Analyzing Hate Speech Dynamics on Twitter Using Text Mining and NLP Techniques

Team Members

- M K Sowmeya
- Harshita Tanksali
- Gughapriyaa Elango
- Mohammed Huzaif Kherani

Project Overview

- **Addressing Online Harm:** The prevalence of hate speech on social media platforms poses significant challenges to fostering a safe and inclusive online environment. This project aims to contribute to the ongoing efforts to mitigate online harm by developing an effective hate speech detection model.
- **User Well-being:** Creating a platform that can automatically identify and flag hate speech contributes to the protection of users from harmful content, promoting a positive online experience.
- **Algorithmic Fairness:** By leveraging state-of-the-art techniques like CNN, the project strives to enhance the fairness of hate speech classification algorithms, minimizing biases and ensuring equitable treatment across diverse user groups.
- **Community Engagement:** Facilitating community discussions on responsible AI use, ethical considerations, and the importance of collaborative efforts in addressing online toxicity.

Motivation

Algorithmic Amplification:

- Social media algorithms unintentionally boost hate speech by prioritizing sensational or divisive content for engagement.

DeepFakes and Manipulated Content:

- The rise of deepfake technology contributes to the spread of hate speech through convincing fake videos or images.

Evasion Tactics:

- Perpetrators use coded language and memes to evade automated detection, posing challenges for content moderation.

Private Messaging Platforms:

- Hate speech is shifting to private messaging platforms, making it harder to monitor compared to traditional public content moderation.

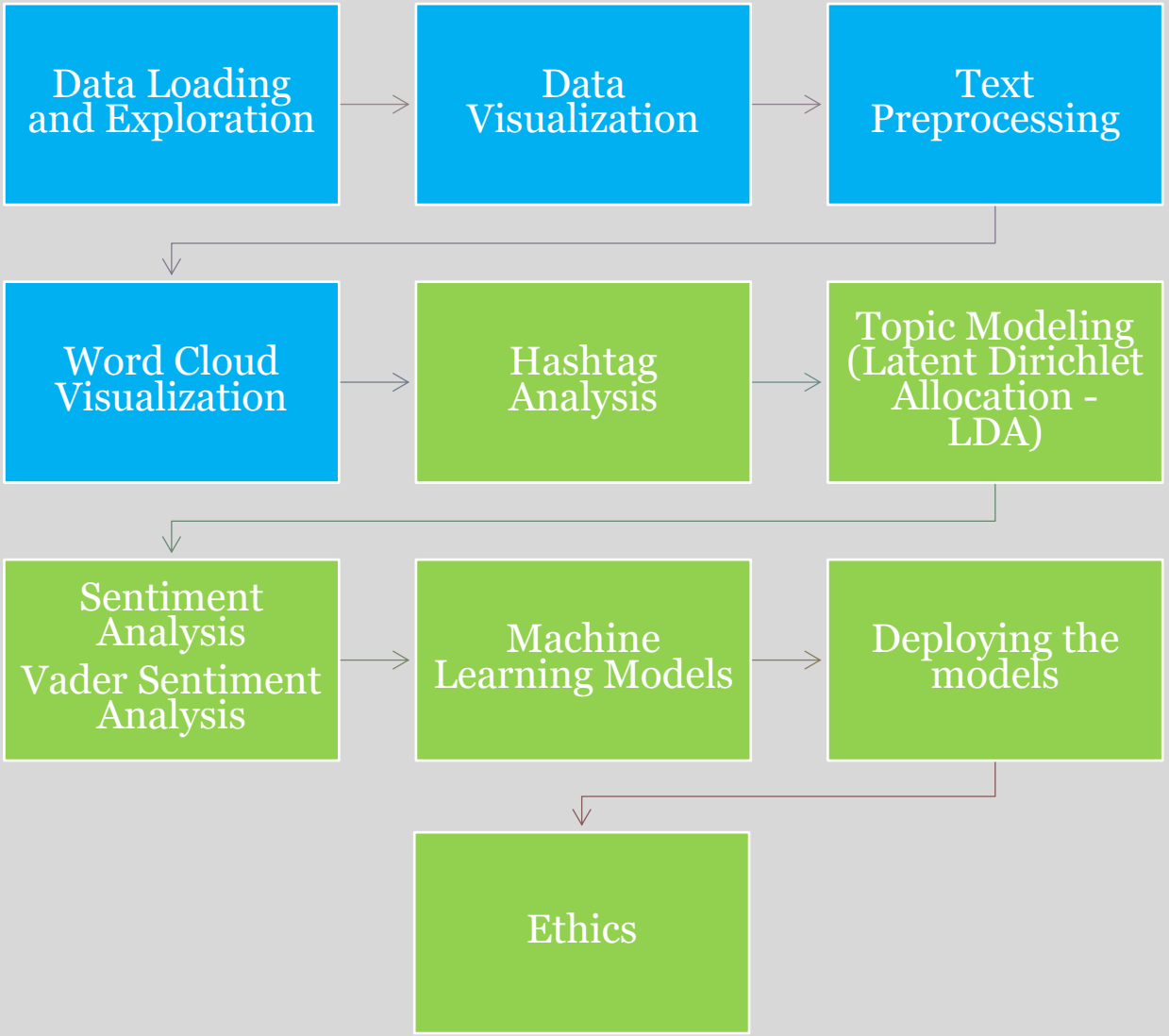
Polarization and Echo Chambers:

- Social media reinforces hate speech within specific communities, intensifying societal polarization.

