- *Subforum IDs* **1345,1371,1391,1346,1388** *are top IDs with more hate speech.*

Top 50 User IDs with Maximum Hate Speech



- *User Id '**572043**' has posted **18%** of hate speech among total.*

User IDs and Hate Speech Distribution



- *Aprroximately **73.2%** users only posted the Hate Speech.*

Natural Language Processing (NLP)

Initially Context undergoes various NLP techniques such as ,

- ***Tokenization (sentence /word)***
- ***Text Normalization***
- ***Stopwords Removal***
- ***Stemming***
- ***Lemmatization,***
- ***Part-of-Speech (POS) Tagging***
- ***Named Entity Recognition (NER)***

as required .In simple words, the context is normalized to the required format to pursue Machine Learning.

Raw context :

```
** As of March 13th , 2014 , the booklet had been downloaded over 18,300 times and counting .
** Thank you in advance. : ) Download the youtube `` description box '' info text file below @ http://www.mediafi
** In order to help increase the booklets downloads , it would be great if all Stormfronters who had YouTube acco
** ( Simply copy and paste the following text into your YouTube videos description boxes. )
** Click below for a FREE download of a colorfully illustrated 132 page e-book on the Zionist-engineered INTENTIO
** Click on the `` DOWNLOAD ( 7.42 MB ) '' green banner link .
```

Processed Context :

```
** march booklet downloaded time counting
** thank advance download youtube description box info text file
** order help increase booklet downloads would great stormfronters youtube account c
** simply copy paste following text youtube video description box
** download colorfully illustrated page intentional destruction western civilizatior
```

Word Cloud :

