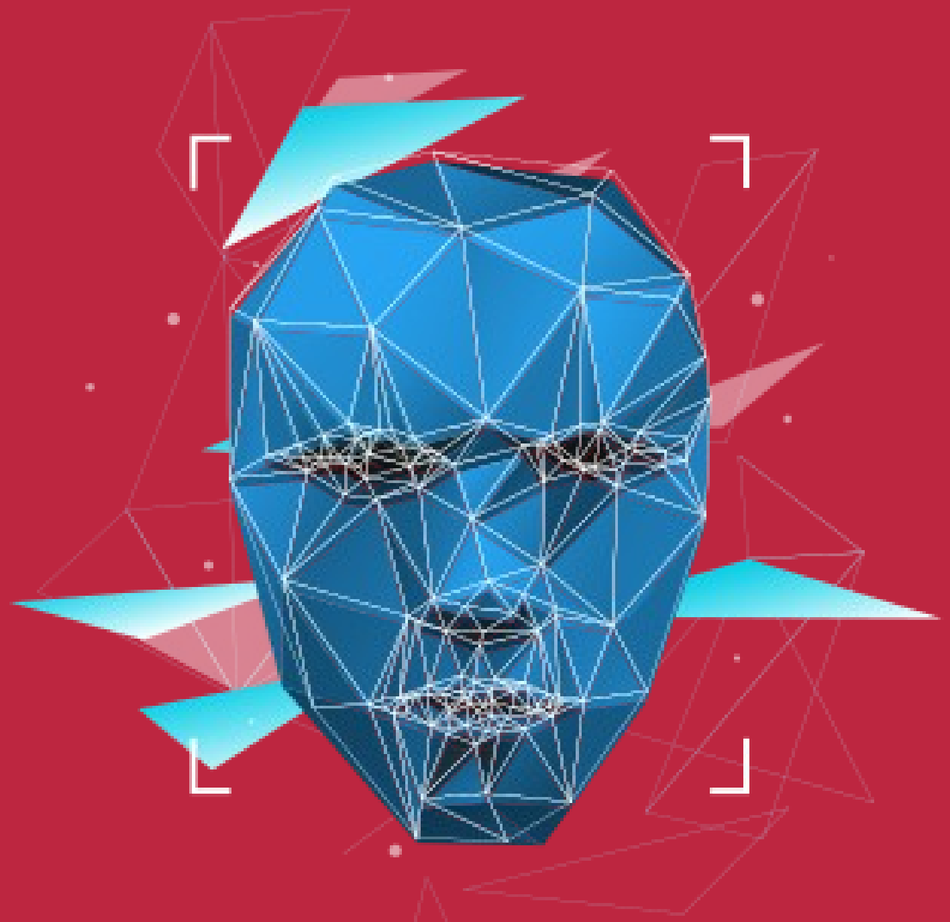
TRI-NIT HACKATHON

Explainable Sexual Harassment Categorization

MACHINE LEARNING



Machine Learning

PRESENTATION OVERVIEW

TODAY'S DISCUSSION

- PS motivation and desc
- Proposed workflow
- Challenges faced in PS
- Results
- Tradeoffs and limitations
- Application Demo

Problem Statement

As sexual harassment becomes more visible, a growing number of courageous victims are stepping forward to tell their stories through online platforms and media outlets. Online platforms offer a powerful tool for raising awareness about sexual harassment, but the current process is often cumbersome for both the readers and victims. Victims often need to manually report these incidents by detailing the occurrences and filling out multiple forms. Readers face an information overload trying to decipher detailed narratives, making it difficult to grasp the nature and severity of incidents quickly. Also, existing methods for processing reports may not effectively categorise and prioritise incidents based on theirseverity, hindering timely and targeted interventions.

Motivation

- The problem statement we chose i.e. , “ Explainable sexual harassment categorization “ is quite unique and less spoken about .
- But it is moral and obvious that it should be one of the main topics that need attention of the privileged.
- There are very few technological solutions currently available to both men and women for this problem.
- The ones that are available are not quite as much specific and are pretty much based to urban areas .
- Our platform takes an advantage of this increase in social media usage to help inform the authorities whenever we detect that an issue needs attention.

Description

- Victims are gaining strengths and speaking about their experiences and more number of cases of sexual harassment are getting into the public eye.

