# SAFE ONLINE SPACES: DETECTING CHILD PREDATORS AND CYBER HARASSERS IN SOCIAL MEDIA ENVIRONMENTS

**K Sowjanya Lakshmi** Dept of Computer science and Engineering Vignan's Foundation for Science, Technology and Research sowjikolasani29@gmail.com
**K Sumana Angel** Dept of Computer science and Engineering Vignan's Foundation for Science, Technology and Research sumana.angel00@gmail.com
**G Bhoomika** Dept of Computer science and Engineering Vignan's Foundation for Science, Technology and Research gavinibhumika@gmail.com
**Dr J Vinoj** Dept of Computer science and Engineering Vignan's Foundation for Science, Technology and Research vinojbu@gmail.com
**Dr S. Gavaskar** Department of MCA, Bharathiar University

***Abstract*—**
In the current digital era, social media platforms have become an integral part of our lives, connecting individuals worldwide. Despite their benefits, these platforms have also exposed children and other vulnerable groups to various online threats, including cyberbullies and child predators. These malicious actors exploit the anonymity and reach of social media to harm others. Traditionally, countering these threats relied heavily on manual reporting and human moderators who investigated reported suspicious activities to determine if they violated platform policies. However, this reactive approach often led to delays, allowing harmful content to spread before action could be taken. Recognizing the need for more proactive and efficient solutions, researchers have turned to machine learning, a subset of artificial intelligence that enables computers to learn from data and make predictions. The goal is to develop an automated system that can swiftly and accurately identify potential online child predators and cyber harassers. This proposed machine learning-based approach offers several advantages over traditional methods. The most significant benefit is the substantial reduction in platform response times, facilitating the rapid removal of offensive content and users.

**Keywords—**
child predators, cyber harassers, Natural language processing, Machine Learning

## INTRODUCTION

Social media platforms have completely changed how we connect and communicate in the current digital era. Along side others. Without a question, these platforms have created a great deal of opportunity for global engagement, but they have also created a number of serious issues. These difficulties include the presence of people who abuse the These malicious actors exploit the internet's anonymity and reach for harmful purposes.such as child predators and cyber harrassers. Protecting people from these internet dangers has become a top priority, especially for kids. Since the early days of the internet, when problems like child predators and cyber harrassers first surfaced, there has been a history of responding to online threats. Numerous initiatives have been taken to counter these risks throughout the years.Legal actions, user-focused education campaigns, and the creation of technology-based solutions have all been a part of these initiatives. With the development of technology, these malicious actors take advantage of the internet's anonymity and reach for harmful purposes. especially in the areas of data analysis and machine learning, new avenues for the more accurate detection and mitigation of these online threats opened up.Practical reasons dictate the necessity for an automated method to detect online child predators and cyber harassers:

**-Scale:** Manual monitoring is not practicable due to the huge amount of internet material. Effective processing and analysis of large datasets requires automated methods.

**-Speed:** Online threats have the potential to grow quickly. To stop damage, quick notice and action are essential.

**-Complexity:** Analyzing language, photos, and user behavior patterns is frequently necessary to identify predatory or abusive activity. Techniques for data analysis and machine learning can greatly improve.

Thus, the " Detection of Online Child Predators and Cyber Harassers " application that is being presented is an advanced web application that was created with the Django framework. To address the complex issues provided by online dangers, it effortlessly incorporates a number of crucial components, such as machine learning algorithms, content monitoring, user registration, and an admin panel. While administrators have the capacity to keep an eye on, evaluate, and take action against dangerous information and people, users can report suspicious activity they come across. To summarise, the creation and utilisation of tools and systems such as the Django-based application that has been shown here play a crucial role in mitigating the enduring issues presented by cyberbullies and child predators within the modern digital environment.