Word Cloud of Processed Context

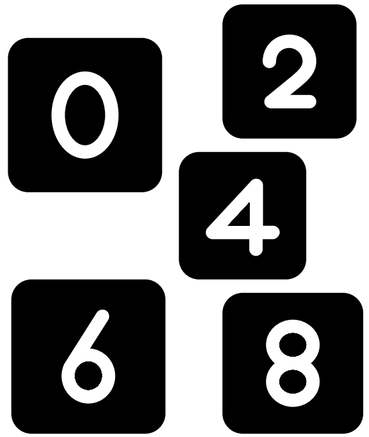
Text Representation

Processed CONTEXT → ***Numerical VECTORS***

Here 3 different vectorization techniques such as,

- ***Count Vectorizer***
- ***Tf-idf Vectorizer***
- ***Word2Vec***

are used and trained with different models and checked for accuracy

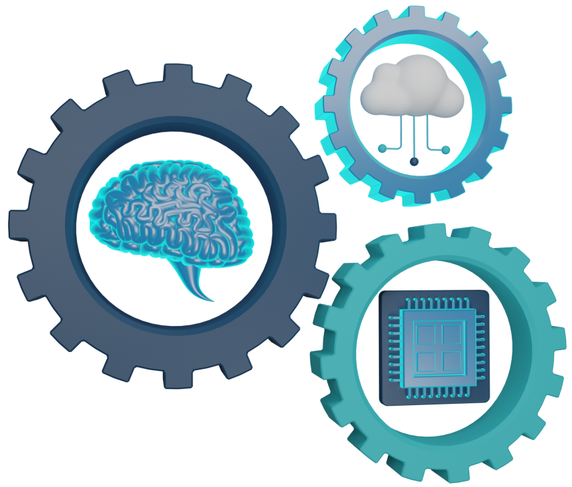


Machine Learning

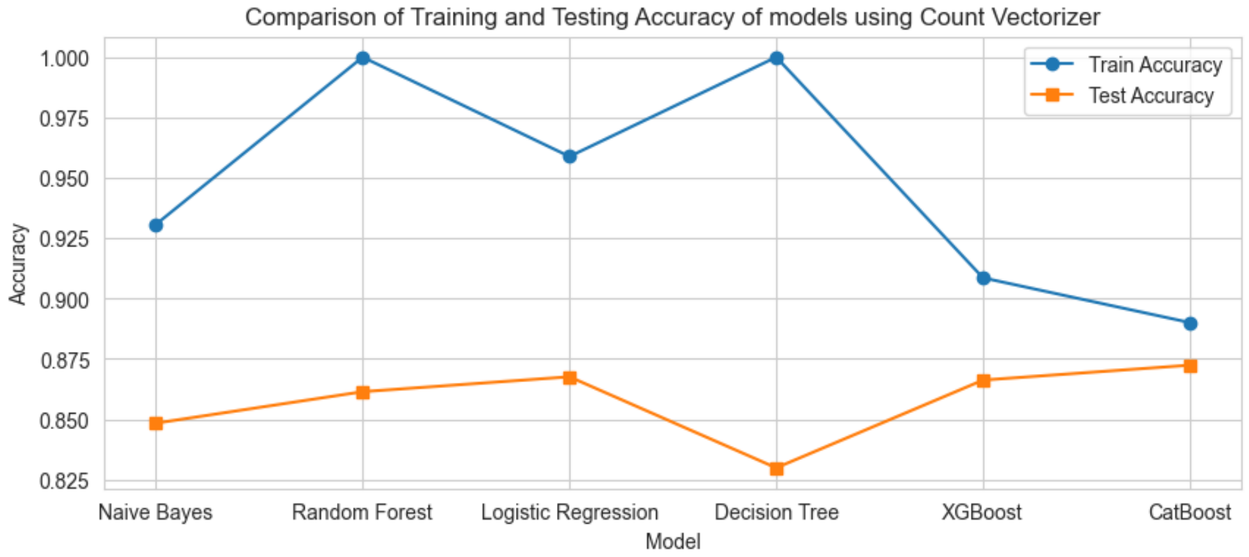
This is a classification problem, so different classification algorithms are used and checked for the performance.

Algorithms used :

- **Naive Bayes**
- **Random Forest**
- **Logistic Regression**
- **Decision Tree**
- **XGBoost**
- **CatBoost**

