From Online Activism to Real-World Change: Leveraging NLP to Empower the Fight Against Sexual Harassment Final Report

Manikanta Gajarao
Computer Science
Georgia State University
Atlanta, United States
mgajarao1@student.gsu.edu

I. Introduction

Harassment has become a significant concern in the digital world with the rise of social media platforms. Social media's anonymity and ease of communication have made it a breeding ground for harassment. Harassment can range from cyberbullying to hate speech and can have severepsychological and emotional impacts on its victims. Automated systems that can recognize and flag incidents of harassment have been developed to address this issue. In this context, Logistic regression, XGBoost, and LightGBM are popular machine-learning algorithms used for harassment detection. These algorithms can be trained on large datasets of labeled instances of harassment, which can then be used to automatically detect and flag potential cases of harassment in real time. Such automated systems can play a critical role in creating safe digital spaces by identifying and acting against instances of harassment.

I. Literature Review

Harassment In both machine learning (ML) and natural language processing (NLP), detection is a crucial task. algorithms have been widely used to detect different types of harassment, such as cyberbullying, hate speech, and offensive language. There has been an increase in interest in developing in recent years. ML algorithms for harassment detection and several studies have reported promising results using various ML algorithms.

Logistic regression is a popular ML algorithm used in many NLP applications, including harassment detection. It is a simple algorithm that models the link between one or more independent variables and a binary dependent variable. In harassment detection, the dependent variable is the label (i.e., whether a message is harassing or not). The independent variables are features extracted from the news (e.g., n-grams, sentiment, and syntactic features). Several studies have reported good results using logistic regression for harassment detection, especially when combined with other Support vector machines (SVMs) and random forests (RFs) areexamples of ML algorithms. Another well-liked ML technique that has been extensively applied in the identification of harassment is XGBoost. It is an ensemble technique built on Trees combine several weak classifiers to produce a powerful classifier.

XGBoost has been shown to outperform many other ML algorithms, including logistic regression, SVMs, RFs, sentiment analysis, text categorization, other NLP tasks, and topic modeling. Several studies have reported promising results using XGBoost for harassment detection, especially when combined with other Convolutional neural networks (CNNs) and long short-term memory (LSTM) networks are examples of ML methods. LightGBM is a relatively new MLalgorithm that has gained popularity in recent years. It is also a tree-based ensemble algorithm, like XGBoost, but it uses a different approach to construct decision trees. LightGBM is, especially when working with large datasets, faster and less memory intensive than XGBoost. Several studies have reported promising results using LightGBM for harassment detection, especially when combined with other ML algorithms, such as word embeddingsand attention mechanisms.

In conclusion, ML algorithms, such as logistic regression, XGBoost, and LightGBM, have shown promising results in harassment detection, and they have been widely used in many NLP applications. However, the performance of these algorithms depends on several factors, such as the quality of the dataset, the choice of features, and the hyperparameter settings. Therefore, further research is needed to investigate the adequacy of these algorithms in different contexts and to identify ways to improve their performance.

II. Description

1) ***Business Understanding:*** Numerous personal accounts of sexual harassment and abuse are now being shared on the internet. To advance the battle against these reprehensibleacts, I propose a task that involves categorizing different types of harassment on the stories posted on the platform SafeCity which is online. This task highlights the benefits of extracting features from these stories, which can assist in the automatic completion of incident reports, identification of unsafe areas, avoidance of hazardous practices, and the identification of perpetrators. (What is the problem).

2) ***Business Problem:*** Understanding and organizing the information contained in these stories is a daunting challenge that can be addressed through natural language processing (NLP). NLP has the potential to bridge the gap between online activism and concrete action. To this end, we introduce neural models that can analyze large volumes of harassment data from social media, making them valuable tools for raising awareness, promoting understanding, and enabling prompt action. This summarizing, analyzing, and aggregating the information automatically can assist activist groups in educating the public. (why is it a problem).

III. Dataset

Dataset link: -
https://drive.google.com/drive/folders/1fMN7b7UrUEetdjLj_mUta4a3YcWwJY-k?usp=drive_link

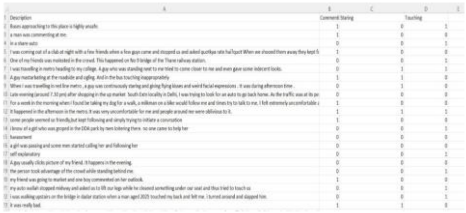
I have done the Classification of single labels and multiple labels.

