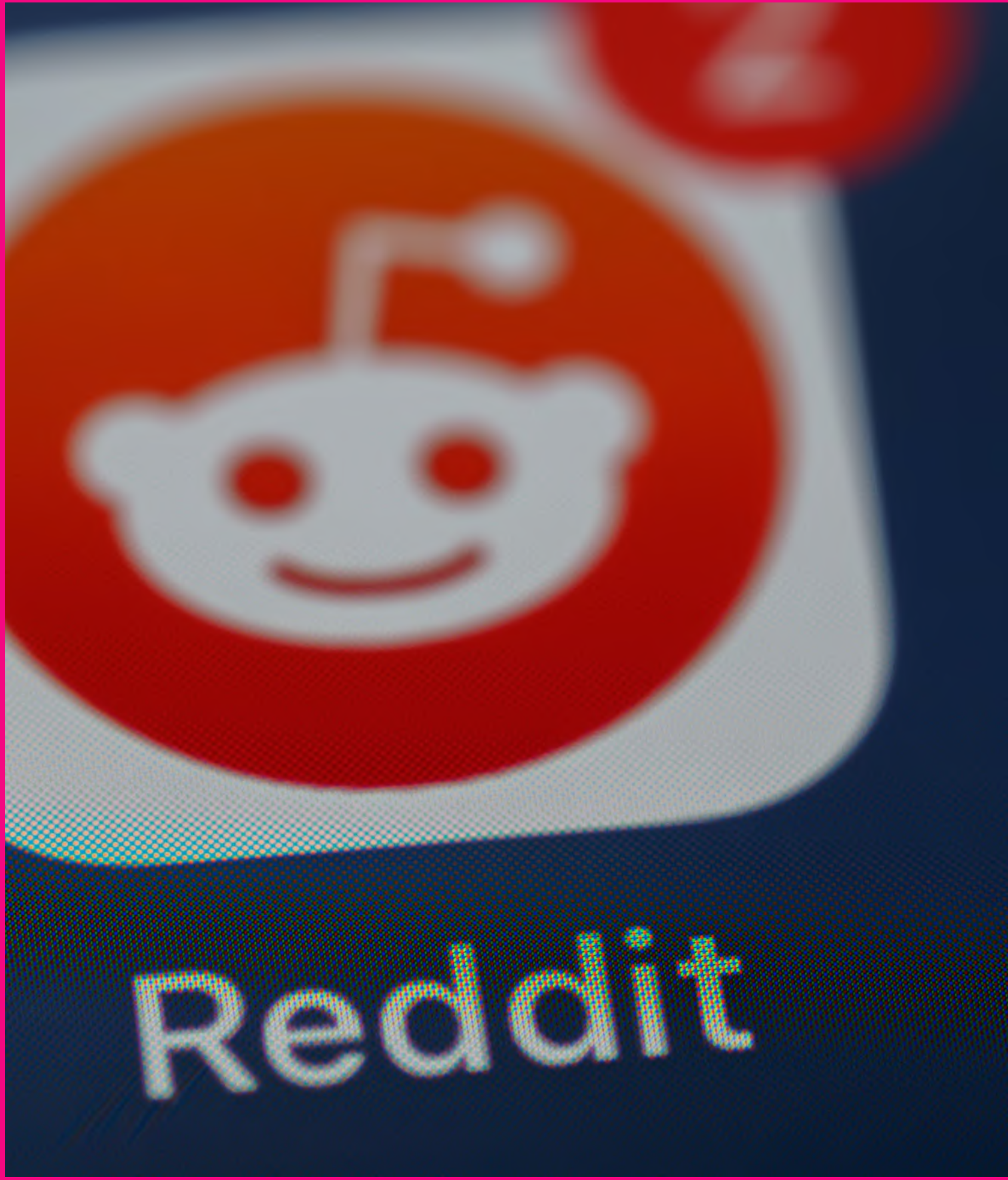
"FEM GUARD :

A Supervised Learning Approach To Segregate Threats Against Women On Reddit With multilingual, multi-label and Automated bot Intervention"

Project by:
Palak Kochey , Kanika Thombre , Pratyush Agarwal

1. Introduction

1. FEM GUARD : A Supervised learning Approach To Segregate Threats Against Women On Reddit With multi-lable, multilingual and Automated Bot Intervention, the main aim to ensures that harmful content is promptly identified and appropriate actions are taken to mitigate potential harm.
2. FEM GUARD allows automated bot intervention. Overall with FEM GUARD the dream of a healthy digital environment can be achieved by leveraging technology to enhance the safety and well-being of women.



2. Problem Statement

A dream of a healthy digital environment can be achieved by leveraging technology to enhance the safety and well-being of women.

BACKGROUND

i. Online harassment and hate speech, particularly directed at women, pervade social media platforms like Reddit, often manifesting in forms such as sexist comments, threats of violence, and sexual harassment.

ii. This widespread issue not only violates individuals' rights and dignity but also perpetuates a toxic online culture that normalizes misogyny and undermines gender equality efforts.

MOTIVATION

i. The motivations behind this project stem from the pressing need to address the pervasive issue of online harassment and hate speech directed towards women on platforms like Reddit.

ii. Online harassment not only violates individuals' rights and dignity but also perpetuates a hostile online environment that undermines inclusivity and contributes to gender inequality.

3. Data Set Description

The Fem Guard Model uses a real world dataset which consists of comments, from diverse subreddits. The data extraction process involved creation of reddit API using standard tools provided by the reddit developer section. This facilitated allocation of a unique script ID and client code, important for user authentication. These credentials were then integrated into the PRAW library in python, enabling the systematic scraping of data from reddit.