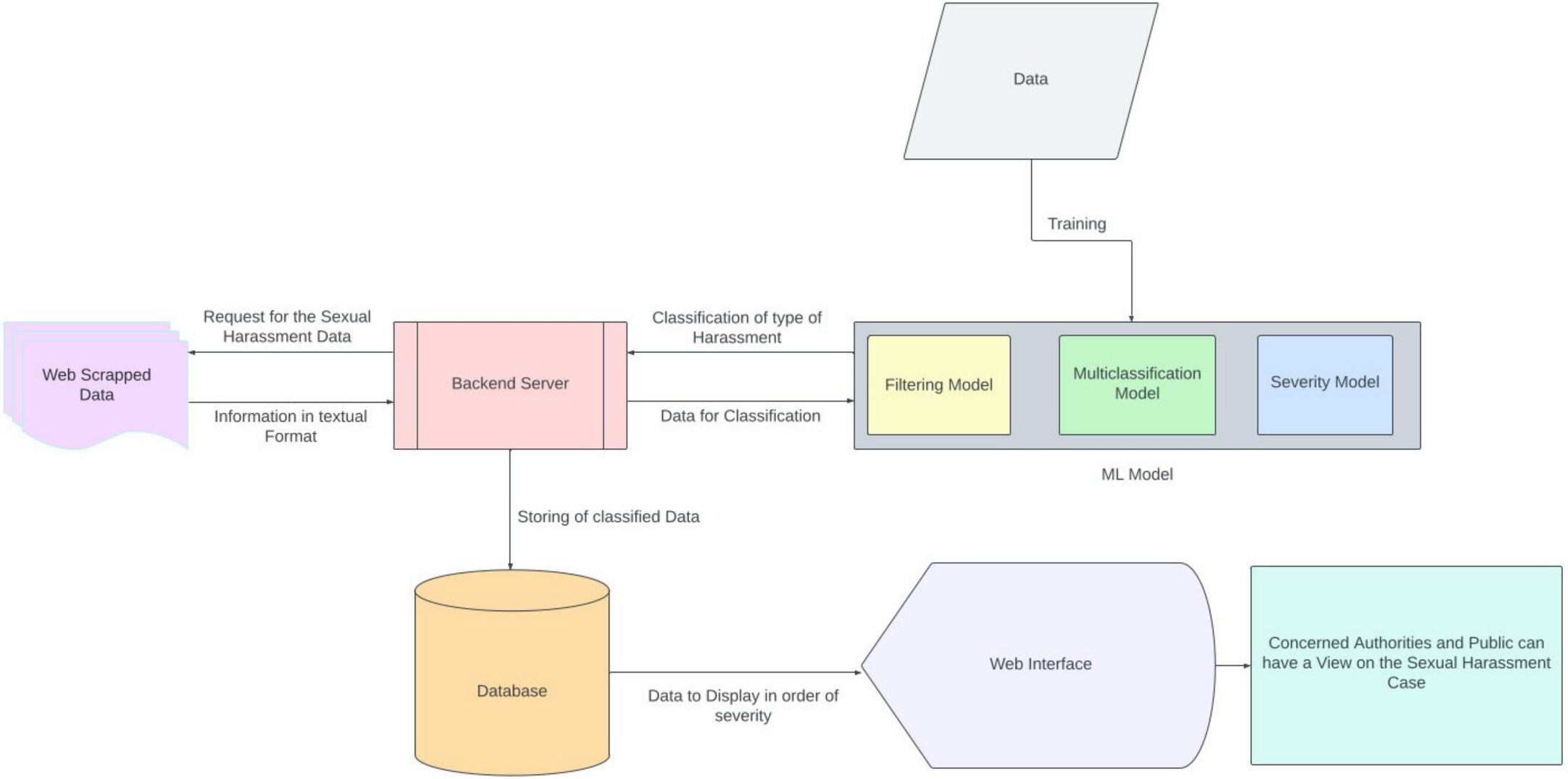
- Victims or witnesses often post about their experiences in social media.

- But the primary issue in this is that people are required to manually report this events and in many cases it is a longand hectic procedure with the primary requirement being that the victims should have the freedom .

- The primary goal is to make a system that minimizes the time and effort that is required to
  report these issues and seek law enforcement .

- The subsequent goals being the categorization of such posts on social media to find the ones that are genuine and actually need help .

PROPOSED WORKFLOW



WorkFlow

1. Web Scraping Component: This component is responsible for extracting data from various online sources. It collects information related to harassment incidents from websites or social media platforms.

2. Machine Learning Model: Trained using the data collected through web scraping, the ML model is responsible for classifying the type of harassment based on various features extracted from the collected data. It helps in automatically categorizing the harassment incidents.

3. Backend Server: The backend server acts as the intermediary between the web scraping component, the machine learning model, and the web interface. It handles incoming requests, processes the data, interacts with the database, and sends responses back to the web interface.

.

WorkFlow

4. Web Interface: The web interface provides a user-friendly platform for users to interact with the system. It displays the classified harassment incidents to the users and allows them to submit new data or queries. Users can view categorized harassment incidents and perform various actions based on the information provided.