Globalization of Hate Speech:

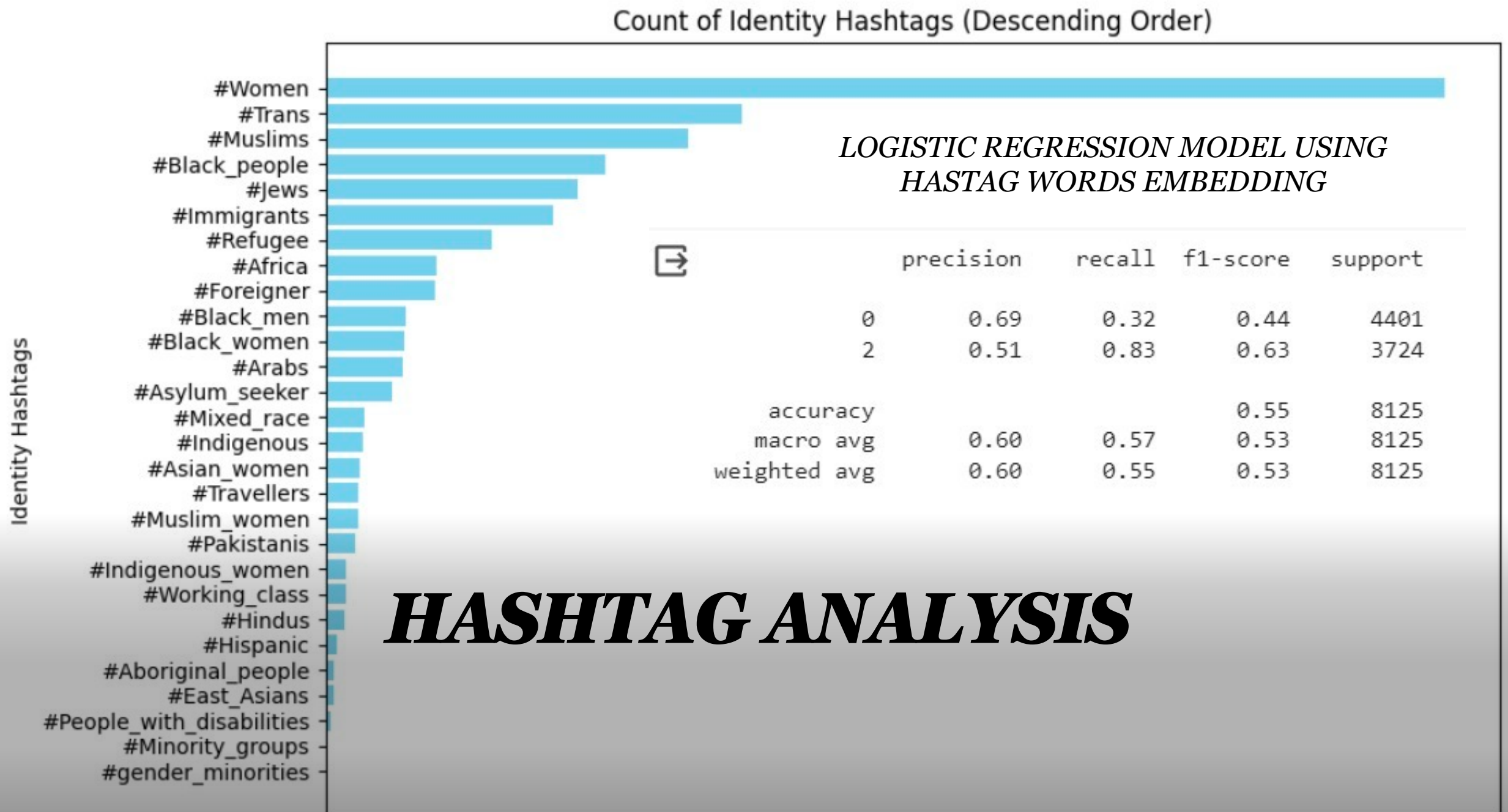
- Social media's borderless nature allows hate speech to transcend geographical boundaries, connecting extremist online communities

Methodology



Challenges Overcome

Traditional models like Mnb and SVM often struggle with capturing the contextual nuances of words.	Traditional models like Mnb and SVM often struggle with capturing the context of words. Fine-tuning and using better embeddings can enhance the models' ability to understand the context of words.
Double negatives pose a challenge for models like Mnb and SVM, as they may not inherently grasp the negation within the sentence.	Using embeddings that capture syntactic and semantic structures can aid in understanding the negation patterns within sentences
Double negatives pose a challenge for models like Mnb and SVM, as they may not inherently grasp the negation within the sentence.	Combination of fine-tuning on domain-specific data and leveraging advanced embeddings, especially contextual embeddings like BERT, can significantly improve the models' performance in handling word ambiguity