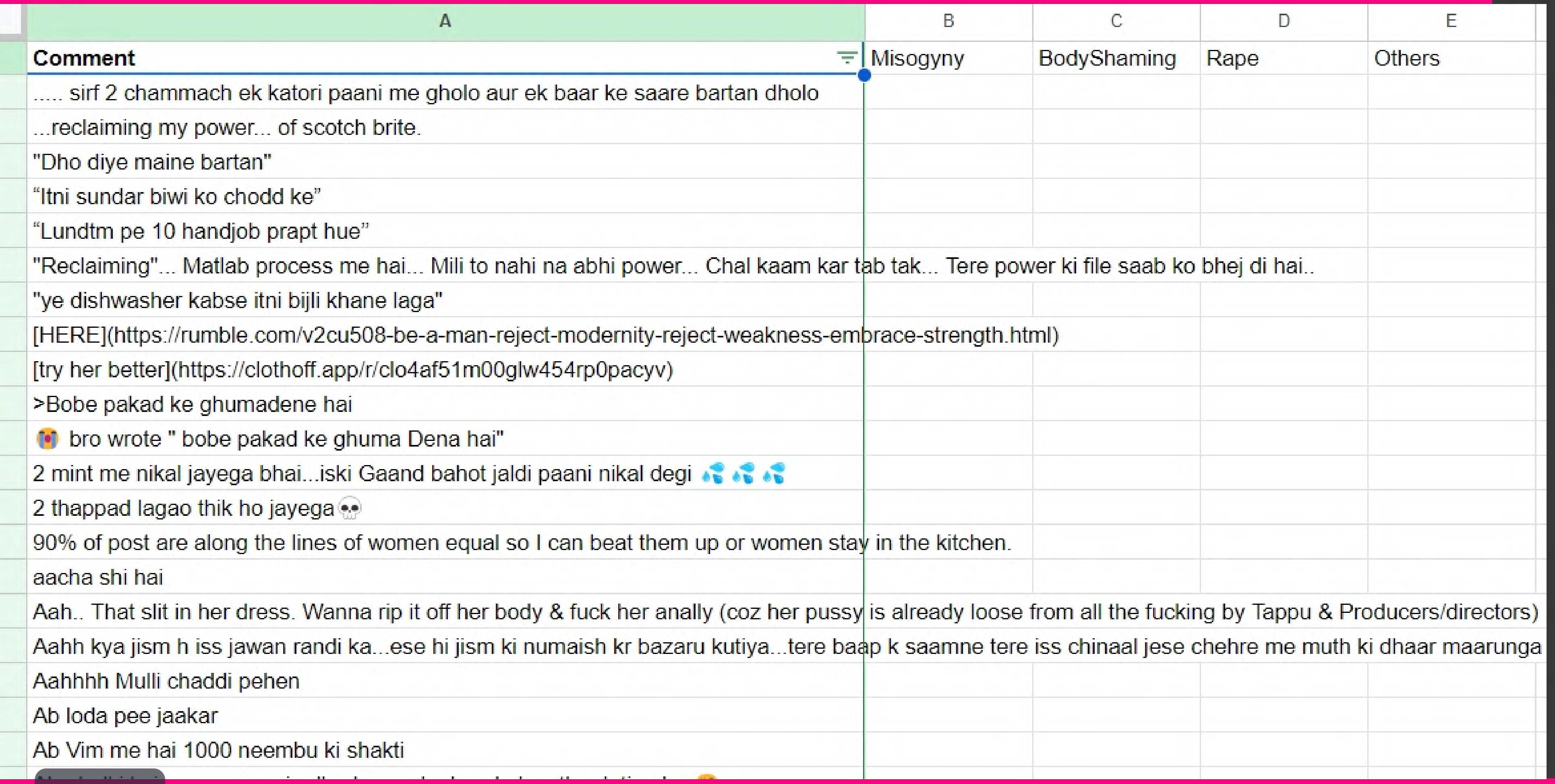
The dataset contains comments against women categorized into 4 broad tones :