5. MongoDB Database: MongoDB is used to store the classified types of harassment. It serves as the persistent storage for storing and retrieving data related to harassment incidents along with their classifications. This allows for efficient data management and retrieval within the system.

Overall, this architecture facilitates the seamless flow of data from web scraping to classification using machine learning, storage in the database, and presentation to users through a web interface, all managed by a backend server. It ensures efficient handling of harassment-related data and provides users with a convenient way to access and interact with the information.

CHALLENGES FACED IN THE PROBLEM STATEMENT

1.Data availability

•The primary challenge faced in the problem statement was the availability of required amount and kind of data.

•In order to properly train the model weneeded datasets other than the ones given by the organizers but since the

•Problem statement is quite rare it was nearly impossible to find the datasets that suited our requirements.

•The solution we found was web scraping but due to increase in issues of privacy breach and misuse of data my third party APIs social media platforms have made it difficult to scrape their sites by blocking libraries that are used for web scraping and also making their official APIs available only to their subscribers.

CHALLENGES FACED IN THE PROBLEM STATEMENT

2. Data preprocessing

•We have primarily considered content from social media platforms like twitter.

•Even after obtaining the data through scrapingwe found that a considerable portion of it did not meet our requirements and we faced challenges in filtering it .

•Also the categorization of data to whether it is related to sexual harassment was a challenging tast as textual data is difficult to categorize and its fine tuning to obtain higher accuracy is complicated .

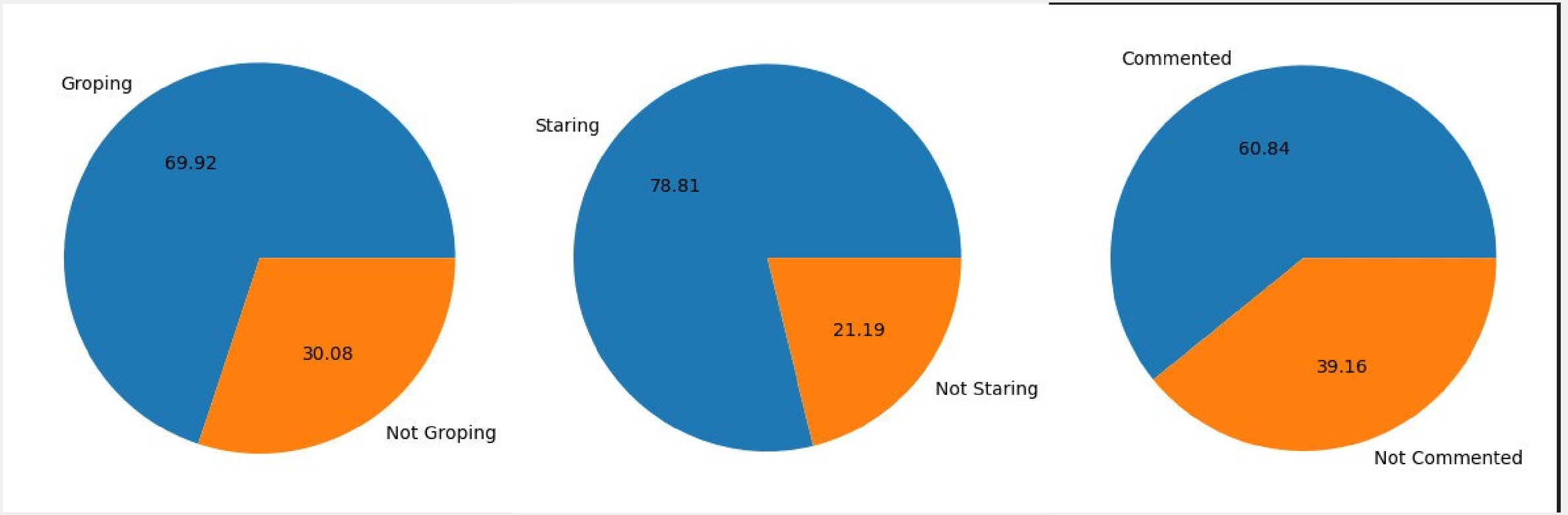
WorkFlow

4. Web Interface: The web interface provides a user-friendly platform for users to interact with the system. It displays the classified harassment incidents to the users and allows them to submit new data or queries. Users can view categorized harassment incidents and perform various actions based on the information provided.

5. MongoDB Database: MongoDB is used to store the classified types of harassment. It serves as the persistent storage for storing and retrieving data related to harassment incidents along with their classifications. This allows for efficient data management and retrieval within the system.