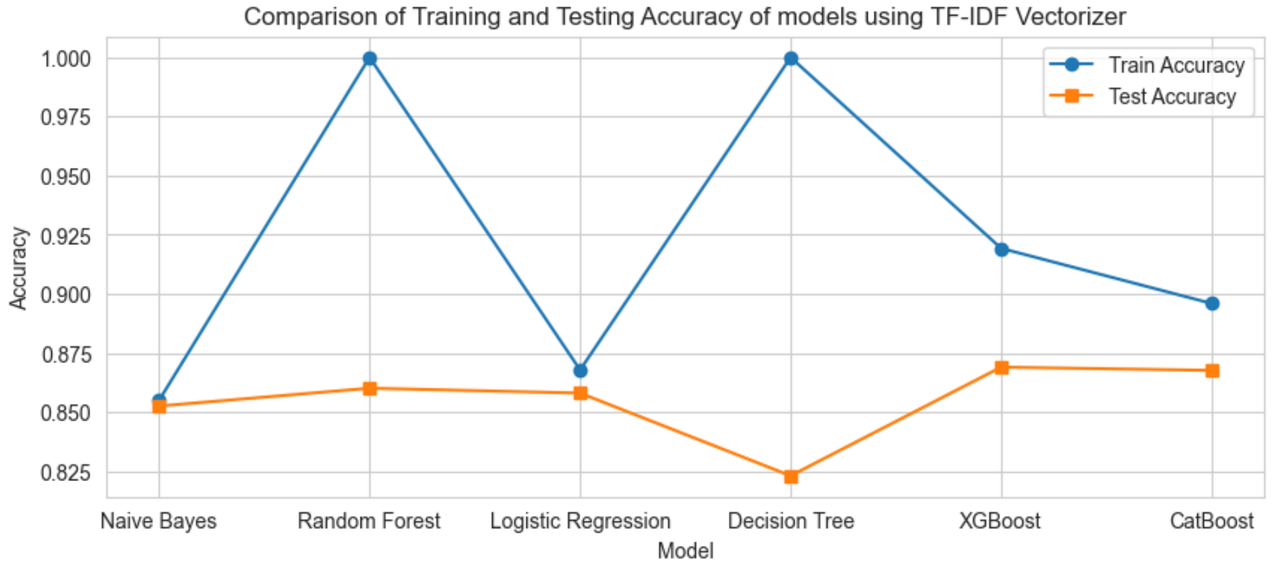


Model perfromance with Count Vectorizer



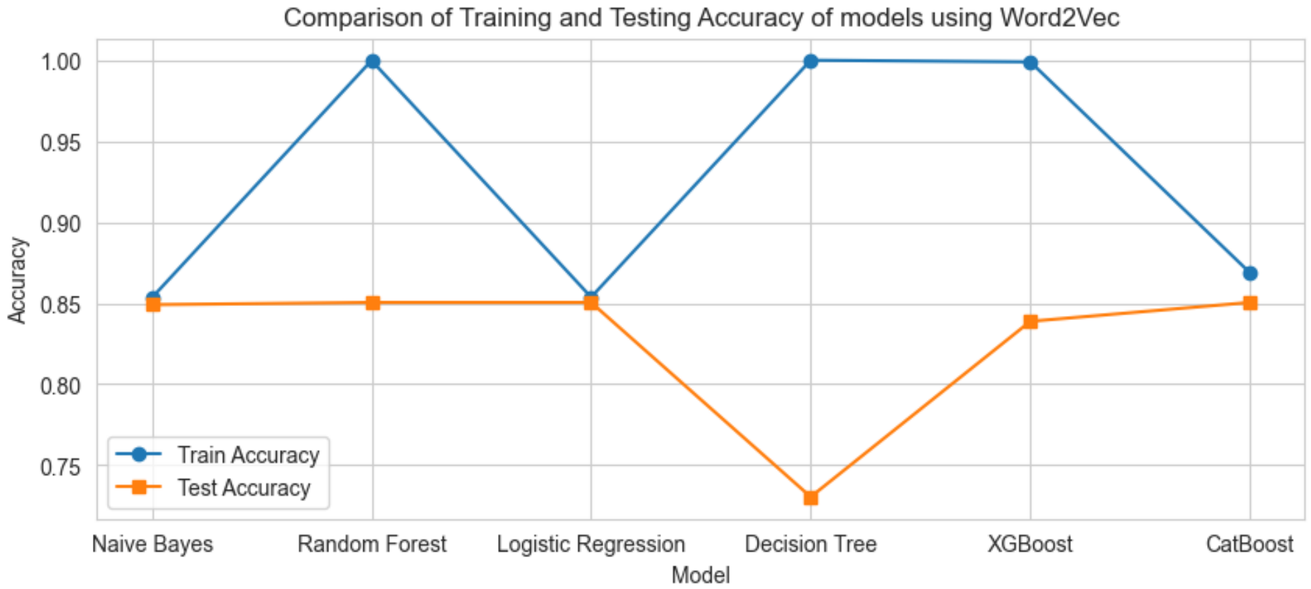
	Model	Train Accuracy_CountVec	Test Accuracy_CountVec	F1 Score
0	Naive Bayes	0.930727	0.848422	0.629261
1	Random Forest	1.000000	0.861454	0.554886
2	Logistic Regression	0.958848	0.867627	0.633859
3	Decision Tree	1.000000	0.827160	0.592903
4	XGBoost	0.908608	0.866255	0.627806
5	CatBoost	0.890089	0.872428	0.607293

Model Performance with TF-IDF Vectorizer



	Model	Train Accuracy_TF-IDF	Test Accuracy_TF-IDF	F1 Score
0	Naive Bayes	0.854938	0.852538	0.460200
1	Random Forest	1.000000	0.860082	0.533340
2	Logistic Regression	0.867798	0.858025	0.505563
3	Decision Tree	1.000000	0.832647	0.618824
4	XGBoost	0.919239	0.868999	0.639835
5	CatBoost	0.895919	0.867627	0.596706