LOGISTIC REGRESSION MODEL USING HASTAG WORDS EMBEDDING

	precision	recall	f1-score	support
0	0.69	0.32	0.44	4401
2	0.51	0.83	0.63	3724
accuracy			0.55	8125
macro avg	0.60	0.57	0.53	8125
weighted avg	0.60	0.55	0.53	8125

HASHTAG ANALYSIS

Comparative Analysis of Sentiment Words and Hate Speech Topics

Top words for each topic:

	Topic 1	Topic 2
0	people	black
1	don	fucking
2	love	people
3	just	like
4	women	white
5	think	women
6	want	fuck
7	like	just
8	really	men
9	know	good

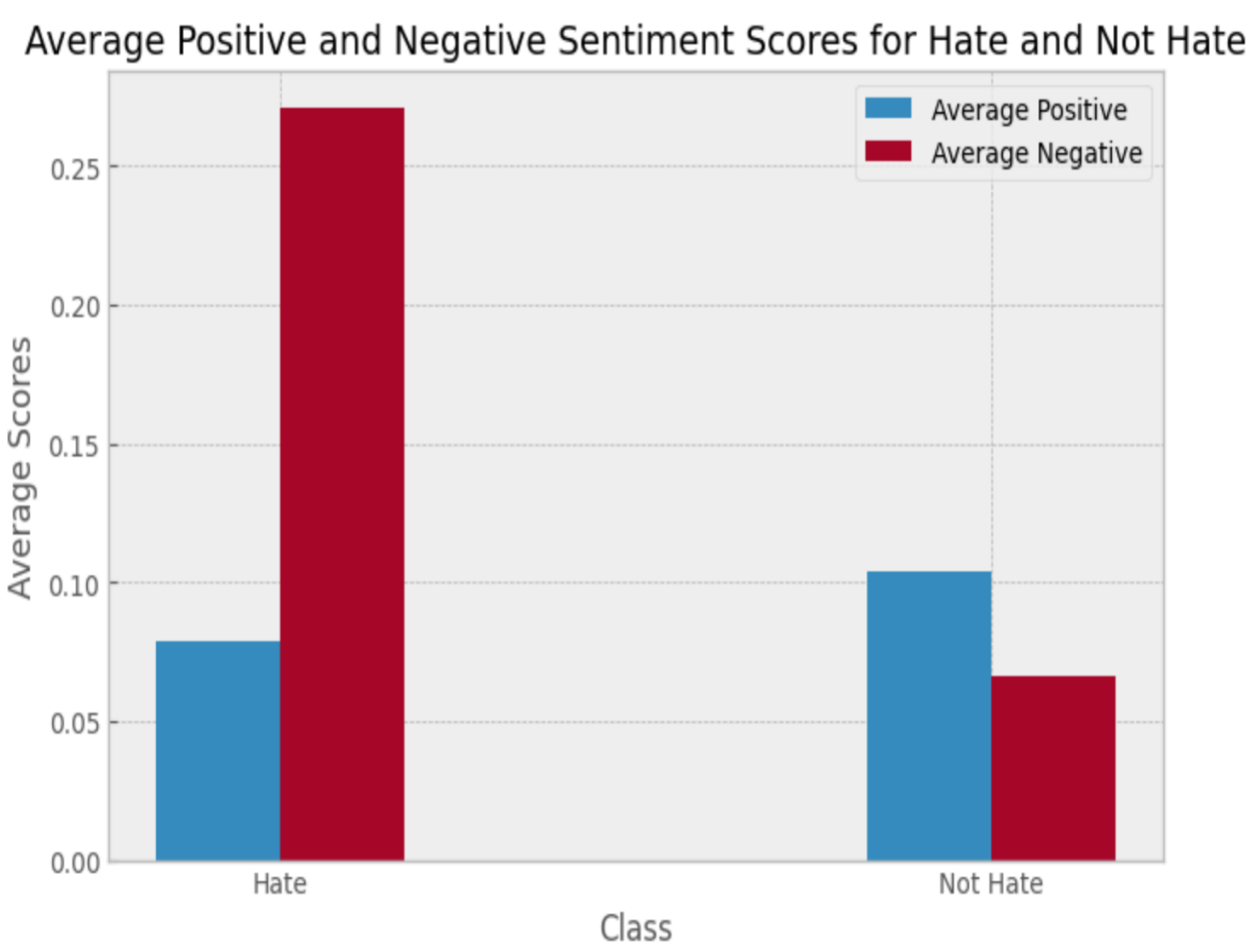
Topic distribution for the new text:
[[0.71193683 0.28806317]]