i) Body shaming Comments

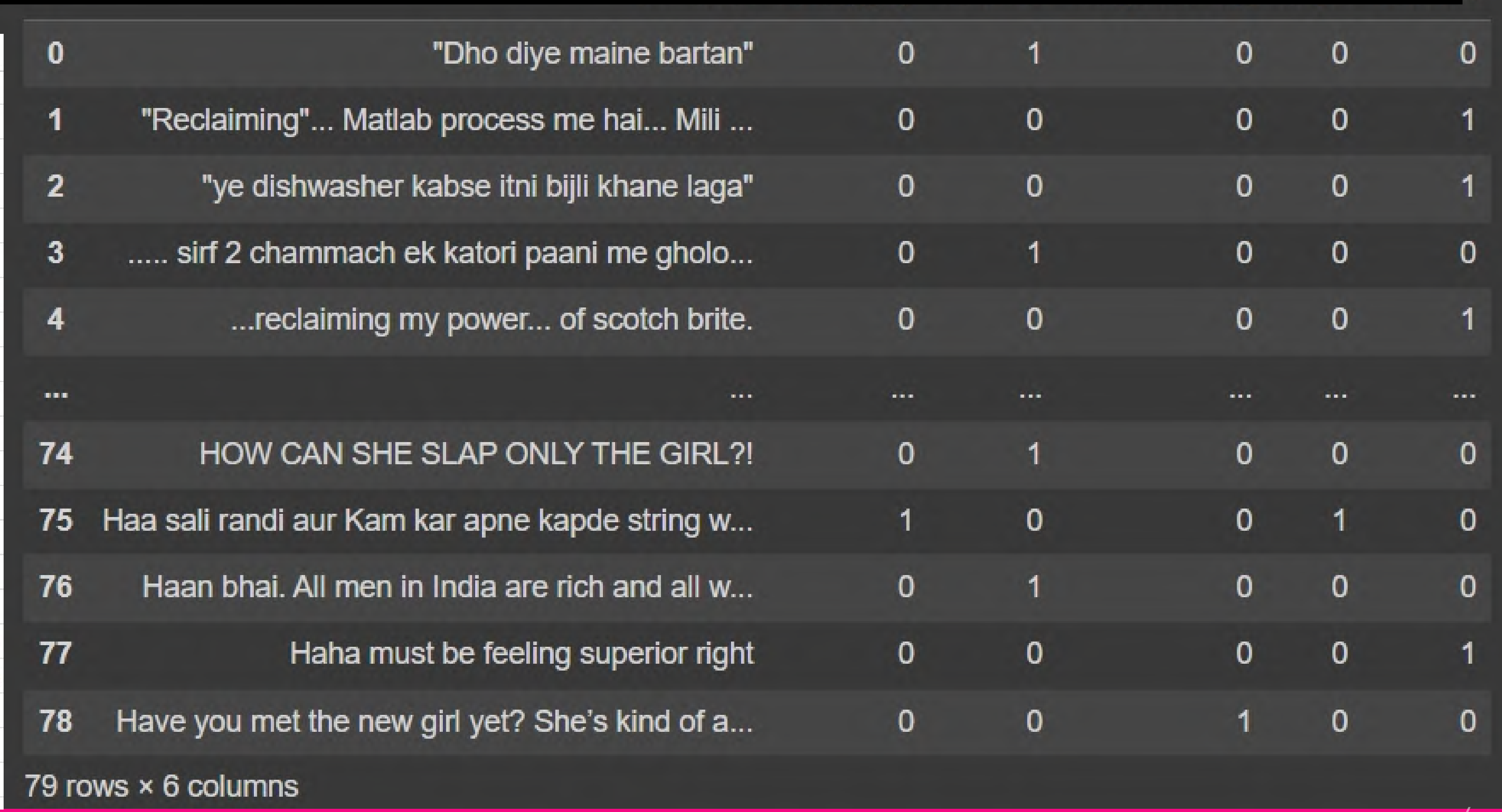
ii) Misogynistic Comments

iii) Comments Containing Rape Threats

iv) Other Uncategorizable Derogatory Comments



	A	B	C	D	E
	Comment	Misogyny	BodyShaming	Rape	Others
	..... sirf 2 chammach ek katori paani me gholo aur ek baar ke saare bartan dholo				
	...reclaiming my power... of scotch brite.				
	"Dho diye maine bartan"				
	“Itni sundar biwi ko chodd ke”				
	“Lundtm pe 10 handjob prapt hue”				
	"Reclaiming"... Matlab process me hai... Mili to nahi na abhi power... Chal kaam kar tab tak... Tere power ki file saab ko bhej di hai..				
	"ye dishwasher kabse itni bijli khane laga"				
	[HERE](https://rumble.com/v2cu508-be-a-man-reject-modernity-reject-weakness-embrace-strength.html)				
	[try her better](https://clothoff.app/r/clo4af51m00glw454rp0pacyv)				
	>Bobe pakad ke ghumadene hai				
	😭 bro wrote " bobe pakad ke ghuma Dena hai"				
	2 mint me nikal jayega bhai...iski Gaand bahot jaldi paani nikal degi 💦💦💦				
	2 thappad lagao thik ho jayega💀				
	90% of post are along the lines of women equal so I can beat them up or women stay in the kitchen.				
	aacha shi hai				
	Aah.. That slit in her dress. Wanna rip it off her body & fuck her anally (coz her pussy is already loose from all the fucking by Tappu & Producers/directors)				
	Aahh kya jism h iss jawan randi ka...ese hi jism ki numaish kr bazaru kutiya...tere baap k saamne tere iss chinaal jese chehre me muth ki dhaar maarunga				
	Aahhhh Mulli chaddi pehen				
	Ab loda pee jaakar				
	Ab Vim me hai 1000 neembu ki shakti				