*1) **Single-Label Classification:*** The incident is described in the first column of the two-column, single-label classification data. Sexual harassment is denoted in the second column by a one rather than a zero. After randomly selecting and setting aside ten percent of each dataset for the test set, the remaining training data were separated into adevelopment set. Each category contains, in turn, 7201 training samples, 990 development samples, and 1701 test samples.

	Description	Category
1	Description	Category
2	Was walking along crowded street, holding mums hand, when an elderly man groped butt, I turned to look at him and he looked away, and did it again after a while. I was 12 yrs old then.	0
3	This incident took place in the evening. I was in the metro when two guys started staring.	1
4	I WAS WAITING FOR THE BUS. A MAN CAME ON A BIKE AND WAS OFFERING LIFT TO YOUNG GIRLS.	0
5	Incident happened inside the train	0
6	I witnessed an incident when a chain was brutally snatched from an elderly lady . The incident took place during evening.	0
7	ONE DAY I WAS WALKING AT [illegible] I SAW A BOY TRY TO TAKE PICTURE OF A GIRL [illegible]	0
8	I have heard about indecent people making indecent comments	0
9	When I entered the subway it was dark, so I decided to go back from there, but suddenly a man came to me and held my hand and forced me to go from that subway, and also asked me my number. [illegible]	0
10	Sexual invites	0
11	poor street lighting	0
12	Catcalls and passing comments were two of the ghastly things the traffic police at the International airport put me and my friend through. It is appalling that the protectors and law enforcers at the airport can make someone so uncomfortable.	1
13	This happened in a bus, public transport. It was in broad daylight. I was on my way to college and the bus wasn't that crowded. There was enough space for two people to stand back to back with big bags on too. This particular male was deliberate	0
14	One night we were returning to our house from party. I saw one boy was trying to touch one young girl. [illegible] the girl was so afraid and was crying. When we came near to them the boy ran suddenly.	0
15	Some people used to stare in a very inappropriate way which is not tolerable. It happened in the evening and night	0
16	local college guys were commenting	0
17	making weird noises and laughing at girls	0
18	I had gone to get ice cream and I noticed 5 men who looked drunk there, I turned around and started walking away, but they made sexual invites from afar	0
19	There were comments and staring when I was going back to my p.g. During the afternoon hours.	1
20	my friend was crossing the road	1

Fig. 1. Single label Classification dataset.

*2) **Multi-Label Classification:*** The incident description is the first column in the data for multi-label categorization, and the second, third, and fourth columns are either 1 or 0, depending on whether the category of sexual harassment is present or not. The reserve for the test set was randomly chosen to be ten percent of the dataset. The other ten percent of the remaining training data were distributed to the development set at random. 7201 training samples, 990 development samples, and 1701 test samples are all available.



	A	B	C	D	E
1	Description	Comment Staring	Touching		
2	Buses approaching to this place is highly unsafe.	1	0	1	
3	a man was commenting at me.	1	0	0	
4	in a share auto	0	0	1	
5	I was coming out of a club at night with a few friends when a few guys came and stopped us and asked quotkya rate hai!quot When we shooed them away they kept f	1	0	0	
6	One of my friends was molested in the crowd. This happened on No 9 bridge of the Thane railway station.	0	0	1	
7	I was travelling in metro heading to my college. A guy who was standing next to me tried to come closer to me and even gave some indecent looks.	0	1	1	
8	A guy masturbating at the roadside and ogling. And in the bus touching inappropriately	1	1	0	
9	When I was travelling in red line metro , a guy was continuously staring and giving flying kisses and weird facial expressions . It was during afternoon time .	0	1	0	
10	Late evening (around 7.30 pm) after shopping in the up market South Extn locality in Delhi, I was trying to look for an auto to go back home. As the traffic was at its pe	0	0	0	
11	For a week in the morning when I found be taking my dog for a walk, a milkman on a bike would follow me and times try to talk to me. I felt extremely uncomfortable a	1	0	0	
12	It happened in the afternoon in the metro. It was very uncomfortable for me and people around me were oblivious to it.	1	1	1	
13	some people seemed us friendly,but kept following and simply trying to initiate a conversation	1	0	0	
14	i know of a girl who was groped in the DDA park by men loitering there. no one came to help her	0	0	1	
15	harassment	0	0	0	
16	a girl was passing and some men started calling her and following her	0	0	0	
17	self explanatory	0	0	1	
18	A guy usually clicks picture of my friend. It happens in the evening.	0	0	0	
19	the person took advantage of the crowd while standing behind me.	0	0	1	
20	my friend was going to market and one boy commented on her outlook.	1	0	0	
21	my auto wallah stopped midway and asked us to lift our legs while he cleaned something under our seat and thus tried to touch us	0	0	1	
22	i was walking upstairs on the bridge in dadar station when a man aged 20/25 touched my back and felt me. i turned around and slapped him.	0	0	1	
23	It was really bad.	1	1	0	