Model Performance with Word2Vec



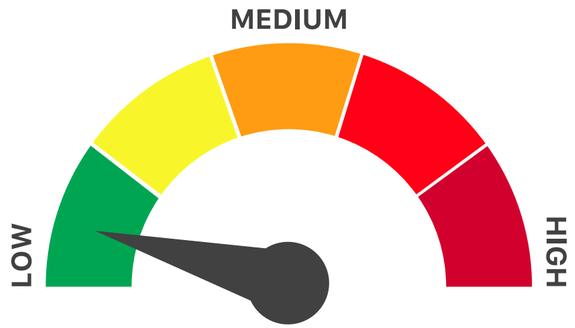
	Model	Train Accuracy_Word2Vec	Test Accuracy_Word2Vec	F1 Score
0	Naive Bayes	0.853052	0.848422	0.458998
1	Random Forest	1.000000	0.850480	0.459600
2	Logistic Regression	0.853738	0.850480	0.459600
3	Decision Tree	1.000000	0.726337	0.504959
4	XGBoost	0.999314	0.840878	0.465191
5	CatBoost	0.868999	0.850480	0.459600

Observation from Metrics :

- *Comparing accuracies of different models with 3 vectorization techniques The Tree based models like **"Random Forest" and "Decision Tree" ,"XGBoost"**models have poor performance except **"CatBoost"**.Their training and testing accuracy around **9%-11%** in difference which indicated the models are "**overfitted**".*

- *Logistic Regression and Naive Bayes performance is **poor** while using **Count Vectorizer**.*

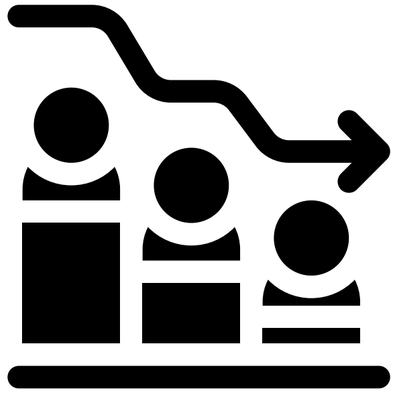
- *Eventhough the accuracy is low, Logistic Regression and Naive Bayes models Performed decently in both training and testing in TF-IDF and Word2Vec.*



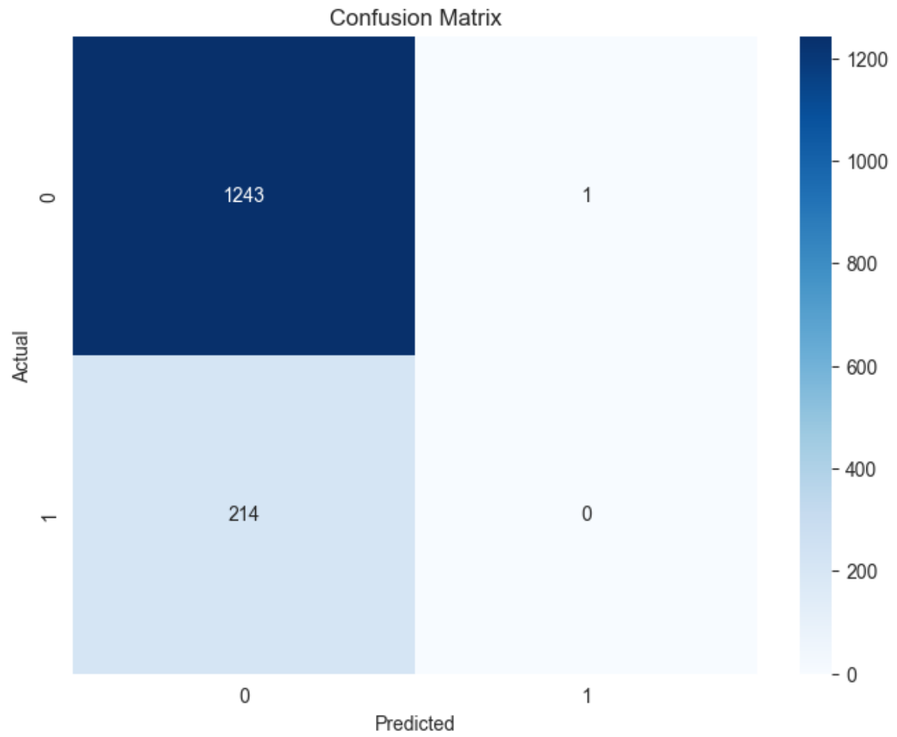
F1-Score :