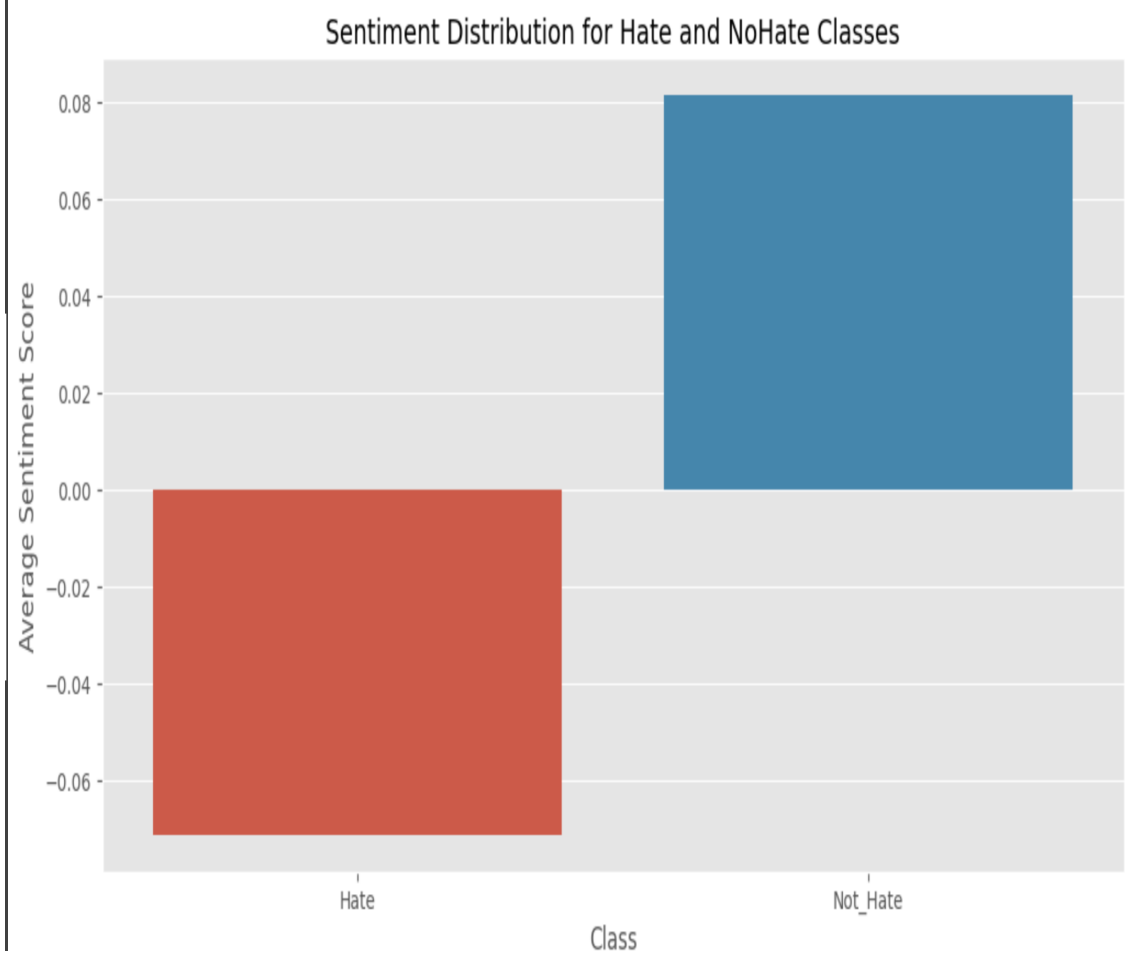
Top 10 Positive Sentiment words in the text

Positive Words	Sentiment Score
ily	0.6597
sweetheart	0.6486
happiest	0.6369
paradise	0.6369
lovingly	0.6369
elated	0.6369
best	0.6369
love	0.6369
euphoric	0.6369
best.	0.6369

Top 10 Negative Sentiment words in the text

Negative Words	Sentiment Score
rapist	-0.7096
raping	-0.7003
slavery	-0.7003
murder	-0.6908
kill	-0.6908
fu	-0.6908
terrorist	-0.6908
rape	-0.6908
terrorism	-0.6808
murderer	-0.6808





VADER SENTIMENT ANALYSIS

Vader Sentiment Analysis Evaluation using Random Forest

Evaluation for tweet2 (Random Forest):

	precision	recall	f1-score	support
0	0.60	0.68	0.64	6636
2	0.55	0.46	0.50	5551
accuracy			0.58	12187
macro avg	0.57	0.57	0.57	12187
weighted avg	0.58	0.58	0.58	12187

Accuracy: 0.5808648559940921

Sensitivity for Class 0 (Hate): 0.6796262808921036
Specificity for Class 0 (Hate): 0.6796262808921036

Sensitivity for Class 2 (Non-hate): 0.46279949558638084
Specificity for Class 2 (Non-hate): 0.46279949558638084

Test Output

0 is hate and 2 is not hate

Tweet: I love spreading kindness and positivity! #SpreadLove | **Predicted Class: 2**

Tweet: This is the worst thing ever! I can't believe people are so hateful. #Angry | **Predicted Class: 2**

Tweet: Why are muslims always fighting with others? Dont they have anything better to do? | **Predicted Class: 0**

Tweet: I can't stand women who act like they know it all. | **Predicted Class: 0**

Tweet: This is unacceptable! How can anyone say such hurtful things? #Angry | **Predicted Class: 2**