These technologies aim to protect privacy and security while fostering safer online settings, especially for children and other vulnerable populations, by integrating technology, user interaction, and regulatory compliance. Since the early days of social media, online harassment has been a widespread problem, and it continues to be so. The initial goal of these experiments was to create an automated system that could identify and report this kind of wrongdoing. Two methods—machine learning and deep learning—have been studied to prevent or identify instances of sexual harassment and shield kids from bullying to provide a secure atmosphere. Using fuzzy logic and genetic algorithms, the authors of this study monitored the incidence of cyberbullying on social media platforms. They recognized and categorized offensive, harassing, racist, and terroristic remarks, as well as other cyberbullying-related words and actions on social media. The F-measure obtained was 0.91. A genetic algorithm was employed to achieve optimal performance and parameter optimization. The authors in reference [3] utilized three weighting systems to filter Facebook messages: entropy, modified TF-IDF for feature selection, and term frequency-inverse document frequency (TFIDF). They measured recall, accuracy, and precision using a Support Vector Machine (SVM). The improved TF-IDF scheme outperformed the previous schemes, with an accuracy of 96.50%, according to test findings.

To analyze online harassment on Twitter messages, this study in reference [4] tested several supervised machine learning algorithms. TF-IDF and Word2Vec embeddings were used to extract features. The results accurately covered more than 80% of all the forms of harassment considered in the data.

This study [5] combines a state-of-the-art approach to sentence vectors with emotion analysis. Word vectors are generated using the Long-Short-Term-Memory, Recurrent Neural Network (LSTM_RNN) linguistic pattern as a new approach to identifying sexual predators. With a recall of 81.10%, extracting the value of emotion from the SoftMax layer outputs resulted in a new achievement in accuracy.

The authors in reference [6] used CNN to extract features from tags to predict a classifier for Twitter posts holding malevolent intent. They analyzed a four-month Twitter dataset to find conditions around stories with evil intentions and used this to create laws against gender-based violence. Sweta Karlekar in [7] described the work of the SafeCity Web Community in categorizing and rating various kinds of sexual harassment. SafeCity Web uses this experience from the victim's exchange to develop online directories, provide more comprehensive safety advice services, and help others find relevant cases to stop further sexual assault. The single-label CNN-RNN model achieved 86.5% accuracy in processing, connecting, and annotating tags.Espinoza [8] developed a new dataset from Twitter in four categories of detecting harassment. They used two models of deep learning architecture, CNN and LSTM, to classify the tweets. The F1 measurement was 55% during training, but only 46% of the test set produced F1 findings.Arijit Josh Chowdhury [9] proposed a disclosure language model. The ULMFiT fine-tuning architecture consists of a linguistic model, a task-specific classifier, and a specific mediator, namely Twitter. The overall comparison showed the benefits of choosing specific, lightweight mean language models supported by LSTMs and enhanced vocabulary by gaining knowledge on the linguistic subtleties in deep text describing sexual harassment. About 10,000 personal accounts of sexual harassment were annotated, and the neural network models produced

excellent results in automatic story classification with a 92.9% accuracy rate. Therefore, more advances were made in classification with further consideration of the importance of features.

## SYSTEM ANALYSIS

The process of identifying malicious information on a site involves combining several Python modules with machine learning methods, such as pandas. The first step is looking at a number of postings in order to use statistical analysis to identify any malicious activity. Those whose degree of suspicion rises over a predetermined cutoff are then categorized as suspects. Next, a thorough examination of the postings made by the alleged user is carried out, including any multimedia content like pictures, videos, and audio files. Artificial intelligence is used in conjunction with picture and audio analysis techniques to perform this analysis and determine whether the suspect is a predator. The outcomes of this procedure help identify trends in child grooming. Lastly, information on possible predators is forwarded to law police.

## EXISTING SYSTEM

There are now techniques for locating child predators on the internet in the areas of gaming, voice chat, and other online entertainment. By using these techniques, parents may shield their kids from sexual exploitation whether they play online games or engage in voice chats.

However, with the prevalence of the internet in today's world, a lot of kids are turning to social networking sites as their main way to communicate with others. Because these sites do not have detection systems set for sexual predators, sexual predators therefore put children in danger. Currently, the method used has five algorithms to classify the conversations. It includes the conversation-centered method, which uses the Ridge or Naive Bayes Classifier while processing the TF-IDF feature set, and Neural Network Classifier, which also processes the TF-IDF feature set.This is our proposed system in which a novel approach will be employed for text and picture categorization. This approach will be a regulated machine learning technique known as the Support Vector Machine (SVM) used to solve problems with two-category categorization.