- *F1 Score all models at different text representation texchniques is "**Very Very Poor**".The model struggles to identify the minority class (hate speech) even if it performs well on the majority class.Model tends to have **low precsion** indicates that the model makes too many false positive predictions. **Low recall** indicates that the model misses many positive instances and has a high rate of false negatives.*

- *This Poor F1-Score is due to "**Class Imbalance**" in the dataset.Well balanced classes in the dataset significantly increases the prediction accuracy as well as F1-Score for classification tasks.In addtion to this **data quality(with less noise)** is also important.*



Confusion Matrix :



Logistic Regression model by using TF-IDF Vectorizer

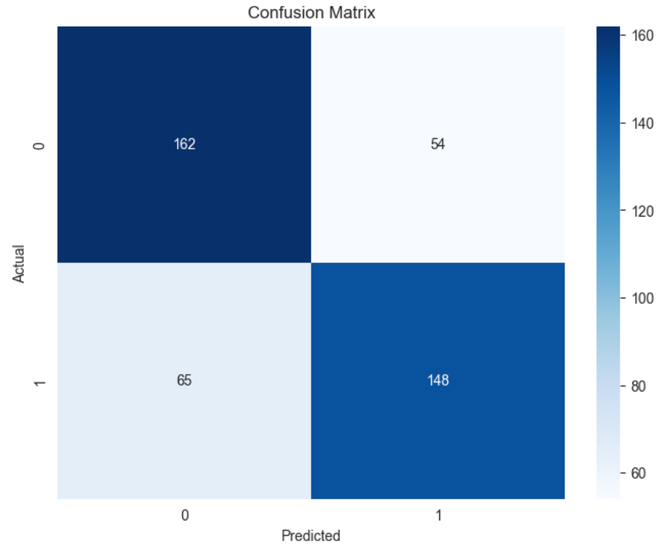
Experiment:

*Trying to train the model by balanced classes by **Undersampling the Majority Class (No Hate)**.*

```
Classification Report :
              precision    recall  f1-score   support

           0       0.71      0.75      0.73       216
           1       0.73      0.69      0.71       213

    accuracy                           0.72       429
   macro avg       0.72      0.72      0.72       429
weighted avg       0.72      0.72      0.72       429
```



	0	1	2	3	4	5	6	7	8	9	10	11
Actual Class	hate	noHate	hate	hate	hate	hate	hate	noHate	noHate	noHate	hate	noHate
Predicted Class	hate	noHate	noHate	hate	noHate	hate	hate	noHate	hate	noHate	noHate	noHate

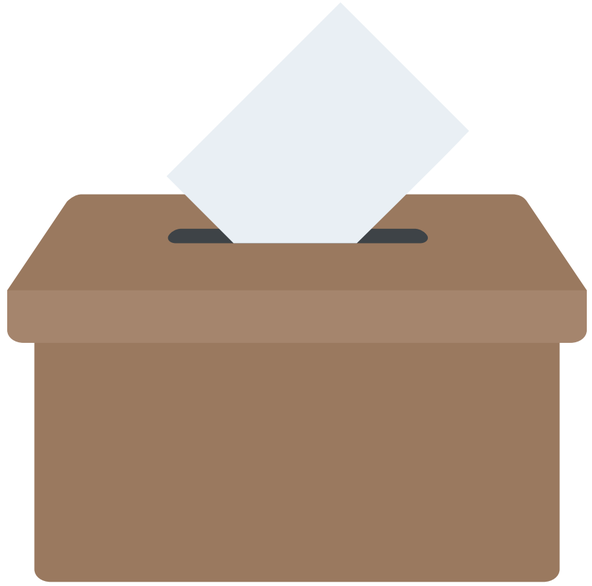
Logistic Regression model by using TF-IDF Vectorizer

Observation :

- *By Training the logistic regression model with balanced class* **F1-score increases** upto some level **(0.72).** *but* ***still it is a low score****.*
- *On the other hand* **accuracy is decreased to 0.72.** *This is because of insufficient data for training.*

Suggestions :

- *In future the model is trained with **sufficienct and detailed data** to achieve good scores.*
- *With huge dataset , **Deep Learning techniques** are also used to train the model to get sigificant performance.*



Banuprakash Vellingiri