Tweet: Gay teachers should not be allowed near children | **Predicted Class: 0**

Tweet: Why are immigrants stealing all our jobs and money | **Predicted Class: 2**

Tweet: It's important to address hate speech and work towards building a more inclusive society. #SocialJustice | **Predicted Class: 0**

Tuned Model Evaluation:

	precision	recall	f1-score	support
hate	0.77	0.71	0.74	6636
nothate	0.68	0.75	0.71	5551
accuracy			0.73	12187
macro avg	0.73	0.73	0.73	12187
weighted avg	0.73	0.73	0.73	12187

Tuned Model Accuracy: 0.726593911545089

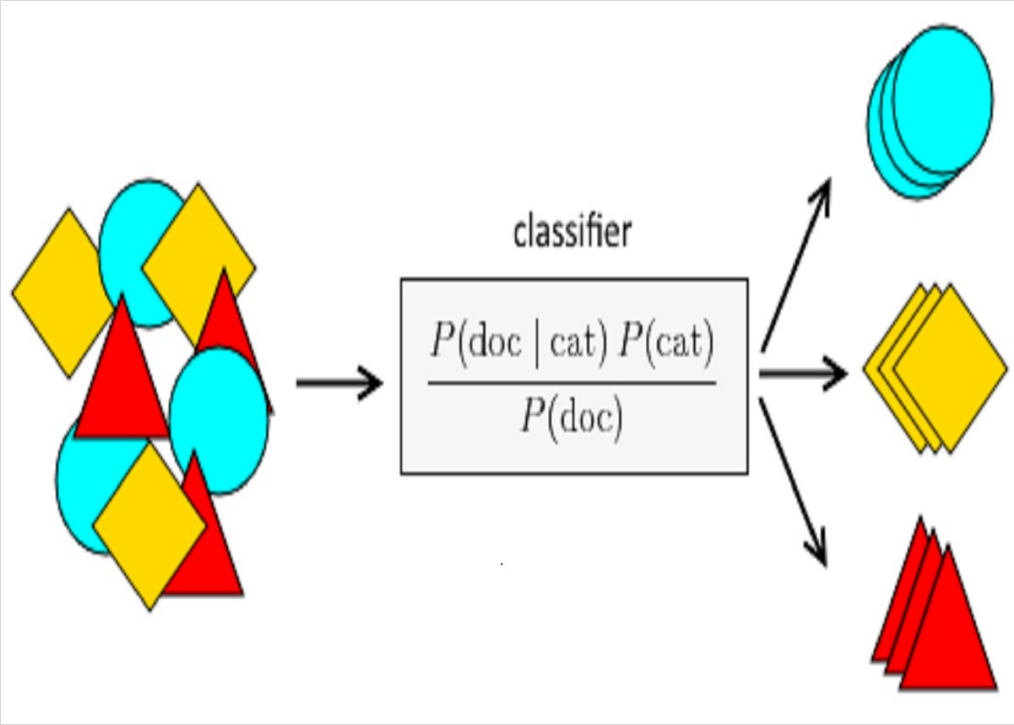
Evaluation for tweet2:

	precision	recall	f1-score	support
hate	0.69	0.72	0.70	6636
nothate	0.65	0.61	0.63	5551
accuracy			0.67	12187
macro avg	0.67	0.66	0.67	12187
weighted avg	0.67	0.67	0.67	12187

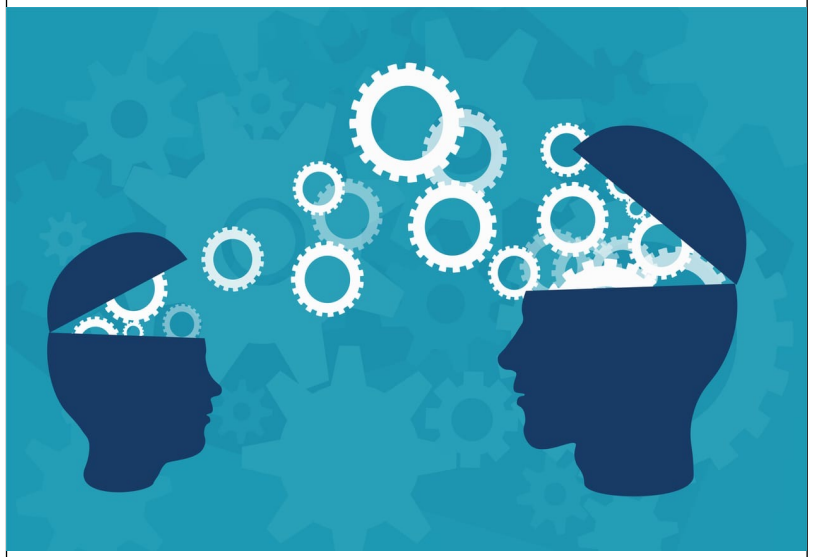
Accuracy: 0.6695659309099861
Confusion Matrix for data:
[[4786 1850]
 [2177 3374]]

Sensitivity for data: 0.6078184110970997
Specificity for data: 0.7212176009644364