## B. PROPOSED SYSTEM

Our project's goal is to find instances of child harassment on social media by applying a number of machine learning techniques, including K-Nearest Neighbors, Random Forest, Support Vector Machine (SVM), Naive Bayes, and Decision Tree. All the models will be trained by combining phrases and messages that are considered normal with those that are harassing. After it has been trained, the model will be applied to user postings in order to identify if they include harassing or regular material. This project uses the Django framework to construct a web application aimed at identifying child predators and cyberbullies in social media settings. It showcases the backend logic of a web application designed to track and detect these threats. Users can register, log in, post messages, and utilize machine learning techniques to categorize text messages as potentially hazardous or not. The application also provides web pages for users and administrators to view the findings.

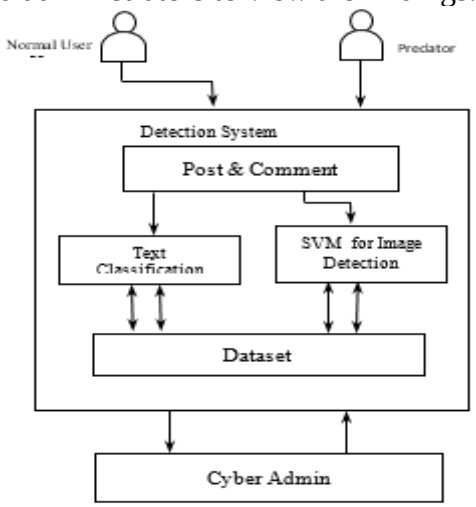


Figure 1: Proposed system architecture

The following are the primary elements and features: Import Statements: The application begins by importing the Django modules and some Python libraries that will be used. These libraries provide web development, data processing, machine learning, and database access utilities.

- Global Variables: A number of global variables are introduced at the outset. These will be utilized for data processing and machine learning, and they include classifier, label_count, X, Y, and corpus.
- Django Views: It defines several Django views, each linked to a unique endpoint in the URL. Index, SendPost, Register, Admin, Login, AddCyberMessages, RunAlgorithms, MonitorPost, AddBullyingWords, Signup, UserLogin, AdminLogin, ViewUsers, ViewUserPost, word_count, prediction, cal_accuracy, and classifyPost are some of the methods that are included in these views. These views manage HTTP requests and provide HTML templates for various application pages.
- Database Access: This utilizes the pymysql package to create a connection to a MySQL database. In order to get and insert data, including user and post information, it communicates with the database.
- Algorithms for machine learning are included in this. A dataset named "dataset.txt" is loaded and preprocessed, text preprocessing is done, and machine learning models (such SVM, Decision Tree, K-Nearest Neighbors, Random Forest, and Naive Bayes) are trained to categorize text data. The classifier variable holds the chosen model for potential usage at a later time.
- Web Forms Handling: AddBullyingWords, Signup, UserLogin, and AdminLogin are some of the functions that handle user-submitted forms and carry out tasks like adding information to the database or confirming user credentials.
- File Upload: This manages the uploading of files, including text files with messages that need to be categorized and pictures from user profiles.
- HTML Templates: To render the user interface, the web application uses HTML templates (such as "index.html," "SendPost.html," "Register.html," and "Admin.html").
- Data processing: This will tokenize words, remove special characters, and convert text to lowercase as part of the preprocessing step.
- Classification: Text messages are classified using machine learning models, and the user is shown the findings.
- Session Management: Upon successful login, the script saves the username in a file called "session.txt" and maintains user sessions.
- Presentation of findings: HTML templates are utilized to offer the user with the categorization findings along with other pertinent data.

**ALGORITHM**

SVM uses a hyperplane to partition the dataset into distinct groups; therefore, it helps determine the largest margin. The main objective of the SVM is to find out that hyperplane of high-dimensional space that will optimally segregate the data points into several classes. The decision boundary is symbolized by the data points on one side of this hyperplane and on the opposite side, the points are representative of another class.

In short, SVM will try to find a hyperplane that maximizes margin, which is the gap between the decision border and the nearest data points on both sides. A maximum margin will improve SVM's ability to generalize, thus his ability to correctly label unknown data. We will feed labeled data to our model to make it learn. In the prediction phase, SVM will match the labeled data with fresh data using the Support Vector Machine.