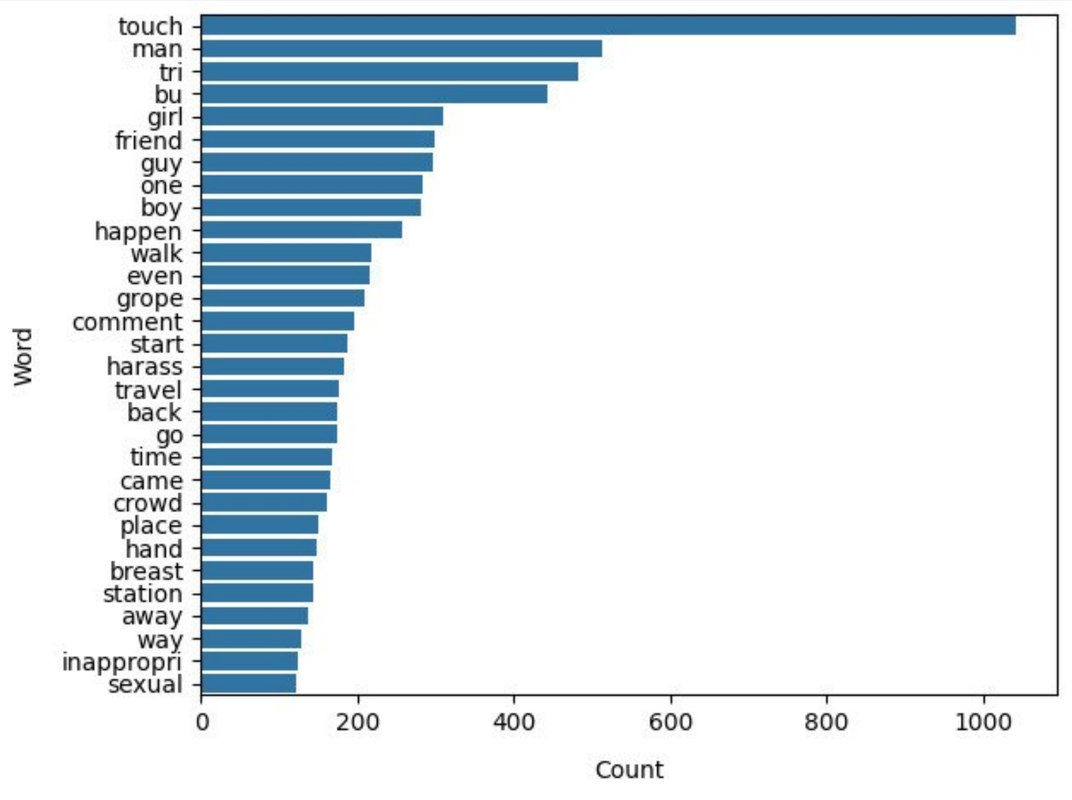
Overall, this architecture facilitates the seamless flow of data from web scraping to classification using machine learning, storage in the database, and presentation to users through a web interface, all managed by a backend server. It ensures efficient handling of harassment-related data and provides users with a convenient way to access and interact with the information.