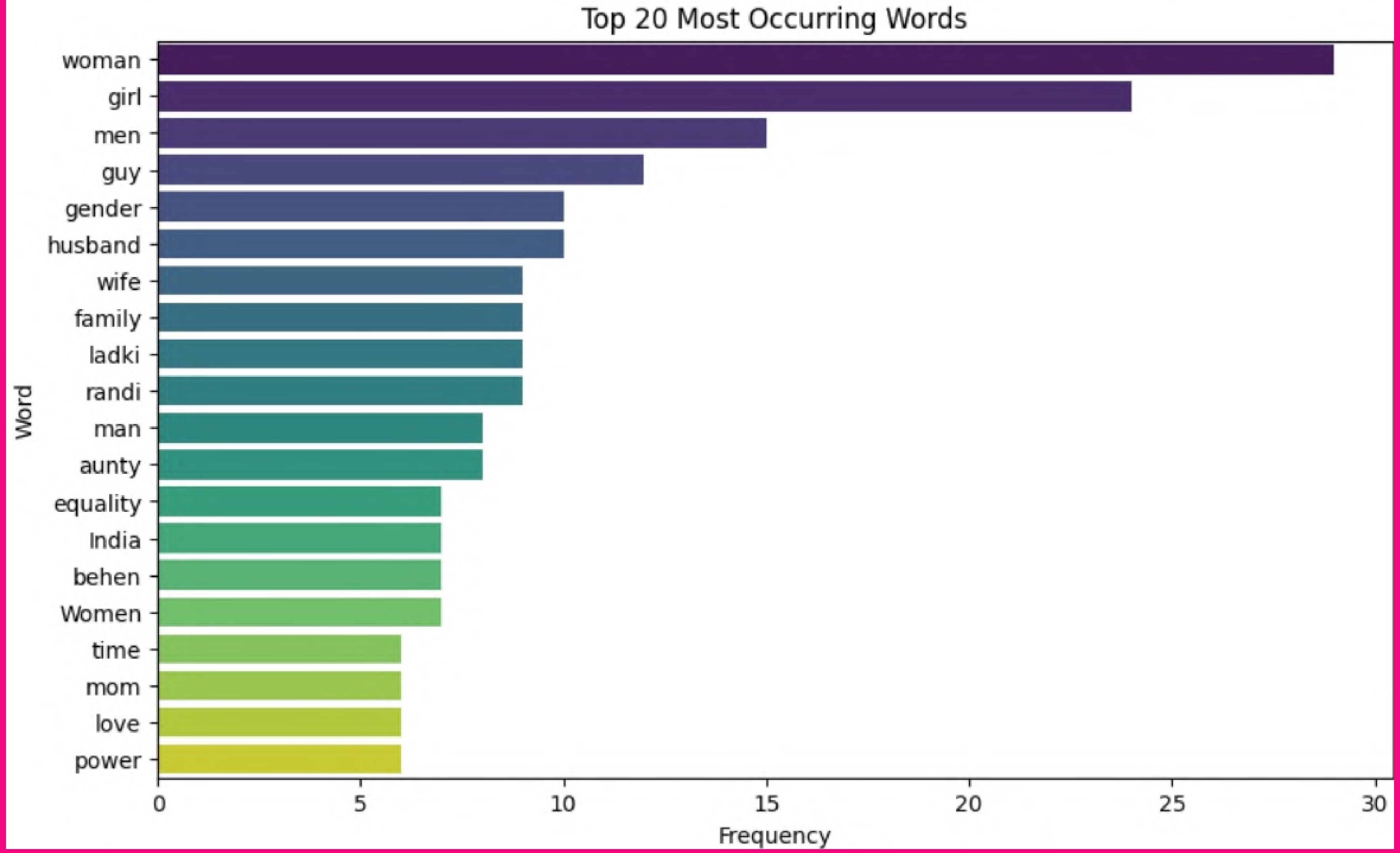


						s
0	"Dho diye maine bartan"	0	1	0	0	0
1	"Reclaiming"... Matlab process me hai... Mili ...	0	0	0	0	1
2	"ye dishwasher kabse itni bijli khane laga"	0	0	0	0	1
3	..... sirf 2 chammach ek katori paani me gholo...	0	1	0	0	0
4	...reclaiming my power... of scotch brite.	0	0	0	0	1
...	...	...	...	...	...	...
74	HOW CAN SHE SLAP ONLY THE GIRL?!	0	1	0	0	0
75	Haa sali randi aur Kam kar apne kapde string w...	1	0	0	1	0
76	Haan bhai. All men in India are rich and all w...	0	1	0	0	0
77	Haha must be feeling superior right	0	0	0	0	1
78	Have you met the new girl yet? She’s kind of a...	0	0	1	0	0