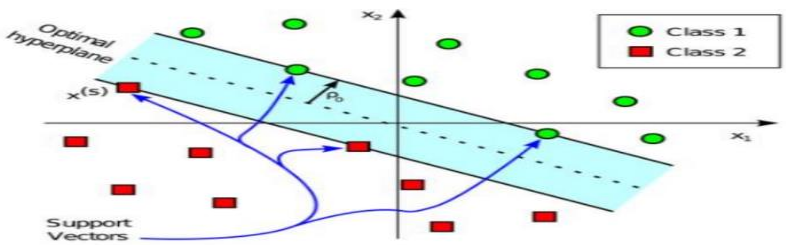



Fig: SVM Algorithm

**DATASET**

The dataset contains two columns: "Tweet" and "Text Label." Here's an explanation of what each column represents:

-**Tweet:** The brief text messages (sometimes known as "tweets") in this section are usually about social networking sites such as Twitter. A user's single tweet is represented by each row in this column. Usually, tweets are only allowed to include a particular number of characters (280, for example).
on Twitter).

-**Text Label:** Each tweet has a label or category listed in this column. The tweets' content may be categorized or classed using these labels. Text labels might signify subjects or themes that are included in the tweet, such "sports," "politics," "entertainment," etc., or they could indicate whether a tweet is favorable, bad, or neutral. Most likely, a label provided to a related tweet in the "Tweet" column correlates with each row in this column.

This dataset is often utilized in machine learning, sentiment analysis, text categorization, and natural language processing (NLP) applications. It enables scholars and analysts to develop and evaluate algorithms that automatically classify or examine twitter content according to the given text labels.

**DATABASE CREATION:**

The first set of SQL queries in this project creates a MySQL database called "cyber" and its two tables, "users" and "posts."

**Table of Users:**

-username (varchar(50)): It is most likely the intention of this column to hold user account usernames. It can be up to 50 characters long at most.

-password (varchar(50)): It is most likely the intention of this column to hold passwords for users. In a production system, passwords ought to be safely hashed and kept, not in plain text.

-contact_no (varchar(12)): It looks that contact numbers for users are kept in this column. It has a maximum character limit of 12. -email (varchar(50)): Email addresses of users are kept in this column. -address (varchar(50)): It appears that user addresses are stored in this column. -status (varchar(30)): It looks that this column is used to store extra information or the status of the user.

**Table for Posts:**

-Sender (varchar(50)): It is most likely the intention of this column to hold the username or identify of the sender.

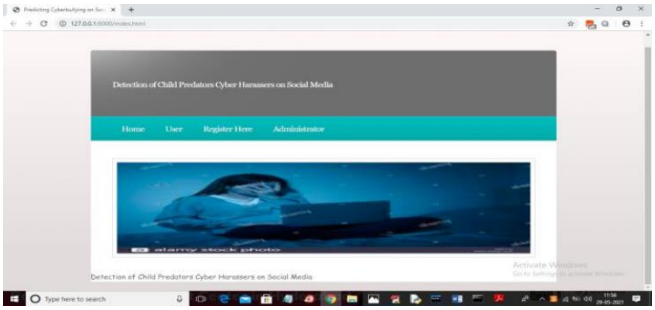
-filename (varchar(50)): The name of a file attached to a post may be stored in this column.

- msg (varchar(300)): It appears that this column is used to store a post's message or content. It has a 300 character limit.

-posttime (timestamp): This column is used to record the post's creation timestamp. When a new post is made, it will automatically log the date and time.

-status (varchar(50)): This column may provide status or other information about posts, much as the "status" field in the "users" table.