Multinomial NB Model



CROSS DOMAIN EVALUATION



Evaluation for tweet1 using the model trained on tweet2:

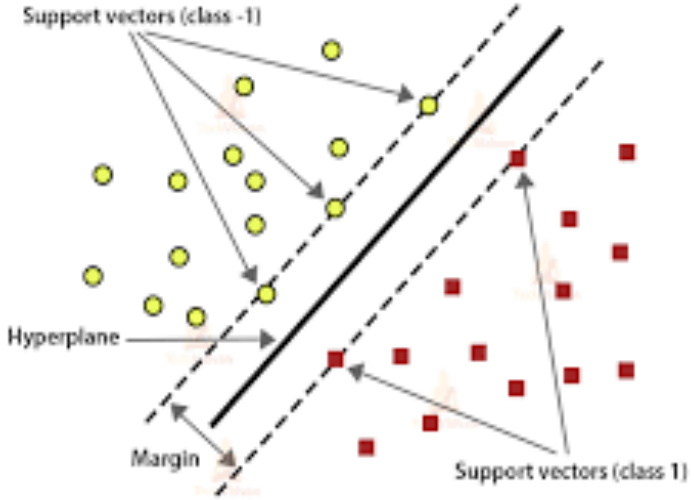
	precision	recall	f1-score	support
0	0.07	0.68	0.13	1430
1	0.00	0.00	0.00	19190
2	0.21	0.55	0.30	4163
accuracy			0.13	24783
macro avg	0.09	0.41	0.14	24783
weighted avg	0.04	0.13	0.06	24783

Accuracy: 0.13134003147318726

Confusion Matrix for tweet2 train and tweet1 test:

```
[[  968     0   462]
 [10902     0  8288]
 [ 1876     0  2287]]
```

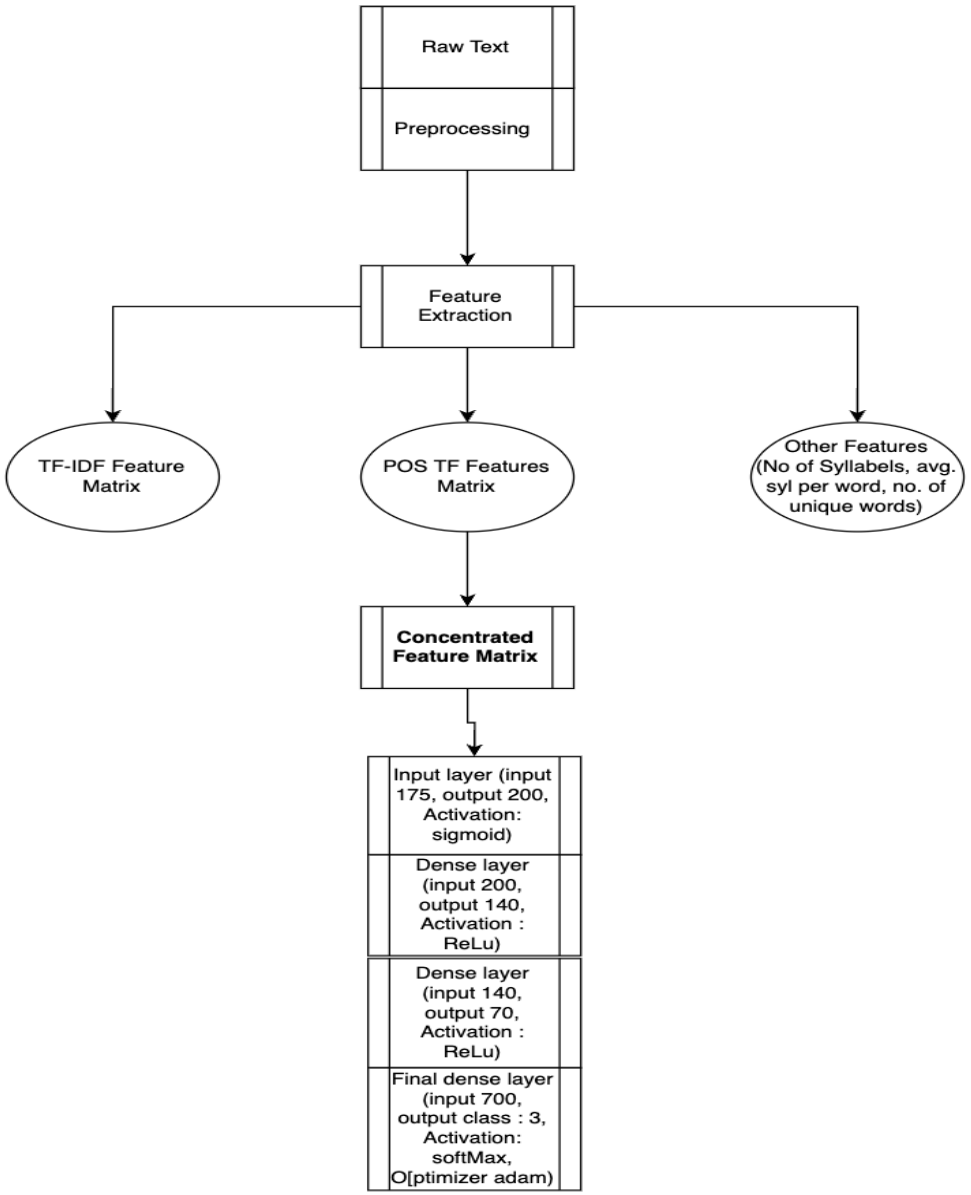
SVM




```
Evaluation for tweet2 (SVM):
              precision    recall  f1-score   support

           0       0.76      0.76      0.76      6636
           2       0.71      0.72      0.71      5551

    accuracy                           0.74     12187
   macro avg       0.73      0.74      0.73     12187
weighted avg       0.74      0.74      0.74     12187

Accuracy: 0.7369327972429638


Confusion Matrix for tweet2 (SVM):
[[5011 1625]
 [1581 3970]]

Sensitivity for tweet2 (SVM): 0.715186452891371
Specificity for tweet2 (SVM): 0.7551235684147076
```