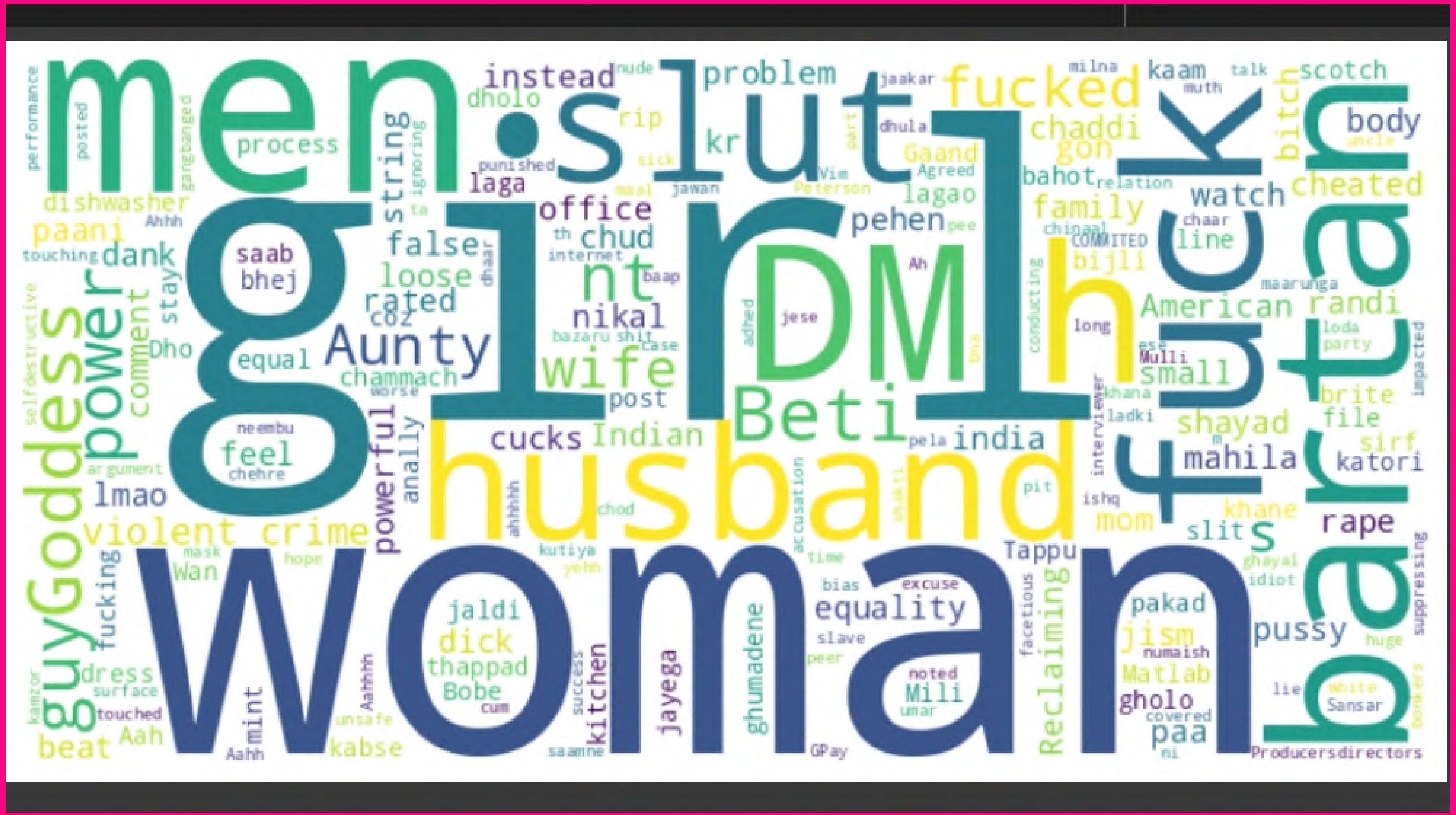
79 rows × 6 columns

3.1 DATASET SNIPPED

4. Dataset visual

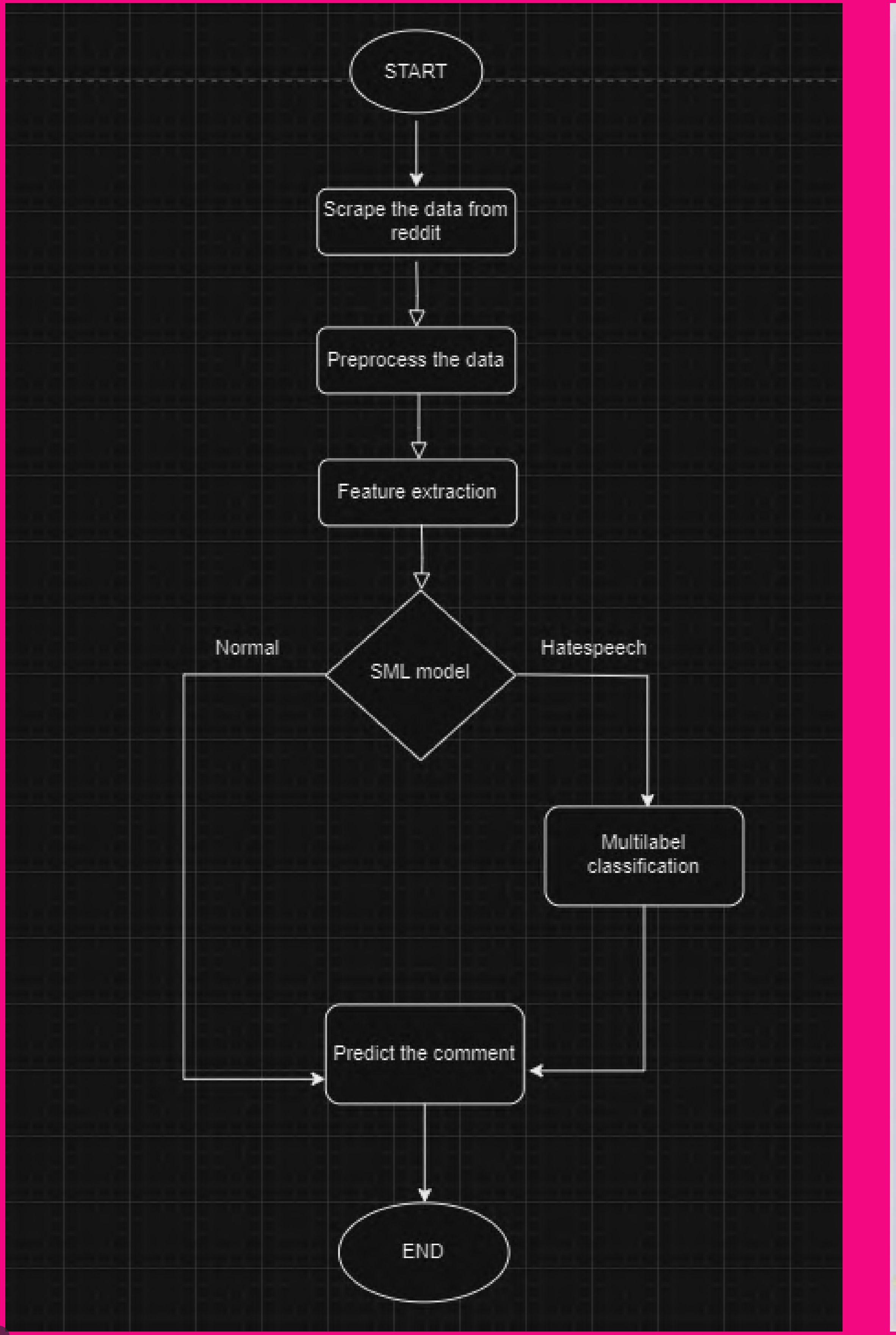


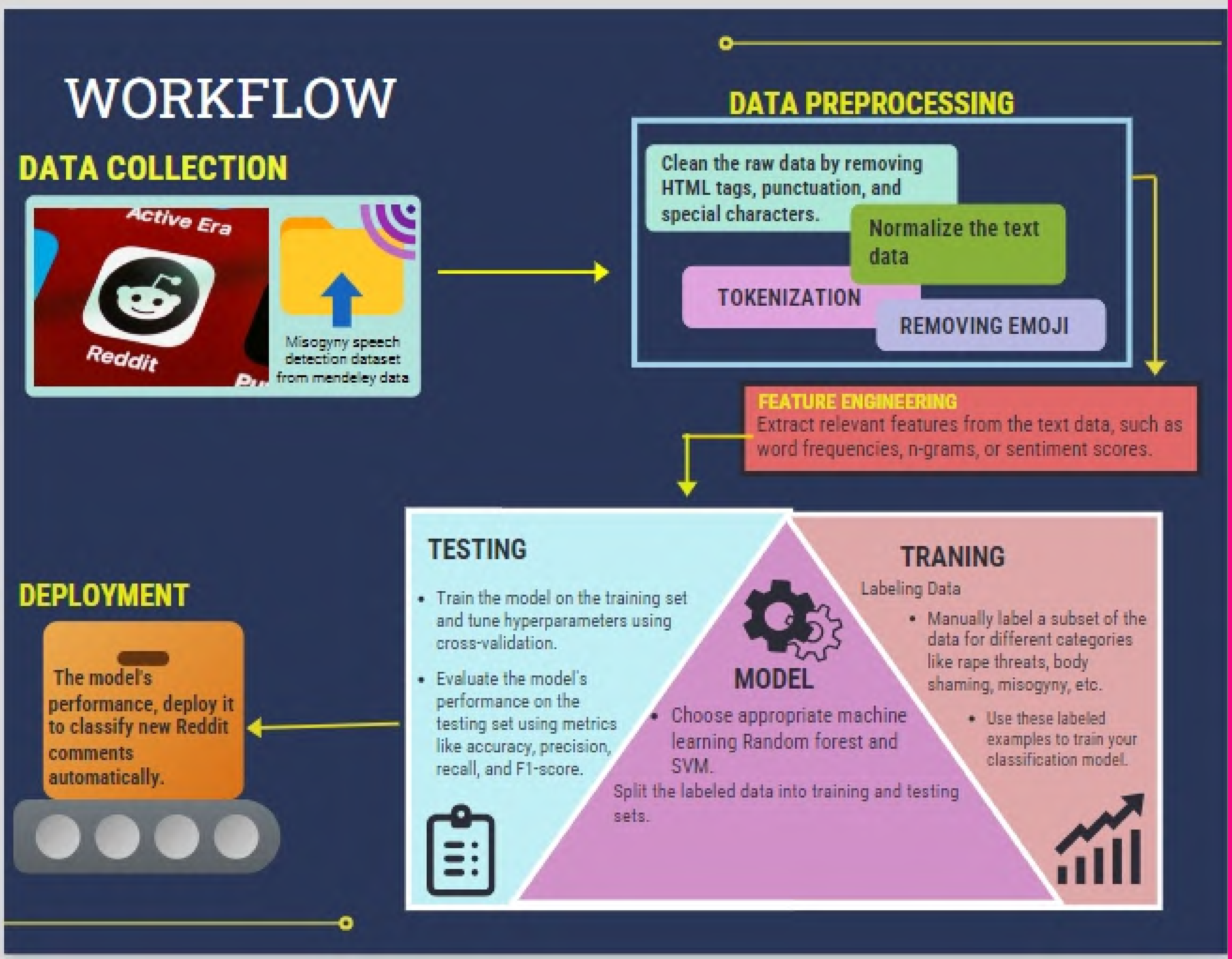
4.1. Representation of most occurring words



4.2. Representation in word cloud format

5. Methodology





WORKFLOW

DATA COLLECTION

Misogyny speech detection dataset from mendeley data

DATA PREPROCESSING

Clean the raw data by removing HTML tags, punctuation, and special characters.

Normalize the text data

TOKENIZATION

REMOVING EMOJI

FEATURE ENGINEERING Extract relevant features from the text data, such as word frequencies, n-grams, or sentiment scores.

TESTING

- Train the model on the training set and tune hyperparameters using cross-validation.
- Evaluate the model's performance on the testing set using metrics like accuracy, precision, recall, and F1-score.

MODEL

- Choose appropriate machine learning Random forest and SVM.

Split the labeled data into training and testing sets.

TRANING

Labeling Data

- Manually label a subset of the data for different categories like rape threats, body shaming, misogyny, etc.
- Use these labeled examples to train your classification model.

DEPLOYMENT

The model's performance, deploy it to classify new Reddit comments automatically.

Try Pitch

5.1 THE FLOWCHART OF THE METHODOLOGY

6. MODEL APPLICATION

1. Preprocessing: The dataset underwent preprocessing to clean and standardize the text data, including removing stopwords, punctuation, and special characters.