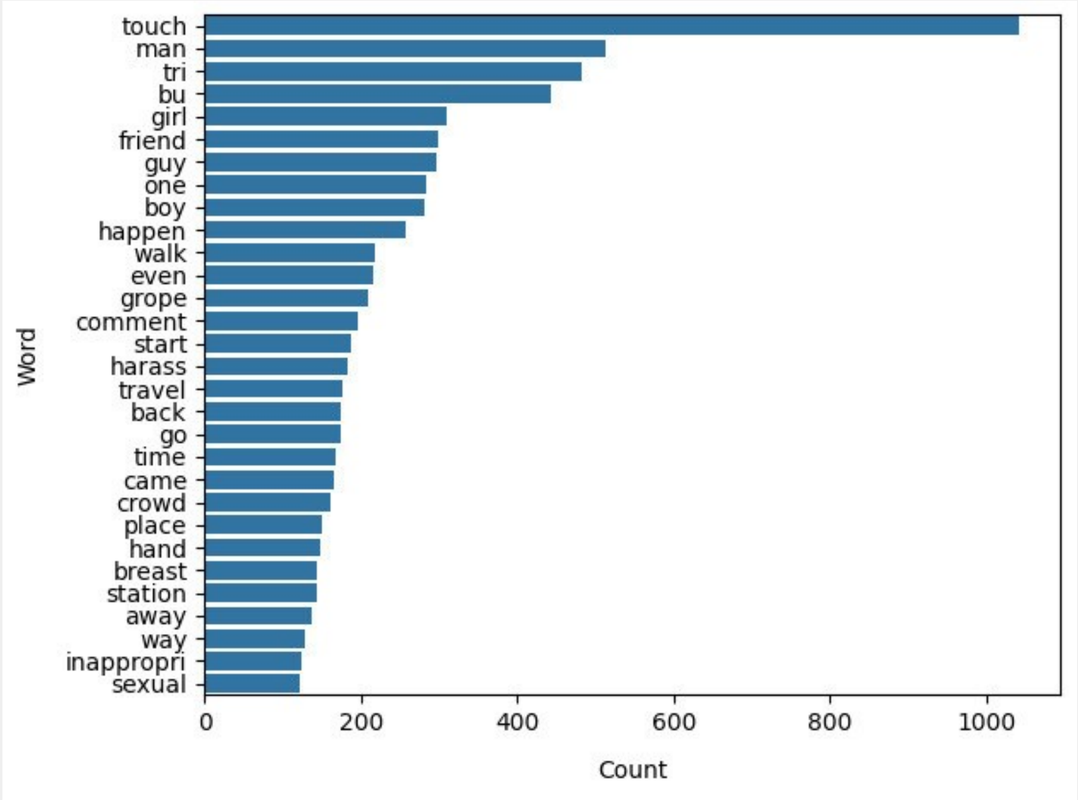
RESULTS



Percentage of the cases involve in one Harassment and
Percentage of people who do not involve in it