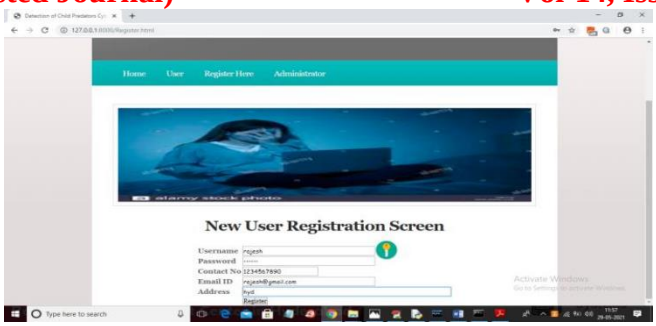
**IMPLEMENTATION AND RESULTS**



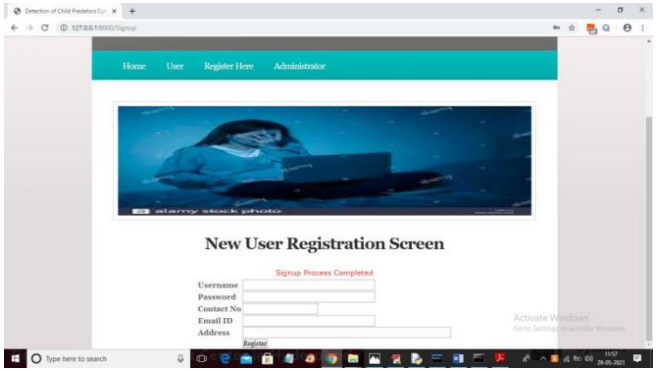

To establish a new user account, navigate to the aforementioned screen and activate the "Register Here" connection.

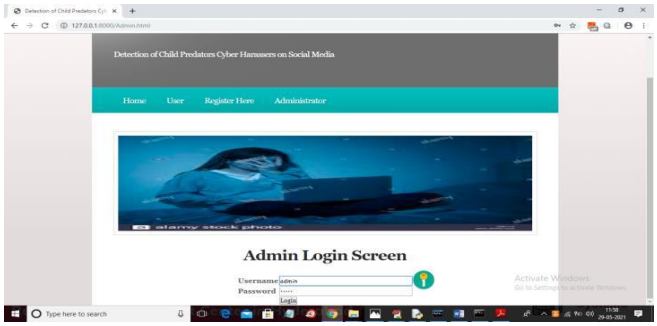
ISSN: 2278-4632
Vol-14, Issue-7, No.02,July: 2024



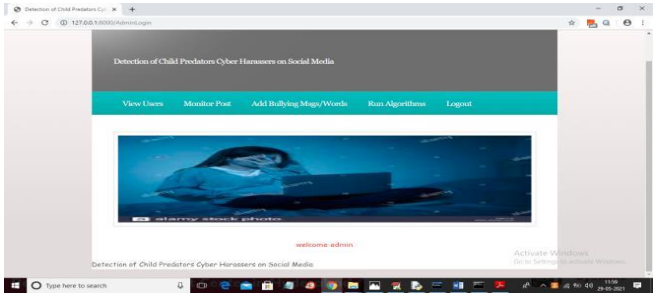

Please proceed to clicking the "Register" button displayed above, in order to input the relevant information.



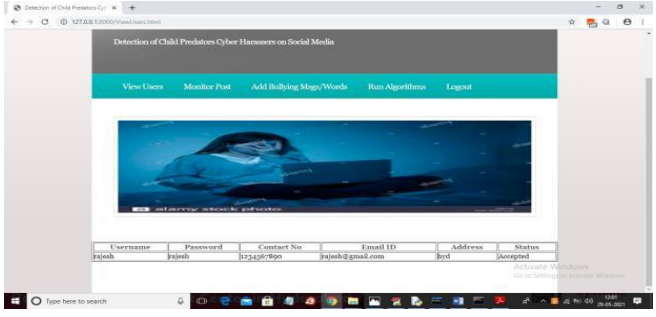

After completing the sign-up process on the aforementioned screen, select the "Administrator" link and log in as the administrator to view the new user details.



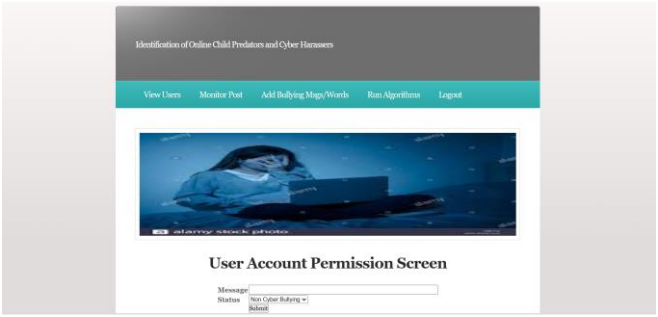

In order to access the below screen, one must log in as the "admin" user on the aforementioned screen by providing "admin" as both the username and password. Upon successful login, the subsequent screen will become accessible.