2. Feature Extraction: Utilized BERT (Bidirectional Encoder Representations from Transformers) for feature extraction, capturing contextual information from the text data.

3. Model Selection: Employed Random Forest, a robust ensemble learning technique, for classification due to its ability to handle high-dimensional data and mitigate overfitting.

4. Model Training: Trained the Random Forest classifier on the extracted BERT features to learn the underlying patterns and relationships within the data.

5. Evaluation: Evaluated the model's performance using standard metrics such as accuracy, precision, recall, and F1-score to assess its effectiveness in classifying the text data.

7. Result Analysis

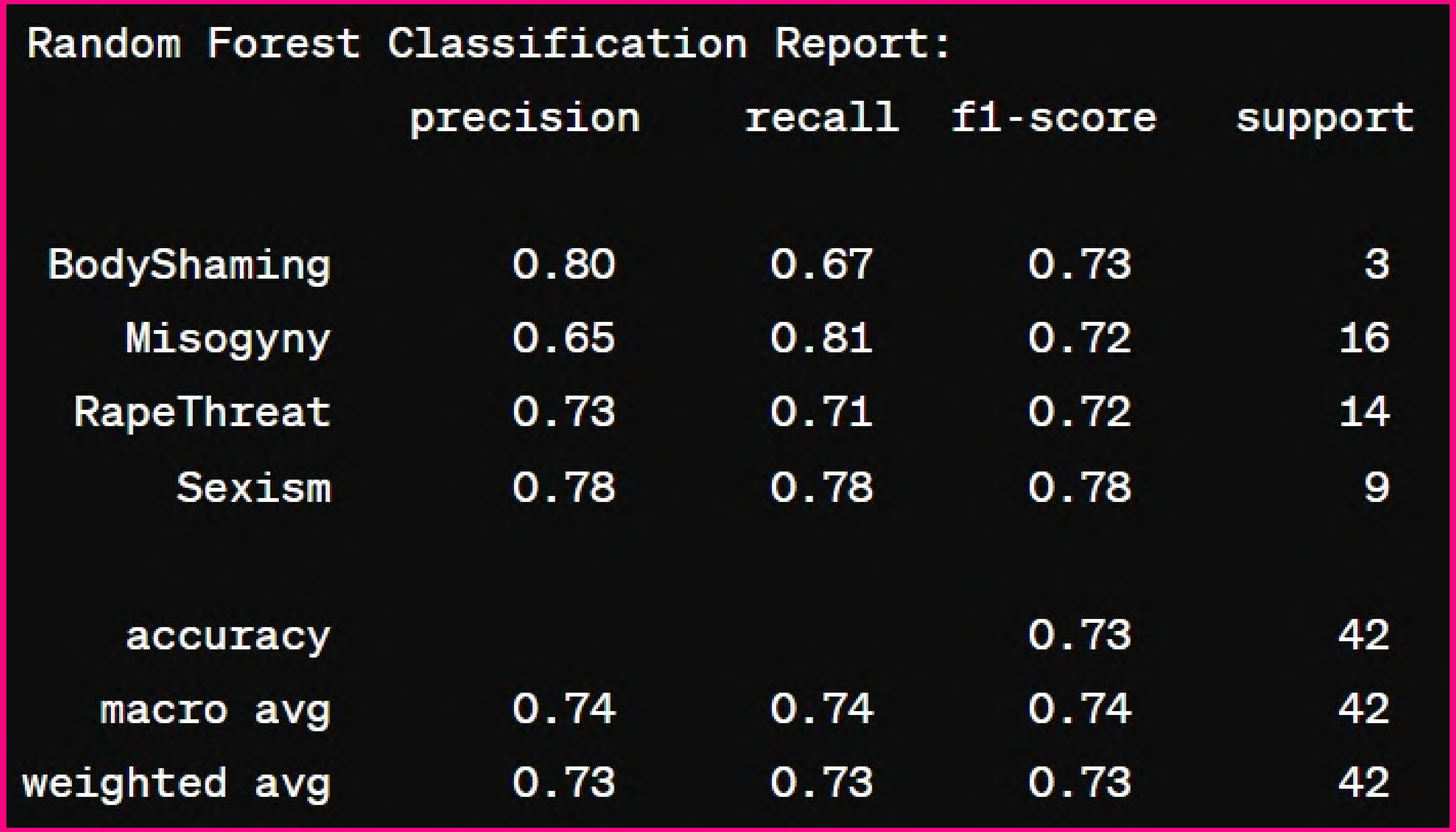
Random Forest displayed higher accuracy, SVM showcased better class-wise metrics, making it a preferable choice for this classification task due to its more consistent and balanced performance across all classes.