Fig. 2. Multi-label Classification dataset

IV. Proposed Analytics Solution

Our study introduces a system for the single-label as well as multi-label classification of various types of harassment, which is sexually, present in the stories shared on the SafeCity. Use this website to report specific instances of sexual assault and harassment using crowdsourcing techniques. where the story includes numerous forms of sexual harassment tags and a thorough account of the incident. For instance, the narrative” university was close by. This frequently occurred. Men making comments, ogling, and attempting to touch”. indicates three categories of sexual harassment: commenting, ogling/staring, and touching or groping.

Numerous forms of sexual harassment have been automatically classified, benefiting both the victim and the authorities. It can speed up filling out online sexual violence reporting forms process, typically requiring the victim to specify the sexual harassment encountered. By partially completing the report, it may increase the likelihood of the victim reporting the incident. Moreover, this system can also be utilized to categorize and summarize numerous online testimonials that describe or report sexual harassment, meeting the need for automatic classification

V. Architecture

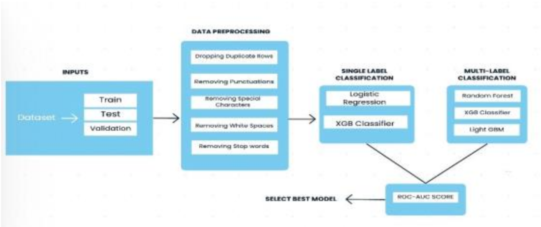


Fig. 3. Architecture

VI. Data Exploration and Pre-Processing

Data exploration involves using statistical and visualization methods to understand the data, allowing users to identifyany issues in the dataset and select the appropriate model for further analysis. Without data exploration, problems within thedataset may go unnoticed, leading to erroneous conclusions. Data preprocessing is data preparation; it is the main step in the data mining process. Data processing transforms data intoa format that is easily processed later. In this section, you should describe the steps taken to make data clean while re-moving missing values and data type conversions. 1. Droppingduplicate description rows 2. Removing punctuation symbols 3. Removing special characters 4. Deleting white spaces and Stopwords 5. Convert all the characters into small letters

VII. Exploratory Data Analysis (EDA):

This section requires performing Exploratory Data Analysis (EDA) to gain insights into the distribution of variables, detect patterns, and uncover any relationships between variables. To effectively communicate these findings, visualizations like histograms, scatter plots, and heat maps should be included.

Single Label classification and Multi-Label classification

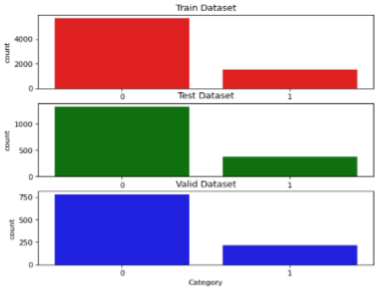


Fig. 4. Commenting

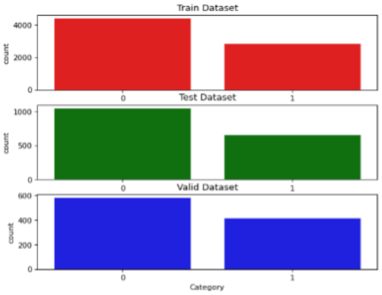


Fig. 5. Ogling.

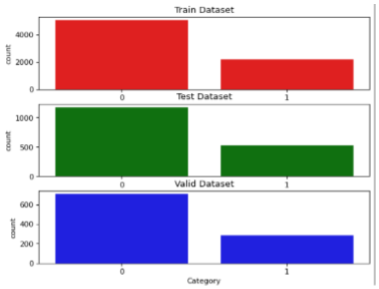


Fig. 6. Groping.

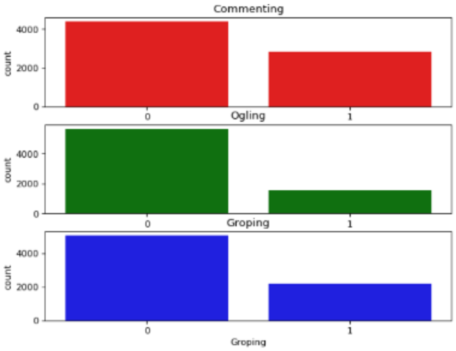


Fig. 7. Commenting vs. Ogling vs. Groping.