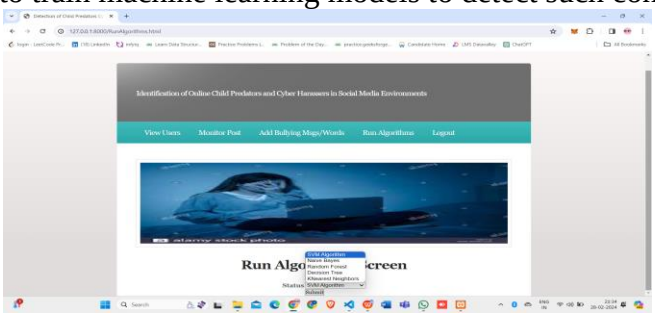

Now, the administrator can click on the 'View Users' link to see the list of all users.



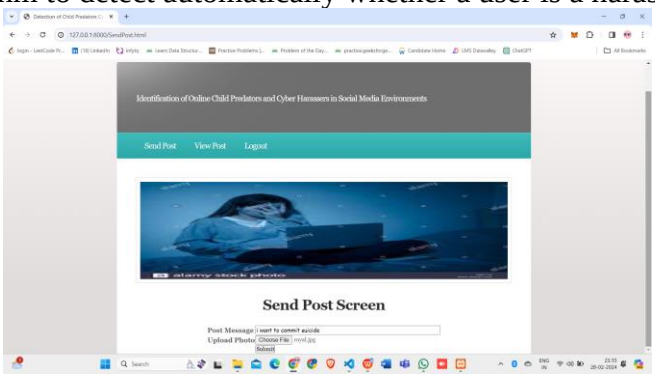
account is demonstrated. The administrator can gain access to a history of posts made by users by clicking the "Monitor Posts" button.




The above figure represents a page or interface where administrators can add specific words or messages to a dataset. These words are typically considered bullying or related to cyber harassment. This dataset is then used to train machine learning models to detect such content.

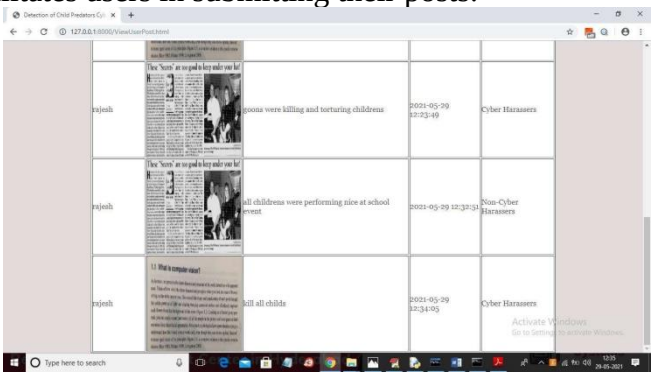



In the above Figure, admin needs to select each algorithm one after one then press the 'Submit' button in order to train the model and the accuracy of each algorithm is shown. Admin has to run these steps again every time he reboots the server or every time new bullying messages are added. He needs to run at least one algorithm to detect automatically whether a user is a harasser or not.




The "Send post" page or module in the user interface is shown in Figure 10.12. Users can write and send a message with a photo on this page. Important components of this page are:

- Text input field: This field allows users to type messages.
- Uploading a file allows users to include a picture or photo in their post.
- The Send button facilitates users in submitting their posts.



The view post page allows users to see all messages with uploaded photos posted by others. As shown in the figure, the proposed system can identify whether a message is from a cyber harasser or a non-cyber harasser using machine learning. Machine learning models are employed to predict whether a user is a harasser based on dataset records. Additionally, through the "add words" module, the admin can update the dataset by adding terms that are identified as potential harasser terms or those that are not.
Following the addition of terms, algorithms are linked to train the model, and the suggested application then automatically forecasts whether the person is harassing or not.