Multi-Layer Perceptron

MLP Based Proposed Model



MLP Model Summary:
Input Layer: Size = 1739 Activation = Sigmoid
Hidden Layer 1: Size = 200, Activation = ReLU
Hidden Layer 2: Size = 140, Activation = ReLU
Hidden Layer 3: Size = 70, Activation = ReLU
Output Layer: Size = 2 Activation = Softmax

```
Accuracy: 0.7504

Classification Report:
              precision    recall  f1-score   support

        hate       0.77      0.76      0.77      4401
     nothate       0.72      0.74      0.73      3724

    accuracy                           0.75      8125
   macro avg       0.75      0.75      0.75      8125
weighted avg       0.75      0.75      0.75      8125


Sensitivity (True Positive Rate): 0.7376476906552094
Specificity: 0.761190638491252
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 200)               100400

 dense_1 (Dense)             (None, 140)               28140

 dense_2 (Dense)             (None, 70)                9870

 dense_3 (Dense)             (None, 3)                 213

=================================================================
Total params: 138623 (541.50 KB)
Trainable params: 138623 (541.50 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Tweet: I love spreading kindness and positivity! #SpreadLove
Prediction: nothate

Tweet: This is the worst thing ever! I can't believe people are so hateful. #Angry
Prediction: nothate

Tweet: Why are muslims always fighting with others? Dont they have anything better to do?
Prediction: hate

Tweet: I can't stand women who act like they know it all.
Prediction: hate

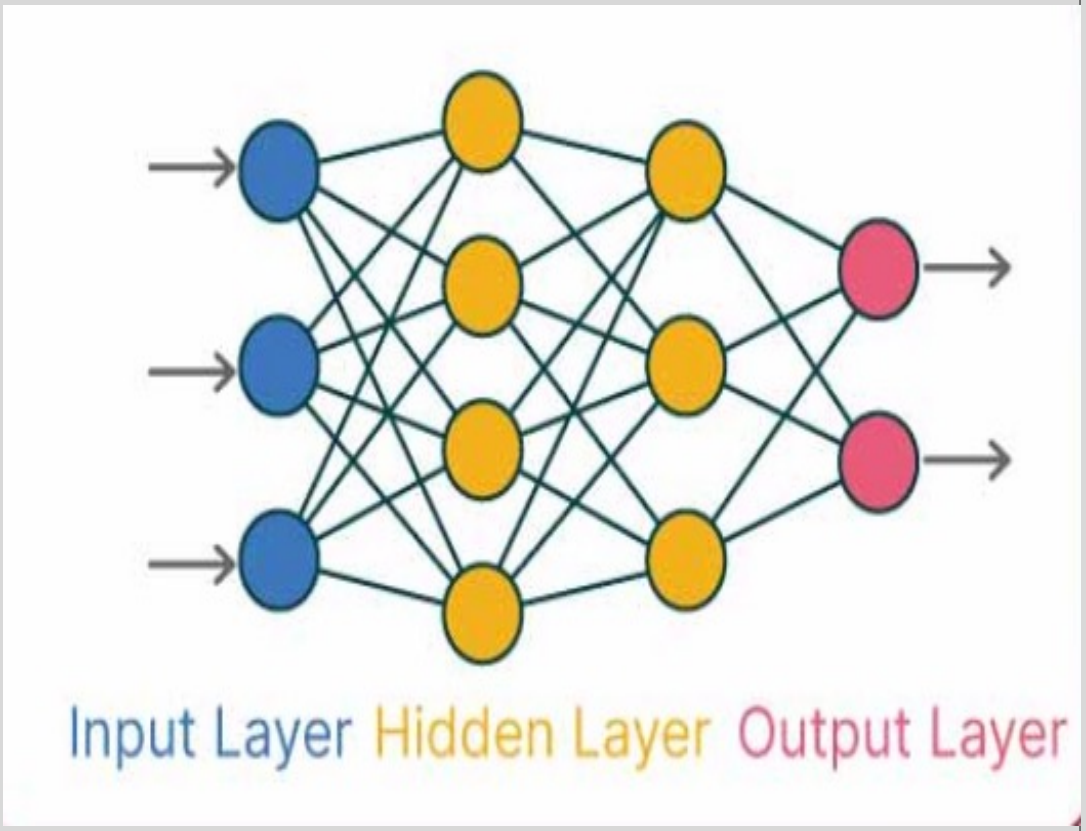
Tweet: This is unacceptable! How can anyone say such hurtful things? #Angry
Prediction: nothate

Tweet: Gay teachers should not be allowed near children
Prediction: hate

Tweet: Why are immigrants stealing all our jobs and money
Prediction: hate

Tweet: It's important to address hate speech and work towards building a more inclusive society. #SocialJustice
Prediction: nothate

Multi-Layer Perceptron



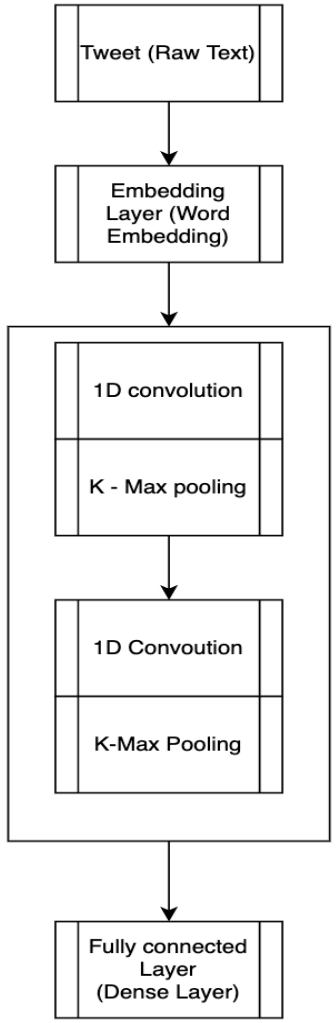
DCNN + GLOVE

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 100, 100)          1000000

 conv1d (Conv1D)             (None, 100, 300)          30300

 conv1d_1 (Conv1D)           (None, 99, 300)           180300

 conv1d_2 (Conv1D)           (None, 97, 300)           270300

 conv1d_3 (Conv1D)           (None, 94, 300)           360300

 global_max_pooling1d (Glob  (None, 300)               0
 alMaxPooling1D)

 dropout (Dropout)           (None, 300)               0

 dense (Dense)               (None, 64)                19264

 dense_1 (Dense)             (None, 1)                 65

=================================================================
Total params: 1860529 (7.10 MB)
Trainable params: 860529 (3.28 MB)
Non-trainable params: 1000000 (3.81 MB)
_________________________________________________________________
```