VIII. TF-IDF VECTORIZER (TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

A common method in (NLP) for transforming text data into a numerical Term frequency-inverse document frequency is a format that machine learning algorithms may employ. (TF-IDF). The algorithm functions in the following manner:

The TF-IDF vectorizer first determines the frequency—often referred to as each word's frequency throughout the text. The algorithm then determines the inverse document frequency, which provides an estimate of the word's rarity over the full corpus of documents by taking the ratio of the logarithm of the overall document count to the number of documents containing a particular word. The TF-IDF score is ultimately calculated by dividing the word by simply dividing the document frequency by its inverse.

TF= Number of repetitions of word/ Number of words in document IDF= log [Number of documents/ Number of documents contains the word]

Reason: The primary limitation of the Bag-of-Words approach can be overcome using the TF-IDF technique, which incorporates the concept of inverse document frequency.

X.ROC-CURVE (RECEIVER OPERATINGCHARACTERISTIC)

The performance of this classification model is graphically represented by the Receiver Operating Characteristic curve. For various model threshold settings, it shows the link betweenthe true positive rate and the rate of false positives.

• **TPR**: The proportion of positive cases that the classification model actually and accurately classifies as positive.
• **FPR**: The ratio of true negative cases that the model incorrectly classifies as positive cases • Threshold value: The probability value at or over which the classification model labels a case as positive.
Reason: These indicators shed light on how well the categorization model is performing. An improved model's performance and capacity to distinguish between positive and negative classes is shown by a higher AUC value. A good AUC score typically ranges from 0.8 to 0.9.

XI: MACHINE LEARNING MODELING:

*1) **Single Label Classification-Logistic Regression:*** It isa frequently employed algorithm in NLP with applications spanning from sentiment analysis and text classification to language identification. Its primary objective is to forecast the probability of a binary outcome (e.g., positive, or negative sentiment) by examining a set of input features (such as words in a text document). Let's explore how logistic regression can be utilized for sentiment analysis in NLP:

Data Preprocessing: The initial stage in text data preparation is to clean and convert it into a modeling-friendly format. This often entails activities such as stemming, lemmatization, stop word removal, and transforming text into numeric features, utilizing techniques like TF-IDF or bag-of-words.

Feature Extraction: The pertinent features are extracted from text data which is preprocessed. Techniques such as n-grams

can be employed to accomplish this, where the frequency of word pairs or triplets is utilized as features.
Model Training: After feature extraction, the data can be utilized to train a logistic regression model. The model will acquire knowledge of the connection between the input features and the binary outcome (positive or negative sentiment).Model Evaluation: Following model training, it can be assessed on a test set to determine its accuracy and performance. This involves employing evaluation metrics like precision, recall, and F1 score.

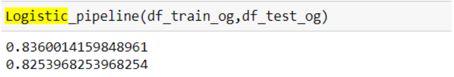
Model Deployment: Following training and evaluation, the model can be put into operation to categorize new text data into positive or negative sentiment categories. Logistic regression is a straightforward and powerful algorithm for NLP applications, particularly for sentiment analysis. It is extensively used in both academic and industrial settings because of its simplicity and interpretability. Nonetheless, for tasks involving extensive and intricate datasets, it may not perform as well as more sophisticated algorithms such as neural networks.

2) ***Random Forest:*** A more reliable and accurate model for regression and classification issues is produced by theensemble learning algorithm Random Forest by combining several decision trees. The training data and input qualities used to build each tree in the forest are varied, and the outputs of several trees are combined to produce predictions. After the algorithm is trained using a bootstrap sample ofthe original dataset, the best split for each node is selected using a criterion like the Gini index or information gain. An effective and well-liked technique called Random Forest can handle missing values and outliers in the data and is resistant to overfitting. Its performance can be further improved by adjusting hyperparameters.

3) ***LightGBM(Light Gradient Boosting Machine):*** Another powerful machine learning library is LightGBM, which is designed for efficient memory usage and computation speed. LightGBM uses decision tree algorithms for classification and regression problems and employs a leaf-wise tree growth strategy that prioritizes the growth of the leafnodes with the largest loss reduction, leading to a deeper and more accurate tree structure. LightGBM is particularly useful for handling large datasets with millions of records and high-dimensional feature spaces, thanks to techniques such as data partitioning, histogram-based feature selection, and gradient-based one-sided sampling. Reason: LightGBM is preferred for its high speed, low memory usage, and focus on accuracy. However, one of its drawbacks is its sensitivity to overfitting, which can easily occur with small datasets. Additionally, Light GBM is six times faster than XGBoost.