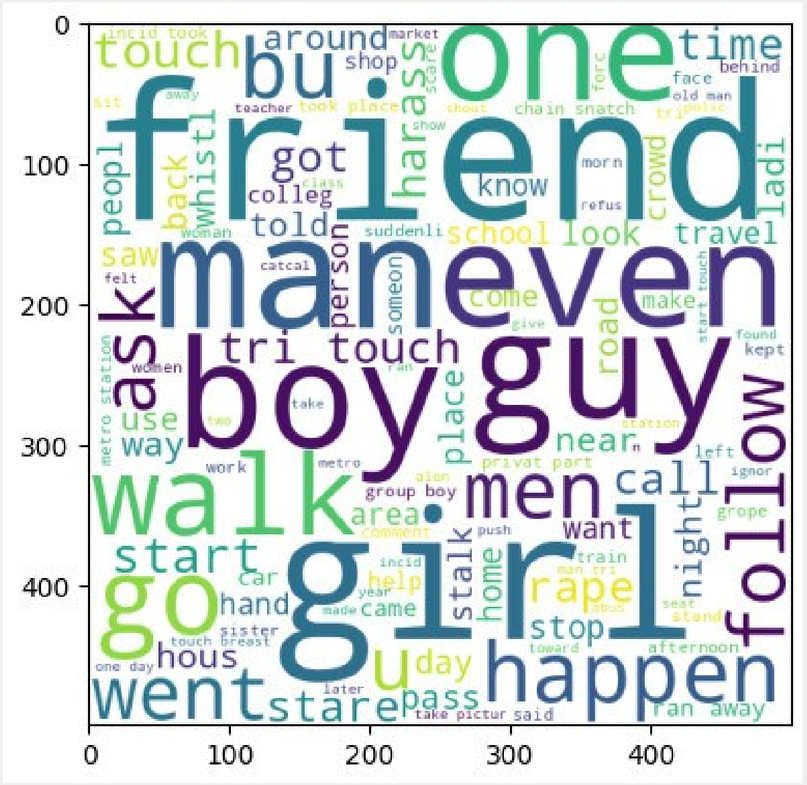
RESULTS

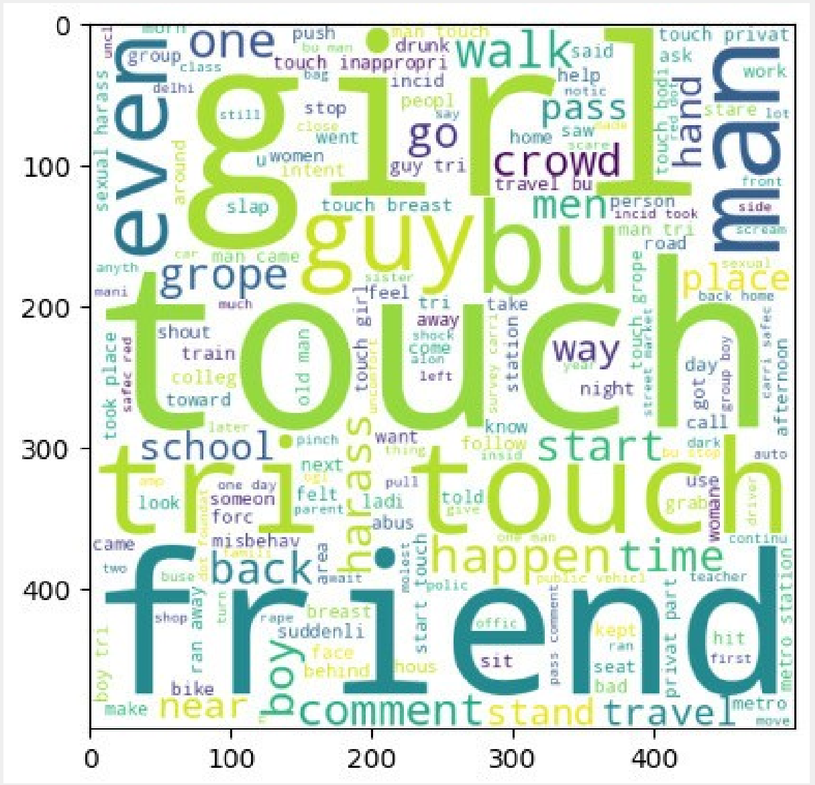




The length of the bars indicate the quantities of the respective attributes

RESULTS



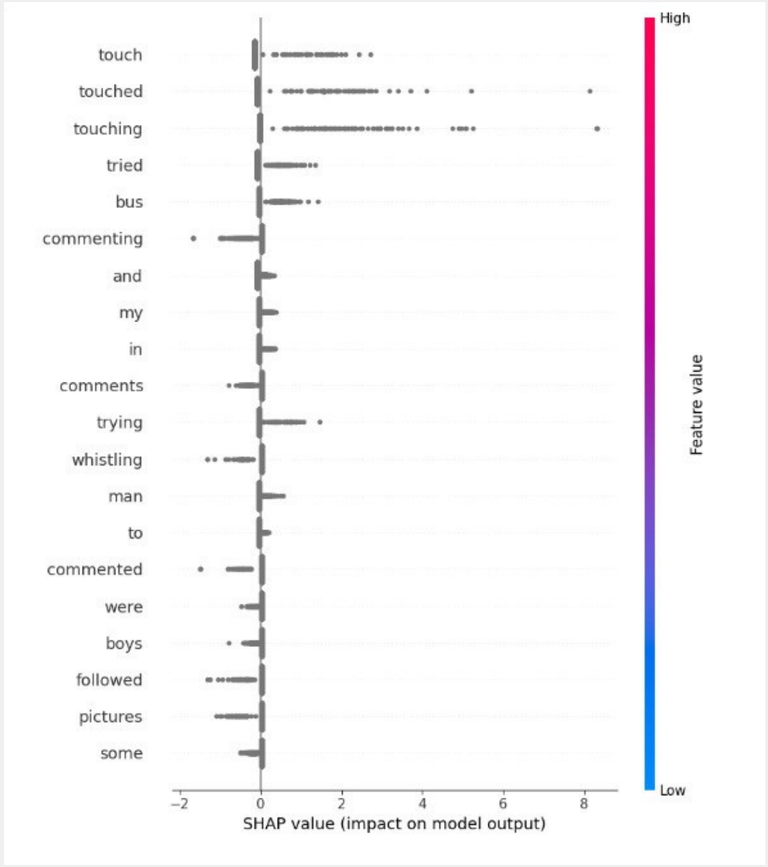


a b c

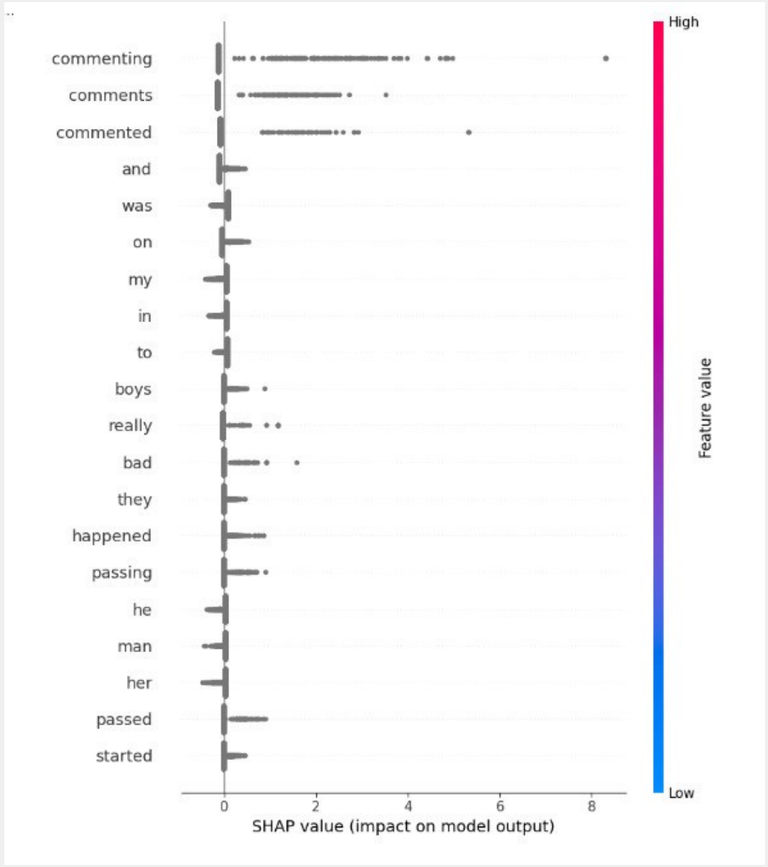
· a. This is the Word Cloud of 1st SHAP graph ·