RANDOM FOREST

ACCURACY : 73%

SVM

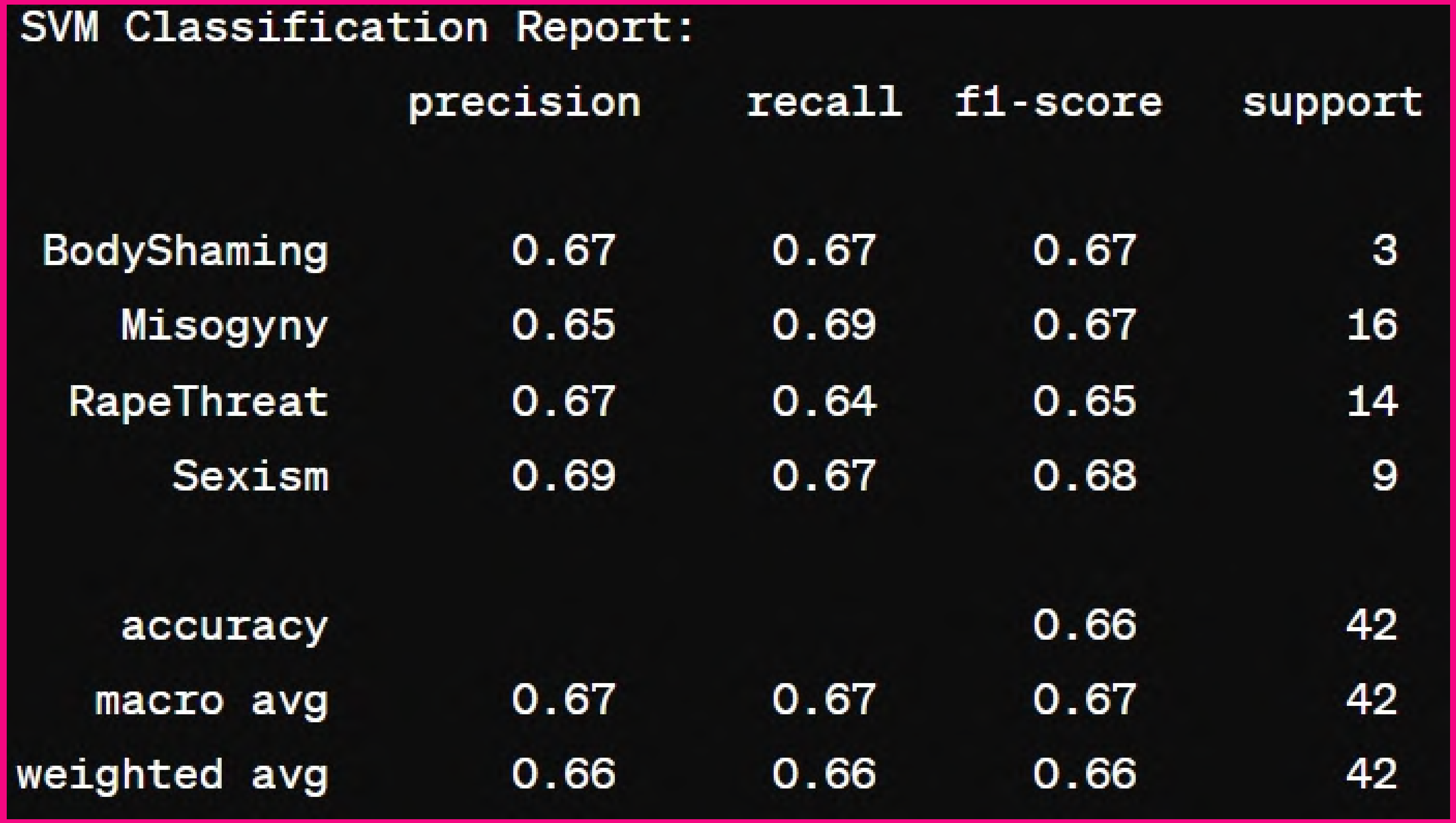
ACCURACY : 66%



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

 BodyShaming       0.80      0.67      0.73         3
    Misogyny       0.65      0.81      0.72        16
  RapeThreat       0.73      0.71      0.72        14
      Sexism       0.78      0.78      0.78         9

    accuracy                           0.73        42
   macro avg       0.74      0.74      0.74        42
weighted avg       0.73      0.73      0.73        42
```



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

 BodyShaming       0.67      0.67      0.67         3
    Misogyny       0.65      0.69      0.67        16
  RapeThreat       0.67      0.64      0.65        14
      Sexism       0.69      0.67      0.68         9

    accuracy                           0.66        42
   macro avg       0.67      0.67      0.67        42
weighted avg       0.66      0.66      0.66        42
```

8. CONCLUSION

In comparing the performance of SVM and Random Forest classifiers on the provided dataset, both models exhibited respectable accuracy levels, with SVM achieving an accuracy of 66% and Random Forest slightly higher at 73%. While Random Forest showed marginally better performance in terms of overall accuracy, SVM demonstrated more balanced precision, recall, and F1-scores across all classes, with all metrics hovering around 67%. Notably, both models struggled with the BodyShaming class, achieving lower precision and recall. Overall, while Random Forest displayed higher accuracy, SVM showcased better class-wise metrics, making it a preferable choice for this classification task due to its more consistent and balanced performance across all classes.

9. Future Scope

FUTURE ADVANCEMENTS

THE FUTURE ADVANCEMENT OF THE FOLLOWING PROJECT INCLUDES :-

1. ADVANCE NATURAL LANGUAGE PROCESSING
2. DEEP LEARNING MODELS
3. COLLABRATION WITH SOCIAL MEDIA PLATFORMS

FUTURE APPLICATIONS

- CONTENT MODERATION
- SENTIMENT ANALYSIS
- RECOMMENDATION SYSTEM
- MARKET RESEARCH
- COMMUNITY ENGAGEMENT

SHORT COMINGS

THE FOLLOWING ARE THE SHORT COMING OF PROJECT :-

1. BIAS AND FAIRNESS CORRECTION
2. FALSE POSITIVE AND NEGATIVIE
3. EVASION STRATEGIES

Thank you

SPECIAL THANKS TO:-

MAYUR GAIKWAD, PREKSHA PAREEK