IX. METHODOLOGY:

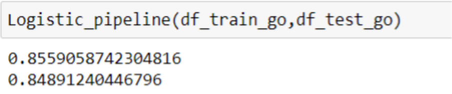
1) ***Single Label Classification-Logistic Regression:*** For this, a logistic regression pipeline has been implemented using the sci-kit-learn library. The pipeline consists of several steps, including splitting the dataset into input and outputvariables, the data conversion of the text using features like Tf-idfVectorizer, fitting the logistic regression assessing the model's performance on the test set by calculating the ROC-AUC score and classifier accuracy. The function 'logistic pipeline' takes two arguments, 'data' and 'test,' which are pandas data frames containing the training and testing data, respectively. Using Tf-idfVectorizer, the text data in both data frames is converted into numerical features. Logistic regression is then used to apply the training data to build the classification model, which helps in assuming the test data. The ROC-AUC score and accuracy, which are shown on the console, are used to assess the classifier's performance.



```
Logistic_pipeline(df_train_og,df_test_og)

0.8360014159848961
0.8253968253968254
```

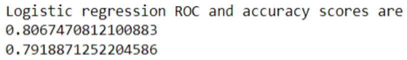
Fig. 8. Ogling



```
Logistic_pipeline(df_train_go,df_test_go)

0.8559058742304816
0.84891240446796
```

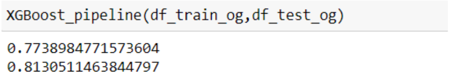
Fig. 9. Groping



```
Logistic regression ROC and accuracy scores are
0.8067470812100883
0.7918871252204586
```

Fig. 10. Commenting

2) ***XGB Classifier:*** The XGBoost pipeline for sci-kit-learn categorization is implemented in the code below. The pipeline includes several phases, such as separating the input and output variables, turning text data into numerical features,fitting the XGBoost model, and assessing the classifier's performance on the test set. The two input arguments" data" and "test," which are Panda's data frames holding the training and testing data, respectively, are passed to the function "XGBoostpipeline." In both data frames, the text information is converted into numerical features by TF- idfVectorization.T On training data, the classification model iscreated by employing the XGBoost classifier, and it is then used to forecast the test data. The ROC-AUC score and accuracy, which are reported to the console, are then usedto assess the classifier's performance.



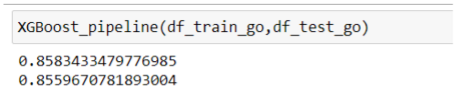
```
XGBoost_pipeline(df_train_og,df_test_og)

0.7738984771573604
0.8130511463844797
```

Fig. 11. Ogling

XII. MULTI-LABEL CLASSIFICATION:

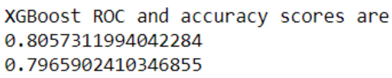
1) ***Random Forest vs. XGB:*** This code performs multi-label classification utilizing (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). It divides the multi-label problem into several binary classification tasks using the



```
XGBoost_pipeline(df_train_go,df_test_go)

0.8583433479776985
0.8559670781893004
```

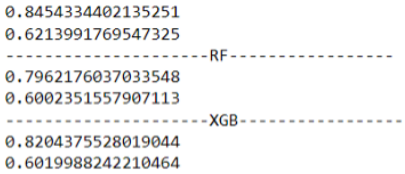
Fig. 12. Groping



```
XGBoost ROC and accuracy scores are
0.8057311994042284
0.7965902410346855
```

