# Abstract

This project aims to develop a model that can accurately identify hate speech in text data. By leveraging machine learning algorithms and natural language processing techniques, the objective is to differentiate between hateful and non-hateful content to mitigate the negative impact of hate speech on social platforms. The developed system is trained and tested using a labeled dataset to achieve optimal performance.

# Objective

The primary objective of this project is to develop a machine learning model capable of detecting hate speech in textual data. This involves preprocessing text, training an appropriate model, evaluating its performance, and ensuring that it can be effectively used for hate speech moderation.

# Introduction

Hate speech is a form of expression that can promote discrimination and violence, especially on social media platforms. As the online community grows, the need for automated tools to detect and mitigate hate speech becomes increasingly vital. This project addresses this need by developing a hate speech detection model using machine learning and natural language processing.

# Methodology

Data Collection: A labeled dataset containing examples of hate speech and non-hate speech was used for training and testing. Data Preprocessing: Text data was cleaned by removing punctuation, converting to lowercase, and removing stop words. Feature Extraction: Techniques like TF-IDF or word embeddings were employed to transform the text into a numerical form suitable for model training. Model Training: Various machine learning models (e.g., logistic regression, SVM, or deep learning) were trained using the preprocessed data. Evaluation: The models were evaluated using metrics like accuracy, precision, recall, and F1-score.

In [1]:

```
import warnings
warnings.filterwarnings("ignore")
```

In [2]:

```
import numpy as np
import pandas as pd
```

In [3]:

```
dataset=pd.read_csv("C:\\Users\\ASUS\\Downloads\\twitter_data.csv")
dataset
```

Out[3]:

|  | Unnamed: 0 | count | hate_speech | offensive_language | neither | class | tweet |
|---|---|---|---|---|---|---|---|
| **0** | 0 | 3 | 0 | 0 | 3 | 2 | !!! RT @mayasolovely: As a woman you shouldn't... |
| **1** | 1 | 3 | 0 | 3 | 0 | 1 | !!!!! RT @mleew17: boy dats cold...tyga dwn ba... |
| **2** | 2 | 3 | 0 | 3 | 0 | 1 | !!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby... |
| **3** | 3 | 3 | 0 | 2 | 1 | 1 | !!!!!!!!! RT @C_G_Anderson: @viva_based she lo... |
| **4** | 4 | 6 | 0 | 6 | 0 | 1 | !!!!!!!!!!!!! RT @ShenikaRoberts: The shit you... |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **24778** | 25291 | 3 | 0 | 2 | 1 | 1 | you's a muthaf***in lie &#8220;@LifeAsKing: @2... |
| **24779** | 25292 | 3 | 0 | 1 | 2 | 2 | you've gone and broke the wrong heart baby, an... |
| **24780** | 25294 | 3 | 0 | 3 | 0 | 1 | young buck wanna eat!!.. dat nigguh like I ain... |
| **24781** | 25295 | 6 | 0 | 6 | 0 | 1 | youu got wild bitches tellin you lies |
| **24782** | 25296 | 3 | 0 | 0 | 3 | 2 | ~~Ruffled \| Ntac Eileen Dahlia - Beautiful col... |

24783 rows × 7 columns

In [4]:

```
dataset.isnull().sum()
```

Out[4]:

```
Unnamed: 0            0
count                 0
hate_speech           0
offensive_language    0
neither               0
class                 0
tweet                 0
dtype: int64
```

In [5]:

```
dataset.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24783 entries, 0 to 24782
```

```
Data columns (total 7 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Unnamed: 0          24783 non-null  int64 
 1   count               24783 non-null  int64 
 2   hate_speech         24783 non-null  int64 
 3   offensive_language  24783 non-null  int64 
 4   neither             24783 non-null  int64 
 5   class               24783 non-null  int64 
 6   tweet               24783 non-null  object
dtypes: int64(6), object(1)
memory usage: 1.3+ MB
```

In [6]:

```
dataset.describe()
```

Out[6]:

| | Unnamed: 0 | count | hate_speech | offensive_language | neither | class |
|---|---|---|---|---|---|---|
| **count** | 24783.000000 | 24783.000000 | 24783.000000 | 24783.000000 | 24783.000000 | 24783.000000 |
| **mean** | 12681.192027 | 3.243473 | 0.280515 | 2.413711 | 0.549247 | 1.110277 |
| **std** | 7299.553863 | 0.883060 | 0.631851 | 1.399459 | 1.113299 | 0.462089 |
| **min** | 0.000000 | 3.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 6372.500000 | 3.000000 | 0.000000 | 2.000000 | 0.000000 | 1.000000 |
| **50%** | 12703.000000 | 3.000000 | 0.000000 | 3.000000 | 0.000000 | 1.000000 |
| **75%** | 18995.500000 | 3.000000 | 0.000000 | 3.000000 | 0.000000 | 1.000000 |
| **max** | 25296.000000 | 9.000000 | 7.000000 | 9.000000 | 9.000000 | 2.000000 |

In [7]:

```
dataset["labels"]=dataset["class"].map({0:"Hate Speech",1:"Offensive",2: "No hate or offensive"})
```