	precision	recall	f1-score	support
0	0.81	0.87	0.84	22124
1	0.83	0.75	0.79	18499
accuracy			0.82	40623
macro avg	0.82	0.81	0.81	40623
weighted avg	0.82	0.82	0.82	40623

```
Accuracy: 0.8167540555842749
Confusion Matrix:
[[19218  2906]
 [ 4538 13961]]
Sensitivity: 0.7546894426725769
Specificity: 0.8686494304827337
```

DCNN BASED PORPOSED MODEL



DCNN + Glove

```
1/1 [==============================] - 0s 81ms/step
Tweet: "I love spreading kindness and positivity! #SpreadLove" - Sentiment: Positive
Tweet: "This is the worst thing ever! I can't believe people are so hateful. #Angry" - Sentiment: Positive
Tweet: "Why are muslims always fighting with others? Dont they have anything better to do?" - Sentiment: Negative
Tweet: "I can't stand women who act like they know it all." - Sentiment: Negative
Tweet: "This is unacceptable! How can anyone say such hurtful things? #Angry" - Sentiment: Positive
Tweet: "Gay teachers should not be allowed near children" - Sentiment: Negative
Tweet: "Why are immigrants stealing all our jobs and money" - Sentiment: Negative
Tweet: "It's important to address hate speech and work towards building a more inclusive society. #SocialJustice" - Sentiment: Positive
```

BERT

```
Epoch:  1
2468/10000.0 loss: 0.8399315525152136
0.0M
```

```
(torch.Size([3, 512]), torch.Size([3, 512, 768]), torch.Size([3, 768]))
```

```
bert_clf = BertBinaryClassifier()
bert_clf = bert_clf.cpu()
```

```
100%|██████████| 407873900/407873900 [00:12<00:00, 33663490.20B/s]
```

```
optimizer = Adam(bert_clf.parameters(), lr=3e-6)
```

Ethical considerations

Freedom of speech vs Hate speech

Striking the right balance is crucial to avoid suppressing legitimate expression.

Bias & Fairness

Ensuring fairness and mitigating biases is essential to avoid discrimination and unintended consequences.

Robustness & Security

The model should not be deceived easily by users or administrators.

Summary

Robust Model Construction:

- Prioritize the construction of models with robustness to mitigate bias.
- Address concerns related to unintentional amplification of stereotypes or discriminatory behavior by users.

Real-time Monitoring Integration:

- Design the model to seamlessly integrate into real-time monitoring systems.
- Enable swift detection and response to potential instances of hate speech as they occur.


THANK YOU