Fig. 13. Commenting

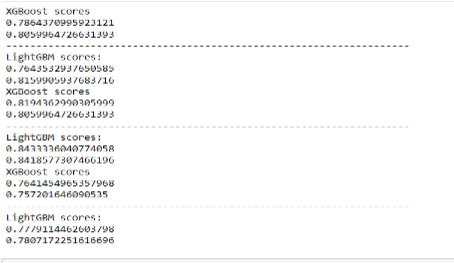
Binary Relevance Approach. Each classifier is trained using training data (X_train y), and then the code makes predictions using test data (X_test. Using the target variables' anticipated values and actual labels (y Gao, S., Chen, H., Zhou, Y. (2018). A sentiment analysis model for cyberbullying detection on social Journal of Computational Social Systems, IEEE, 6(4), 902-912.test)," Commenting," Ogling/Facial Expressions/Staring," and" Touching/Groping," it then determines the ROC-AUC score and accuracy for each of the three classifiers. The SVM classifier is initialized with the Binary Relevance wrapper to enable binary classification on each target variable separately. Like this, Binary Relevance is also used to wrap the Random Forest and XGBoost classifiers. The outcomes demonstrate the accuracy and ROC AUC. score for each classifier on the test data. This code helps compare how well various classifiers perform when dealing with problems involving many labels.



```
0.8454334402135251
0.6213991769547325
---------------------RF-----------------
0.7962176037033548
0.6002351557907113
---------------------XGB----------------
0.8204375528019044
0.6019988242210464
```

Fig. 14. RF vs XGB

*2) **LightGBM vs XGBoost:*** The performance of the XGBoost and LightGBM classifiers on text data that has been encoded into sentences using the Universal Sentence Encoder is contrasted in this code. Using the embed_test () function, the input data (X test and_X test) are first transformed into sentence embeddings. Then, using the transformed data, XGBoost and LightGBM classifiers are trained. Initial hyperparameters for the XGBoost classifier include the n test,learning test, and max test. Predictions are made using the testdata after the classifiers have been trained, and the accuracy and ROC AUC score of each classifier is determined using the original labels (y test) and predicted values. Accuracy and ROC AUC scores for both classifiers are shown in the final output. This code helps assess and contrast the effectiveness of various classifiers on text data.



```
XGBoost scores
0.7864370995923121
0.8059964726631393
----------------------------------------------------------------------
LightGBM scores:
0.7643532937650585
0.8159905937683716
XGBoost scores
0.8194362990305999
0.8059964726631393
----------------------------------------------------------------------
LightGBM scores:
0.8433336040774058
0.8418577307466196
XGBoost scores
0.7641454965357968
0.757201646090535
----------------------------------------------------------------------
LightGBM scores:
0.7779114462603798
0.7807172251616696
```

Fig. 15. LightGBM vs XGBoost

XIII: CONCLUSION

An essential use of machine learning (ML) and natural language processing (NLP) techniques is the detection of harassment. In this task, I identified instances of harassment in text, which can help prevent such behavior and promote a safer online environment. I applied several ML algorithms to develop harassment detection models, including Logistic regression, XGBoost, and LightGBM. In terms of single-label classification, the XGB classifier demonstrated the best performance. On the other hand, for multi-label classification, LightGBM exhibited the best results. These algorithms are popular due to their high performance and interpretability, which makes them suitable for real-world applications. Overall, the performance of ML algorithms for harassment detection can depend on several factors, such asthe quality and size of the dataset, the choice of features and hyperparameters, and the evaluation metrics used. To get the optimum performance for the work at hand, it is crucial to correctly choose and optimize the ML algorithms.

XIV: Future Scope

A) While the report focused on Logistic Regression, XGBoost, and LightGBM, there may be other machine learning algorithms worth exploring for this task, such as deep learning models (e.g., LSTM, Transformers) that could potentially capture more nuanced features of harassment.

B) The report suggests there is room for improving the performance of the harassment detection models. Experimenting with different feature engineering techniques, hyperparameter tuning, and ensemble methods could help boost the accuracy and robustness of the models.

C) Allowing users to provide feedback on the model's performance and decisions could help refine the system over time through active learning techniques.

XV: REFERENCES

[1] Mohapatra, P., Jena, S. K., Sahoo, B. K. (2020). Detection of cyber-bullying utilizing algorithms for machine learning. Journal of Computer Science and Mobile Computing International, 9(8), 93-102.

[2] Chand, S. S., Madhavi, K. (2019). Cyberbullying Detection using methods of machine learning. Journal of Advanced Computer Science Research International, 10(4), 74-78.

[3] Saini, M., Bansal, R. (2020). A comparison of machine learning techniques for detecting cyberbullying. An international journal of computer science and engineering research, 8(1), 1-7.

[4] Sheikh, N. A., Gupta, A. (2020). Sentiment Analysis for Cyberbullying Machine learning techniques for detection. Journal of Advanced Science and Technology International, 29(11), 1575-1583.

[5] Sari, E. Y., Syarifuddin, D. (2019). Cyberbullying detection on Indonesian social media using XGBoost. Journal of Physics: Conference Series, 1238(1),01234