In [8]:

```
dataset
```

Out[8]:

| | Unnamed: 0 | count | hate_speech | offensive_language | neither | class | tweet | labels |
|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 3 | 0 | 0 | 3 | 2 | !!! RT @mayasolovely: As a woman you shouldn't... | No hate or offensive |
| **1** | 1 | 3 | 0 | 3 | 0 | 1 | !!!!! RT @mleew17: boy dats cold...tyga dwn ba... | Offensive |

|  | Unnamed: 0 | count | hate_speech | offensive_language | neither | class | tweet | labels |
|---|---|---|---|---|---|---|---|---|
| **2** | 2 | 3 | 0 | 3 | 0 | 1 | !!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby... | Offensive |
| **3** | 3 | 3 | 0 | 2 | 1 | 1 | !!!!!!!!! RT @C_G_Anderson: @viva_based she lo... | Offensive |
| **4** | 4 | 6 | 0 | 6 | 0 | 1 | !!!!!!!!!!!!! RT @ShenikaRoberts: The shit you... | Offensive |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **24778** | 25291 | 3 | 0 | 2 | 1 | 1 | you's a muthaf***in lie &#8220;@LifeAsKing: @2... | Offensive |
| **24779** | 25292 | 3 | 0 | 1 | 2 | 2 | you've gone and broke the wrong heart baby, an... | No hate or offensive |
| **24780** | 25294 | 3 | 0 | 3 | 0 | 1 | young buck wanna eat!!.. dat nigguh like I ain... | Offensive |
| **24781** | 25295 | 6 | 0 | 6 | 0 | 1 | youu got wild bitches tellin you lies | Offensive |
| **24782** | 25296 | 3 | 0 | 0 | 3 | 2 | ~~Ruffled \| Ntac Eileen Dahlia - Beautiful col... | No hate or offensive |

24783 rows × 8 columns

In [9]:

```
data=dataset[["tweet","labels"]]
```

In [10]:

```
data
```

Out[10]:

|  | tweet | labels |
|---|---|---|
| **0** | !!! RT @mayasolovely: As a woman you shouldn't... | No hate or offensive |
| **1** | !!!!! RT @mleew17: boy dats cold...tyga dwn ba... | Offensive |
| **2** | !!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby... | Offensive |
| **3** | !!!!!!!!! RT @C_G_Anderson: @viva_based she lo... | Offensive |
| **4** | !!!!!!!!!!!!! RT @ShenikaRoberts: The shit you... | Offensive |
| **...** | ... | ... |

|  | tweet | labels |
|---|---|---|
| **24778** | you's a muthaf***in lie &#8220;@LifeAsKing: @2... | Offensive |
| **24779** | you've gone and broke the wrong heart baby, an... | No hate or offensive |
| **24780** | young buck wanna eat!!.. dat nigguh like I ain... | Offensive |
| **24781** | youu got wild bitches tellin you lies | Offensive |
| **24782** | ~~Ruffled \| Ntac Eileen Dahlia - Beautiful col... | No hate or offensive |

24783 rows × 2 columns

In [11]:

```
import re
import nltk
import string
nltk.download('stopwords')
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\ASUS\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[11]:

```
True
```

In [12]:

```
from nltk.corpus import stopwords
stopwords=set(stopwords.words("english"))
```

In [13]:

```
Stemmer=nltk.SnowballStemmer("english")
```

# Data Cleaning:

In [14]:

```
def clean(text):
    text = str(text).lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub(r"\@w+|\#",'',text)
    text = re.sub(r"[^\w\s]",'',text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    text = [word for word in text.split(' ') if word not in stopwords]
#removing stopwords
    text = " ".join(text)
    text = [Stemmer.stem(word) for word in text.split(' ')]
    text = " ".join(text)
    return text
```

In [15]:

```
data['tweet'] = data['tweet'].apply(clean)
```

```
data
```

Out[15]:

| | tweet | labels |
|---|---|---|
| **0** | rt mayasolov woman shouldnt complain clean ho... | No hate or offensive |
| **1** | rt boy dat coldtyga dwn bad cuffin dat hoe ... | Offensive |
| **2** | rt urkindofbrand dawg rt ever fuck bitch sta... | Offensive |
| **3** | rt cganderson vivabas look like tranni | Offensive |
| **4** | rt shenikarobert shit hear might true might f... | Offensive |
| **...** | ... | ... |
| **24778** | yous muthafin lie coreyemanuel right tl tras... | Offensive |
| **24779** | youv gone broke wrong heart babi drove redneck... | No hate or offensive |
| **24780** | young buck wanna eat dat nigguh like aint fuck... | Offensive |
| **24781** | youu got wild bitch tellin lie | Offensive |
| **24782** | ruffl ntac eileen dahlia beauti color combin... | No hate or offensive |

24783 rows × 2 columns

In [16]:

```
X=np.array(data['tweet'])
y=np.array(data['labels'])
```

In [17]:

```
X
```

Out[17]:

```
array([' rt mayasolov woman shouldnt complain clean hous amp man alway take
trash',
       ' rt  boy dat coldtyga dwn bad cuffin dat hoe  place',
       ' rt urkindofbrand dawg rt  ever fuck bitch start cri confus shit',
       ..., 'young buck wanna eat dat nigguh like aint fuckin dis',
       'youu got wild bitch tellin lie',
       'ruffl  ntac eileen dahlia  beauti color combin pink orang yellow amp
white coll '],
      dtype=object)
```

In [18]:

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
```

In [19]:

```
cv =CountVectorizer()
X=cv.fit_transform(X)
```

In [20]:

```
X
```

Out[20]:

```
<24783x25697 sparse matrix of type '<class 'numpy.int64'>'
        with 197861 stored elements in Compressed Sparse Row format>
```

In [21]:

```
X_train ,X_test,y_train,y_test=train_test_split(X,y,test_size=0.33,random_sta
te=42)
```

In [22]:

```
X_train
```

Out[22]:

```
<16604x25697 sparse matrix of type '<class 'numpy.int64'>'
	with 132620 stored elements in Compressed Sparse Row format>
```

In [23]:

```
from sklearn.tree import DecisionTreeClassifier
```

In [24]:

```
dt=DecisionTreeClassifier()
dt.fit(X_train,y_train)
```

Out[24]:

```
DecisionTreeClassifier()
```

In [25]:

```
y_pred=dt.predict(X_test)
```

In [26]:

```
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(y_test,y_pred)
cm
```

Out[26]:

```
array([[ 154,   38,  273],
       [  33, 1161,  185],
       [ 227,  256, 5852]], dtype=int64)
```

In [27]:

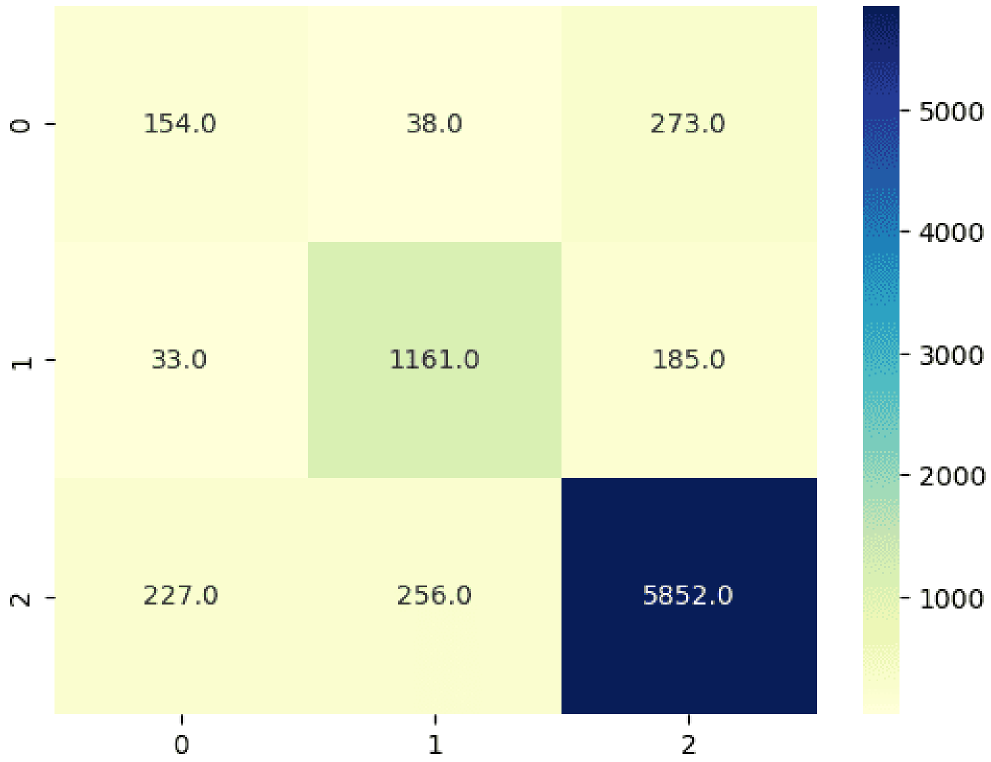
```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [28]:

```
sns.heatmap(cm,annot = True,fmt=".1f",cmap="YlGnBu")
```

Out[28]:

```
<AxesSubplot:>
```



In [29]:

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,y_pred)
```

Out[29]:

```
0.8762684924807433
```

# Model Sample :

In [30]:

```
sample="kill all the people"
sample=clean(sample)
```

In [31]:

```
sample
```

Out[31]:

```
'kill peopl'
```

In [32]:

```
data1=cv.transform([sample]).toarray()
```

In [33]:

```
data1
```

Out[33]:

```
array([[0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [34]:

```
dt.predict(data1)
```

Out[34]:

```
array(['Hate Speech'], dtype=object)
```

# Conclusion

The project successfully developed a model that can classify text as hate speech or non-hate speech with reasonable accuracy. The performance metrics indicate that the model can be a helpful tool for moderating online content, but further improvement can be achieved by increasing the dataset size and exploring more advanced models.

In [ ]: