PROJECT : Hate-Speech-detection-using-Transformers-Deep-Learning

Group Name: Hate Speech Detective

Members:

Name: VARSHIT MANEPALLI

Email: varshitmanepalli1810@gmail.com

Country: INDIA

College/Company: Stevens Institute of Technology

Specialization: NLP

Batch code : LISUM32

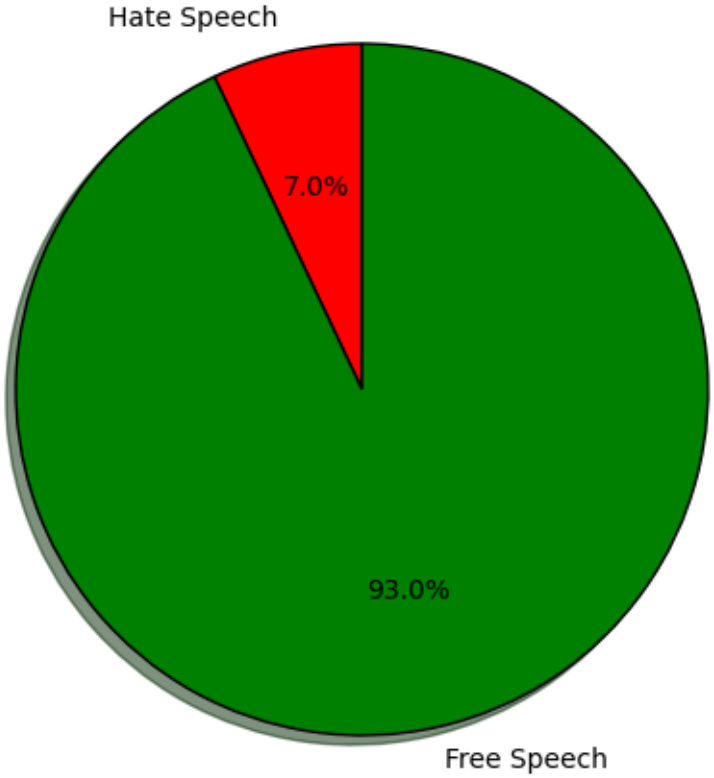
Problem Description

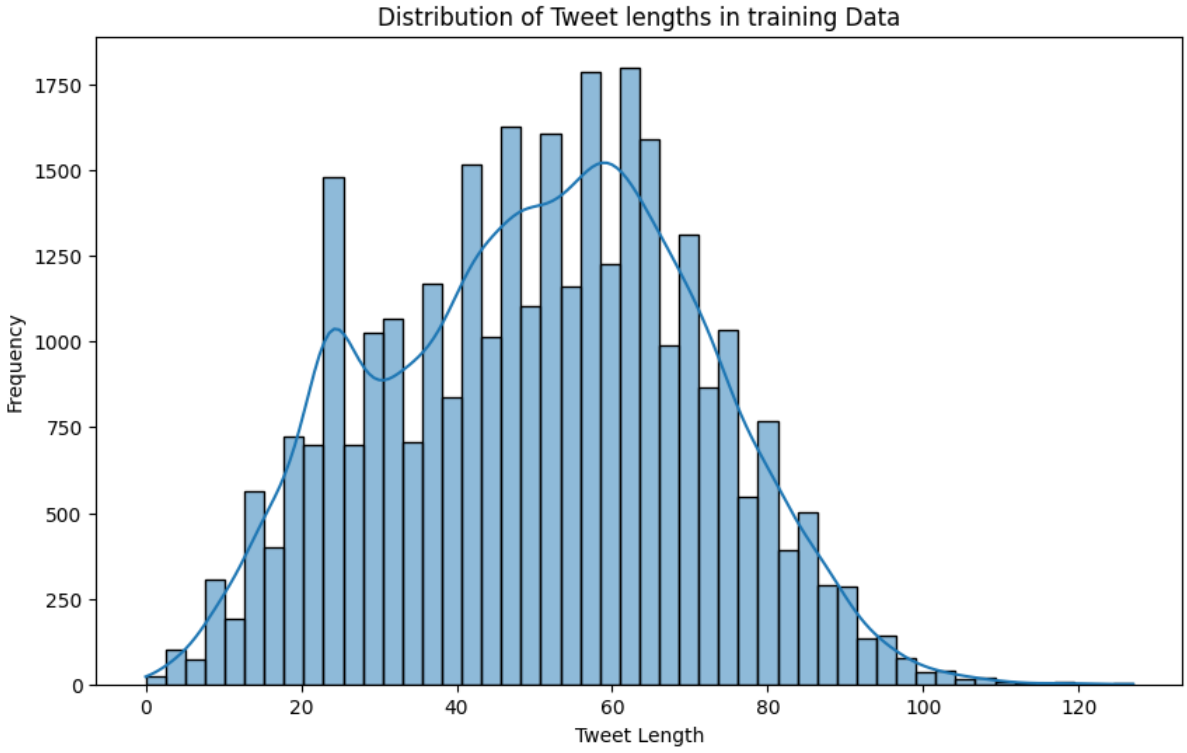
Hate speech detection aims to identify and classify statements that contain offensive, derogatory, or discriminatory language directed towards individuals or groups based on their identity factors such as religion, ethnicity, nationality, race, color, ancestry, sex, or other identity factors. This project involves developing a machine learning model to detect hate speech in Twitter tweets.

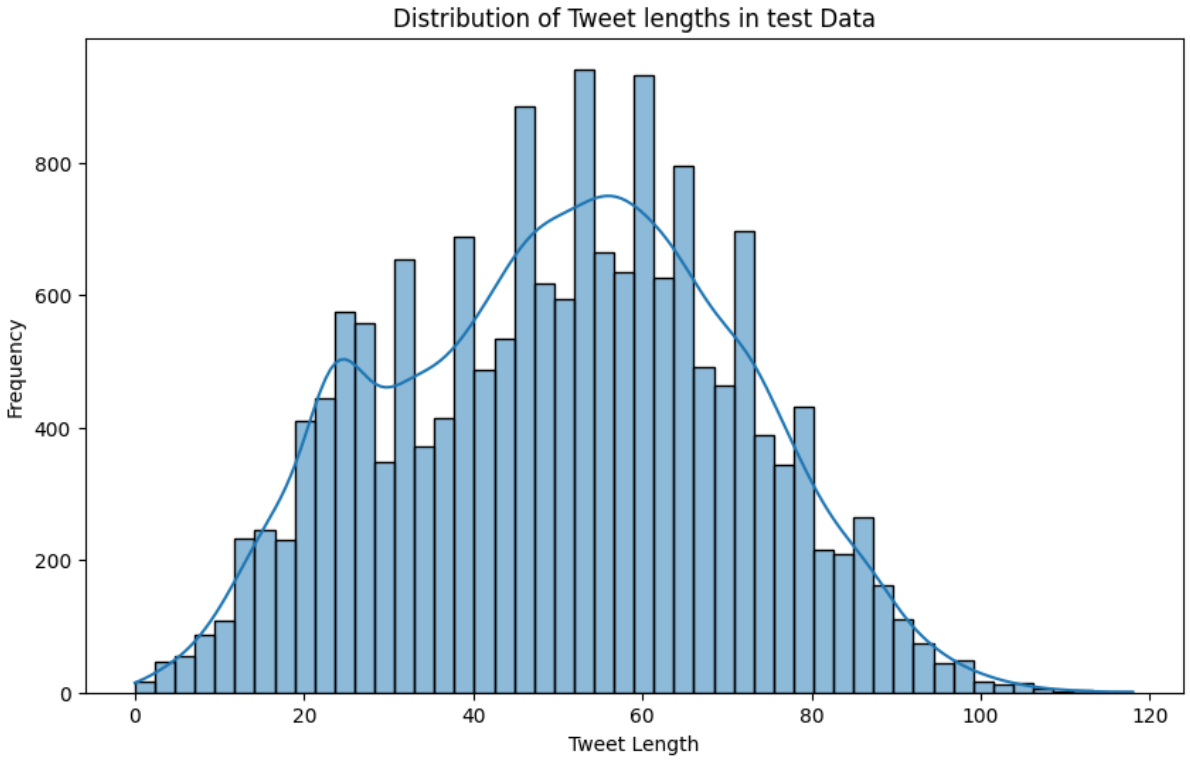
Business Understanding

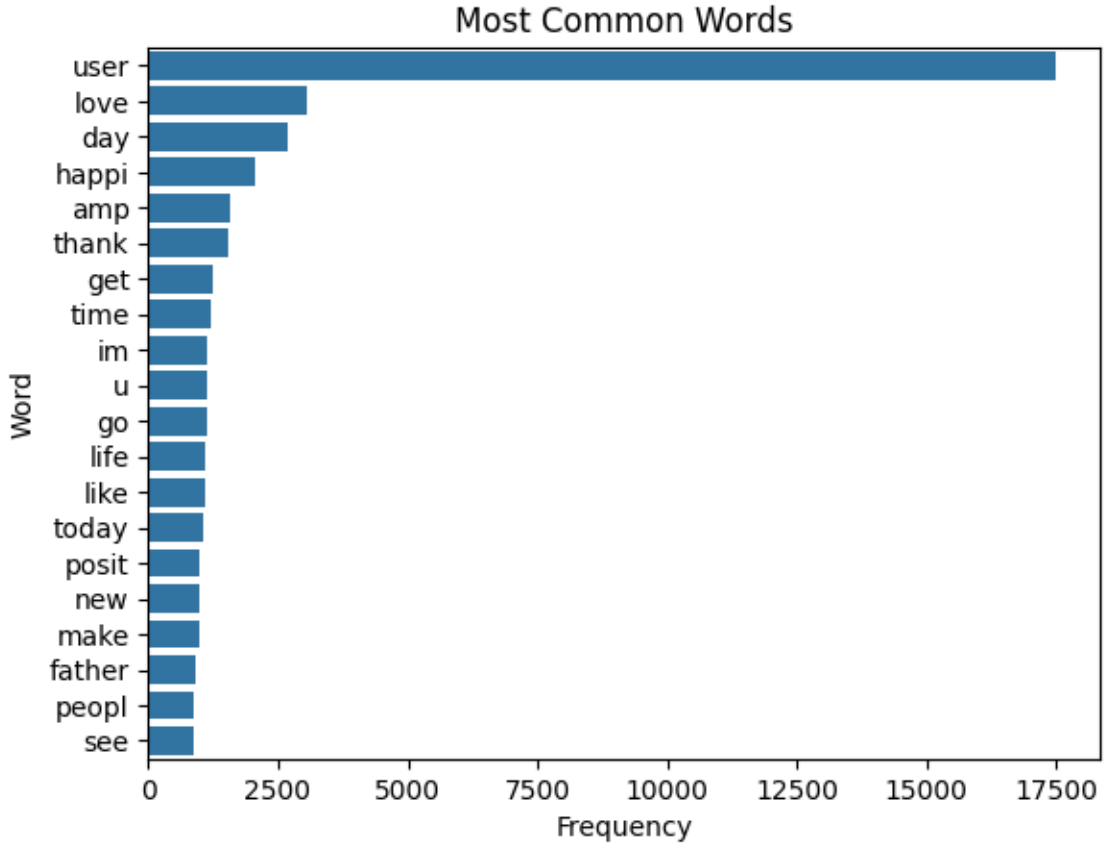
Hate speech can have serious consequences, including perpetuating discrimination, inciting violence, and causing psychological harm. Detecting hate speech on social media platforms like Twitter is crucial for maintaining a safe and inclusive online environment. By identifying and flagging hate speech, we can help prevent the spread of harmful content and protect vulnerable individuals and communities.

EDA


Hate Speech
7.0%
93.0%
Free Speech

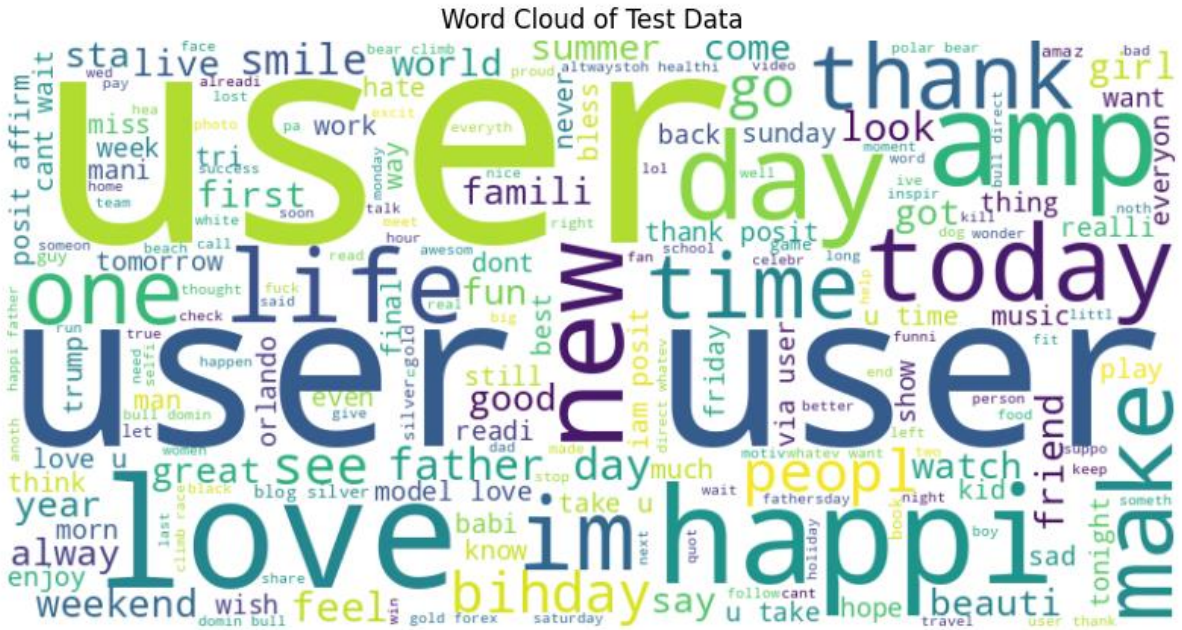

Distribution of Tweet lengths in training Data
Frequency
Tweet Length
0
250
500
750
1000
1250
1500
1750
0
20
40
60
80
100
120


Distribution of Tweet lengths in test Data
Frequency
Tweet Length
0
200
400
600
800
0
20
40
60
80
100
120


Most Common Words
Word
Frequency
user
love
day
happi
amp
thank
get
time
im
u
go
life
like
today
posit
new
make
father
peopl
see
0
2500
5000
7500
10000
12500
15000
17500

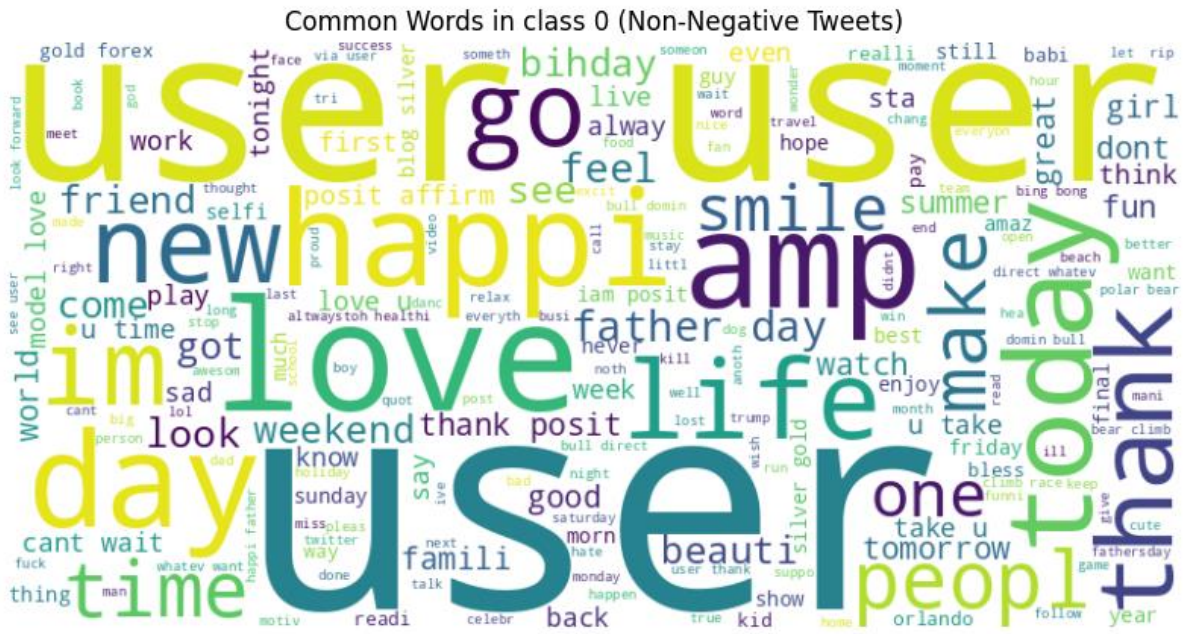
Word Cloud

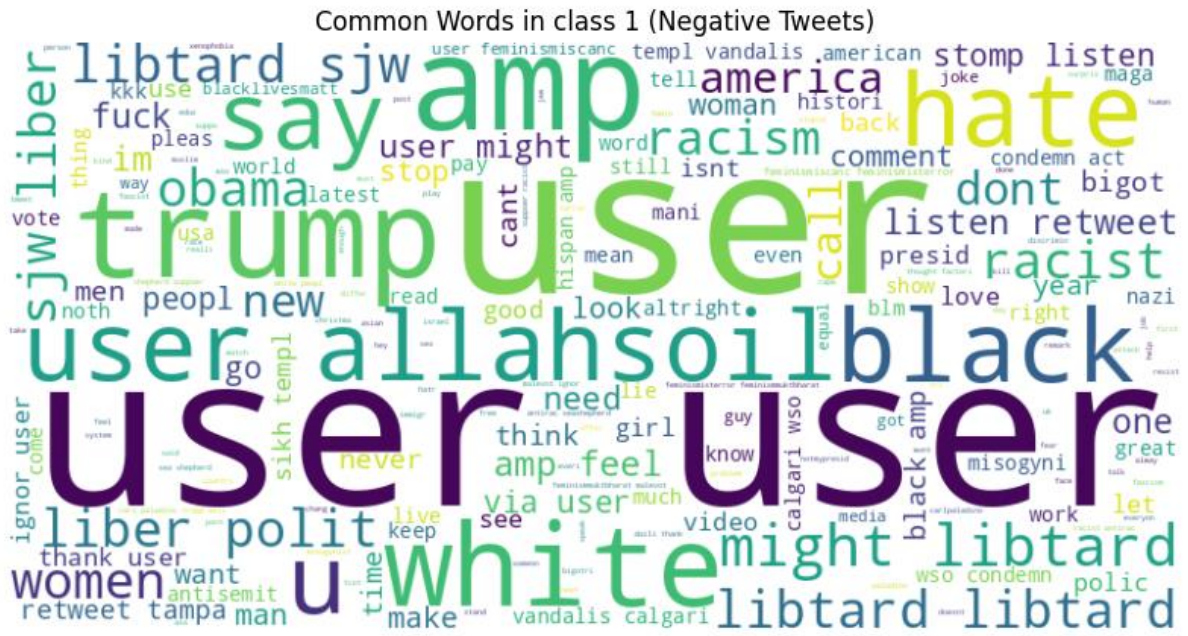

Word Cloud of Training Data
user
day
love
new
amp
thank
life
happi
today
peopl
im
smile
make
time
one
go
look
father
beauti
weekend
bihday
summer
friend
watch
tomorrow
feel
work
world
good
first
know
year
dont
got
sad


Word Cloud of Test Data
user
love
new
amp
thank
day
time
today
life
happi
make
im
one
look
go
peopl
bihday
weekend
smile
summer
come
first
friend
watch
tomorrow
sunday
thing
realli
music
girl
want
everyon
sad
hope
beauti
tonight
play
kid
say
feel
work
world
hate
miss
week
tri
mani
famili
dont
fun
still
good
readi
great
see
father
day
much
model
babi
know
wish
enjoy
alway
morn
year
think
orlando
trump
via user
friday
iam posit
cant wait
posit affirm

Class Distribution

Common Words By Class





Key Findings from EDA

- **Class Distribution :** The class distribution is imbalanced with significantly more non-hate speech tweets compared to hate speech tweets. This needs to be addressed in the model training phase.
- **Common Words :** Common words include generic terms like "user" indicating that further cleaning might be needed to remove such non-informative words.
- **Common Words :** Common words include generic terms like "user" and "thanks," indicating that further cleaning might be needed to remove such non-informative words.

Recommendations

- We will use techniques such as oversampling (e.g., SMOTE) and under sampling to create balance in the dataset.
- We will do additional preprocessing steps to remove non-informative words and symbols to improve the quality of the features.
- We will be using the advanced models such as MLP classifier or Transformer based models such as BERT for this project.