**Conclusion**

This study presents the backend logic of a web application focused on detecting cyberbullying and managing users, offering a comprehensive approach to addressing cyberbullying-related issues and enhancing user experiences. A key feature of the application is its robust user management capabilities, allowing users to check their profiles, log in, and sign up. User data, including usernames, passwords (which should be further protected using techniques like hashing and salting), contact details, and status, is securely stored in a MySQL database. This feature forms the foundation for both user experience and the administration of user interactions.
One of the application's most notable features is its ability to detect cyberbullying. The program extracts relevant characteristics from user-submitted text messages, storing this data in a dataset for training and inference. Users can submit posts with sender names, messages, timestamps, statuses, and filenames for attached photos.
A thorough record of user interactions is formed by the careful recording of these posts in a database. The program provides a set of user-friendly web pages for user interaction. Users have access to register, log in, see profiles, write posts, run machine learning algorithms for detection, add phrases linked to cyberbullying to the dataset, and monitor postings. User involvement is made easier by this user interface's accessibility and intuitiveness.

Online begging poses hazards, but the costs of sexual exploitation of children and society are too great to ignore. Child groomers usually pose as kids with similar interests and hobbies in order to build relationships with children and obtain access to them. The goal is to establish a trusting relationship with the youngster. For the purpose of protecting children, our initiative looks for these predators and, if it finds any, notifies the cyber administrative authorities right away so that the proper action may be taken.
Examining questionable information on a platform entails the following sequential steps:

- Getting information from the postings made by the suspected person, including multimedia content like pictures, music, and videos.
- analysis of the acquired data using the NSFW library, artificial intelligence, Urllib, and the IGPL Python module.
- Identifying the suspect as a predator or a suspect, depending on the results of the investigation.
- Analyzing kid grooming practices and statistical data to identify the person as a predator.
- automatic transmission of the predator categorization to a server-stored Gmail account.

**Future Work**

Even though the current application is already very good, there is still a lot of space for improvement and growth.
Some of these include improved user authentication and ongoing machine learning model optimization for Cyber bullying detection through the investigation of various algorithms, feature engineering approaches, and hyper parameter tuning.

**References**

[1] Amer, N. Arabic-sexual-harassment-dataset. Available from https:// github.com/Nooramer8/Arabic-sexual-harassment-dataset. [Accessed 09-10- 2023].

[2] Nandhini BS, Sheeba J. Online social network bullying detection using intelligence techniques. Proc Comput Sci 2015;45:485–92. doi: https://doi. org/10.1016/j.procs.2015.03.085.
[3] Al-Katheri ASA, Siraj MM. Classification of sexual harassment on Facebook using term weighting schemes. Internat J Innov Comput 2018;8(1):15–9. doi: https://doi.org/10.11113/ijic.v8n1.157.
[4] M.Saeidi, S.Sousa, E.Milios, N.Zeh, L.Berton. Categorizing online harassment on Twitter. in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2019, 3, 283-297. https://doi.org/10.1007/978-3-030 43887-6_22.
[5] Liu, D., C.Y. Suen, and O. Ormandjieva. A novel way of identifying cyber predators. 2017, 1712.03903,1-6. https://doi.org/10.48550/arXiv.1712.03903
[6] Pandey, R., et al. Distributional semantics approach to detect intent in Twitter conversations on sexual assaults. in 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). 2018, 1 270-277. https://doi.org/ 10.1109/wi.2018.00-80.
[7] S.Karlekar, and M. Bansal.. Safecity: Understanding diverse forms of sexual harassment personal stories, arXiv preprint arXiv. 2018, 2,1-7. https://doi.org/ 10.18653/v1/d18-1303.
[8] Espinoza I, Weiss F. Detection of harassment on Twitter with deep learning techniques. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases 2019;1168:307–13. doi: https://doi.org/10.1007/978- 3-030-43887-6.
[9] Liu Y et al. Sexual harassment story classification and key information identification. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. p. 2385–8. https://doi.org/10.1145/ 3357384.3358146.
C. H. Ngejane, G. Mabuza-Hocquet, J. H. P. Eloff, and S. Lefophane, “Mitigating online sexual grooming cybercrime on social media using machine learning: A desktop survey,” in 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Aug 2018, pp. 1–6.