b.&c. These are the Word Clouds based on 2nd SHAP graph

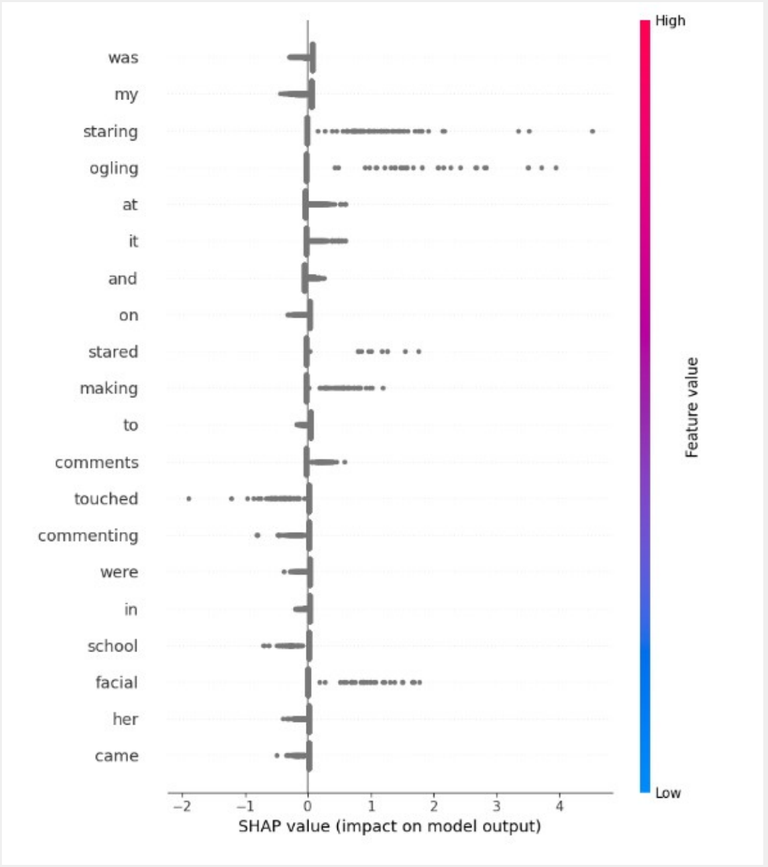
RESULTS



SHAP Values for Touching

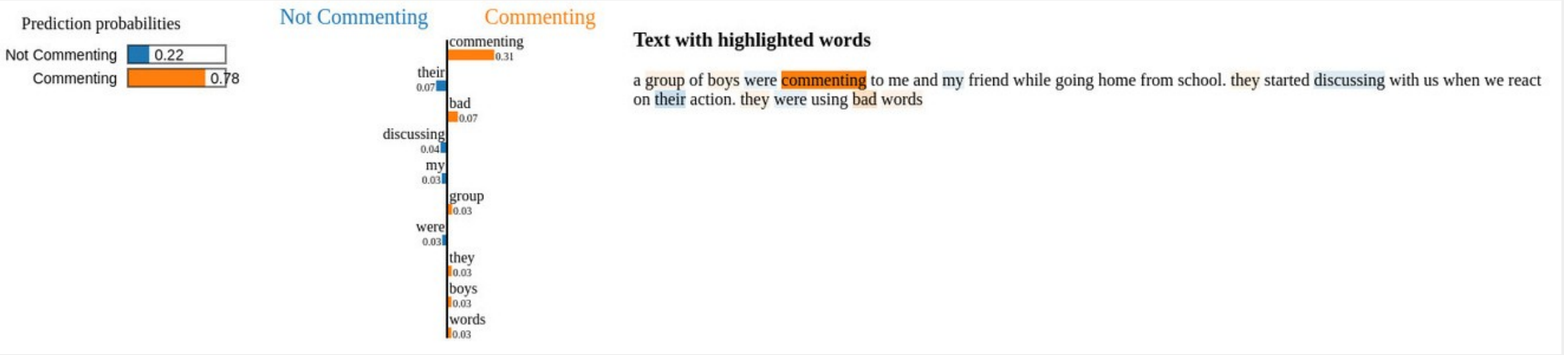


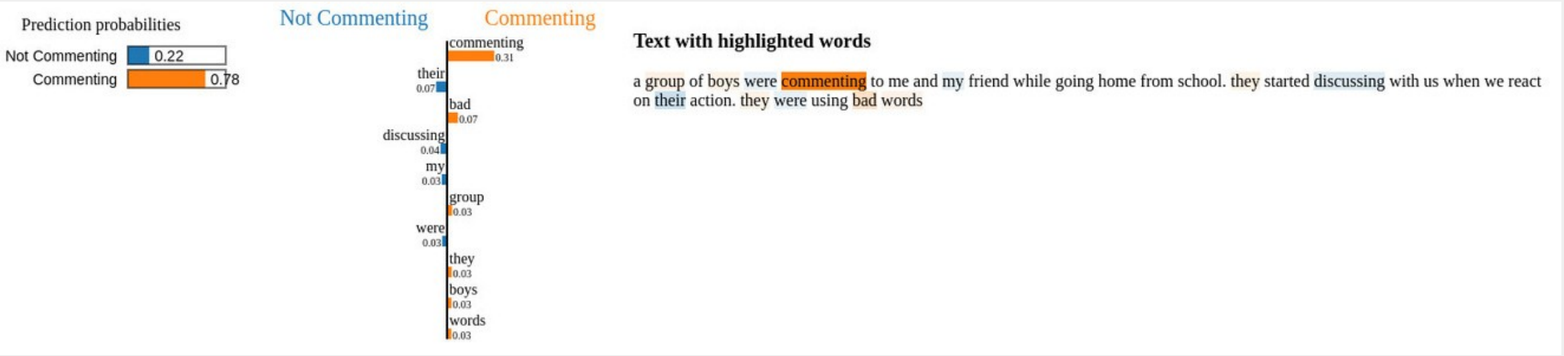
SHAP Values For Commenting

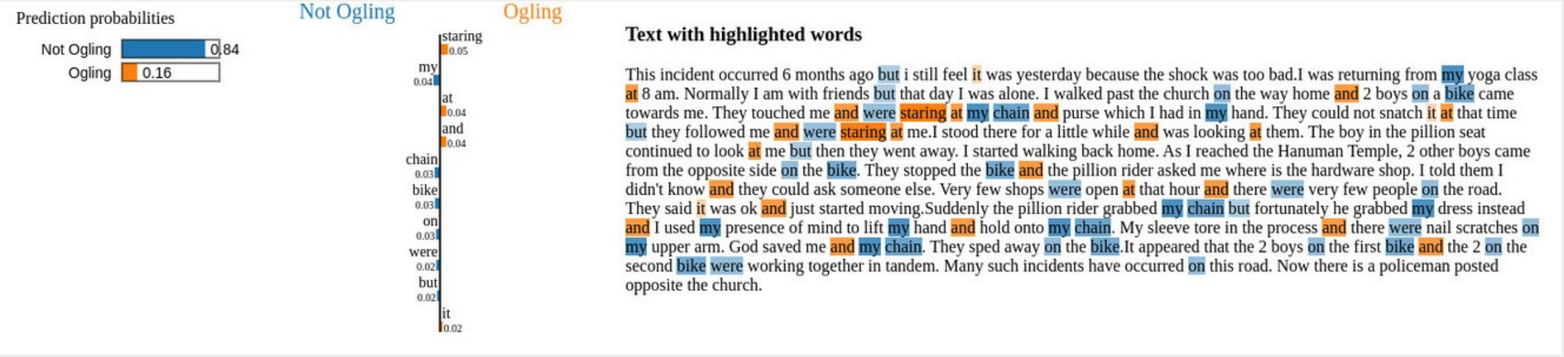


SHAP Values For Staring

RESULTS







Lime Value Analysis

RESULTS

In [101…

```
y_pred_voting=voting.predict(X_test)
print("Accuracy",accuracy_score(y_test,y_pred_voting))
print("Precision",precision_score(y_test,y_pred_voting))
```

```
Accuracy 0.8521859819569744
Precision 0.8782894736842105
```

Accuracy and Precision Score

Tradeoffs and limitations

Tradeoffs:

Accuracy vs. Interpretability:

Complex models may offer higher accuracy but lower interpretability.

Simpler models sacrifice some accuracy for greater interpretability.

Scalability vs. Resource Requirements:

Highly scalable models may require significant computational resources.

Balancing scalability with resource constraints is crucial.

Automation vs. Human Oversight:

Full automation can lead to faster responses but requires human oversight.

Human oversight ensures ethical considerations and addresses contextual nuances.

Tradeoffs and limitations

Limitations:

Data Bias:
Biases in training data may lead to inaccurate categorizations or reinforce existing biases. Addressing data bias is essential for model effectiveness.

Data Source Bias:
Tweets and messages collected from various social media pages vary in terms of their internal meaning, sarcasm, humor. Instagram and X/twitter differ in the context of their tweets.

Application Demo

Thank You

- Team ByteBusters