Identification of Online Child Predators and Cyber Harassers in Social Media Environments

Given Name Surname   
*dept. name of organization   
(of Affiliation)*  
*name of organization   
(of Affiliation)*City, Country  
email address or ORCID

*Abstract*—Social media platforms have become an essential aspect of our lives in the current digital era, linking individuals from throughout the globe. Although these platforms have been very beneficial, they have also exposed children and other vulnerable people to a variety of online threats.These dangers include cyberbullies and child predators who prey on children online and use social media's anonymity and reach to hurt other people. In the past, manual reporting and human moderators were mostly depended upon to counter these dangers. After users reported suspicious activity, human moderators looked into the content to see whether it broke any site rules. But this reactive strategy frequently caused delays, allowing dangerous information to proliferate before appropriate action could be done. Researchers have resorted to machine learning, a subfield of artificial intelligence that enables computers to learn from data and make predictions, in recognition of the need for more proactive and effective solutions. The objective is to create an automated system that can quickly and effectively identify possible online child predators and cyber harassers by utilizing machine learning techniques. The suggested machine learning-based strategy has a number of benefits over conventional techniques. The main benefit is that it speeds up platform reaction times considerably, allowing for the quick removal of offensive content and users. Seeing the need for more proactive and efficient solutions, researchers have turned to machine learning, a branch of artificial intelligence that allows computers to learn from data and make predictions. The goal is to use machine learning techniques to develop an automated system that can rapidly and efficiently identify potential online child predators and cyber harassers. Compared to traditional methods, the machine learning-based approach that has been proposed has several advantages. The primary advantage is that it significantly accelerates platform response times, enabling the prompt removal of objectionable information 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 anonymity and reach of the internet for malicious reasons, 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, 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 "Identification 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.

# LITERATURE REVIEW

Since the early days of social media, online harassment has been a widespread problem, and it still is. 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 in order to provide a secure atmosphere. Using fuzzy logic and genetic algorithms, the authors of this study [4] 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 that was obtained was 0.91. To get the right performance and optimize parameters, a genetic algorithm is employed.

The authors in ref [5] employed three weighting approaches for Facebook message filtering: entropy, term frequency-inverse document frequency (TFIDF), and modified TF-IDF for feature selection. A Support Vector Machine (SVM) was employed to assess a support vector's recall, accuracy, and precision. According to the test findings, the modified TF-IDF works better than the other schemes, with an accuracy of 96.50%. This study in [6] compared supervised machine learning (ML) algorithms for the evaluation of online harassment in Twitter messages as a component of social media competitiveness and harassment (a feature) and for natural language learning. With the use of Word2Vec embeddings and TF-IDF, features were extracted.

The outcomes accurately encompassed more than 80% of all the forms of harassment that were taken into account in the data. This study [7] integrates a contemporary method of sentencing vectors with emotion analysis. Word vectors are created using the Long-Short-Term-Memory, Recurrent Neural Network (LSTM\_RNN) linguistic pattern as a novel method of predators' sexual identification. With a recall of 81.10%, the last step of extracting the emotion value from the SoftMax layer outputs has produced a record-breaking accuracy rate.  
 Convolution neural networks (CNN) are used by the authors in ref. [8] to extract features from tags and build a model for classifying Twitter posts based on malevolent intent. In order to investigate the story circumstances that conveyed malevolent intent, they examined a four-month Twitter dataset. They discussed the importance of these incidents in creating laws against gender-based violence. The work of the SafeCity Web Community to classify and evaluate various forms of sexual harassment is presented by Sweta Karlekar in [9]. By exchanging victim experiences, SafeCity Web utilizes this data to assist victims in creating online directories, offer more in-depth safety advise services, and assist others in finding pertinent instances to stop more sexual assault.

The accuracy of the single-label CNN-RNN model in processing, connecting, and annotating tags is 86.5%. Espinoza [10] utilizes Twitter to create a fresh data set that includes four categories for detecting harassment. They used CNN and LSTM, two different deep learning architectures, to categorize the tweets. The measurement for F1 was equivalent to 55 percent when the data was being trained, however only 46 percent of the test set produced findings for F1.   
Arijit Josh Chowdhury [11] suggests a disclosure language model. A linguistic model, a task-specific classifier, and a particular mediator (Twitter) make up the ULMFiT fine-tuning architecture. The entire comparison demonstrates the advantages of adopting specific, lightweight mean language models based on LSTMs, as well as an enhanced vocabulary that takes into account linguistic subtleties found in the deep text that addresses sexual harassment.

About 10,000 personal accounts of sexual harassment were annotated to extract key features and automatically classify the stories using the neural network models that produced excellent results. With a 92.9% accuracy rate, more advances in classification were made by examining the specifics of important characteristics.

# 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.

Children are therefore at danger of damage from sexual predators on these platforms due to the lack of a dedicated detection system.The current technique employs five classification algorithms: the conversation-centered approach uses the Ridge or Naive Bayes Classifier, which works on the TF-IDF feature set, and the Neural Network Classifier, which functions on the TF-IDF feature set.

In order to improve accuracy over current systems, our proposed system will use a novel approach for text and picture categorization. This approach will be a regulated machine learning technique called the Support Vector Machine (SVM), which is used to solve problems with two-category categorization.

## 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. 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. The software seems to have something to do with identifying child predators and cyberbullies in social media settings. This project also shows the backend logic of a web application intended to track and detect child predators and cyberbullies on social media. Registering, logging in, posting, and using machine learning techniques to categorize text messages as potentially hazardous or not are all options available to users. Web pages displaying the findings are available for users and administrators to view.

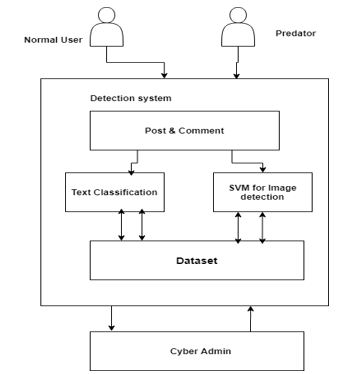


Figure 1: Proposed system architecture

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

To determine the largest margin, SVM uses a hyperplane to partition a dataset into distinct groups. The fundamental objective of SVM is to find the hyperplane in high-dimensional space that optimally separates the data points into different classes. The decision boundary is represented by data points on one side of this hyperplane, whereas points on the other side are representative of a different class.   
 The goal of SVM is to find the hyperplane that maximizes margin, which is the distance between the decision border and the nearest data points for each class. By maximizing margin, SVM aims to improve generalization and its ability to correctly categorize unknown data. Labeled data will be fed into our model to train it. Using the Support Vector Machine (SVM), labeled data will be matched with fresh data during the prediction phase.

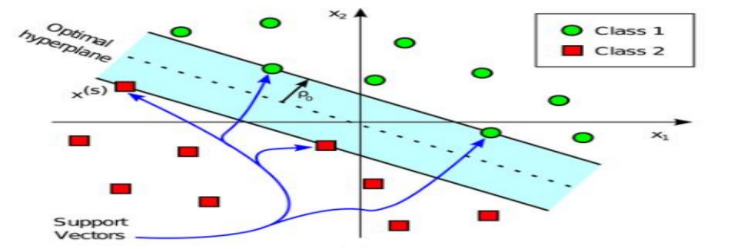


Fig:SVM Algorithm

# 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."

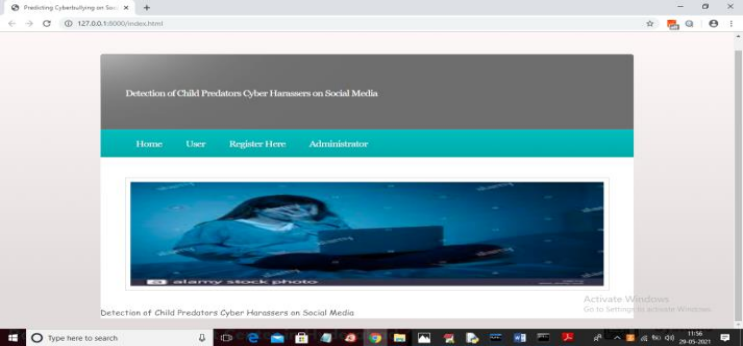
**Table of Users:**   
-username (varchar(50)): It is most likely the intention of this column to hold user account usernames. It can be up to 50 characters long at most.   
-password (varchar(50)): It is most likely the intention of this column to hold passwords for users. In a production system, passwords 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.

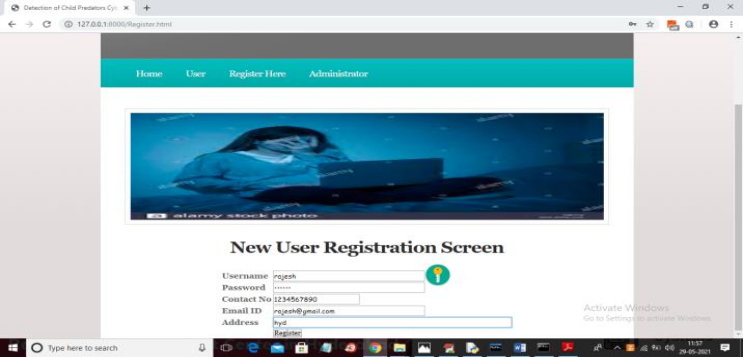
# IMPLEMENTATION AND RESULTS



To establish a new user account, navigate to the

aforementioned screen and activate the "Register

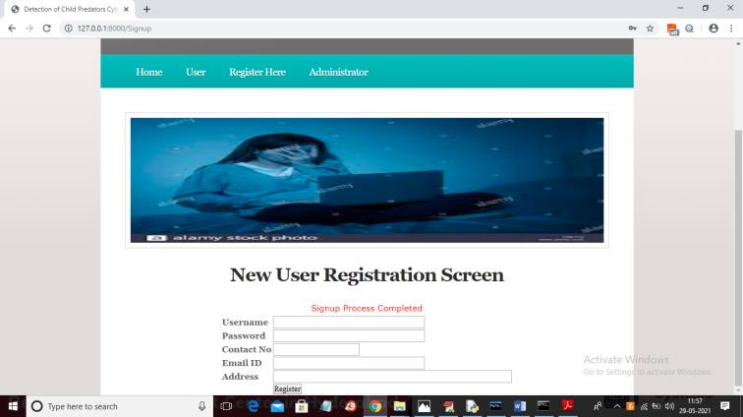
Here" connection.



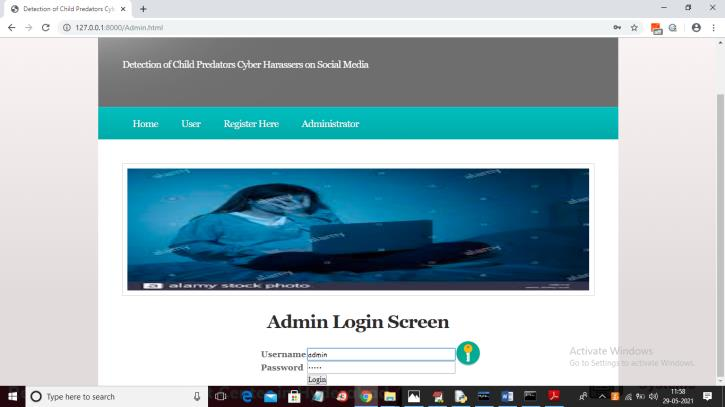
Please proceed to clicking the "Register" button

displayed above, in order to input the relevant

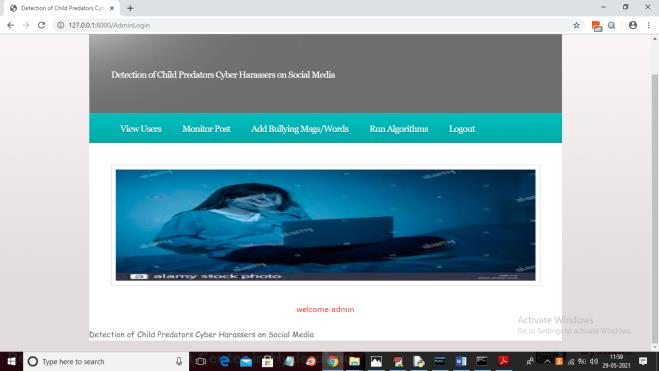
information.



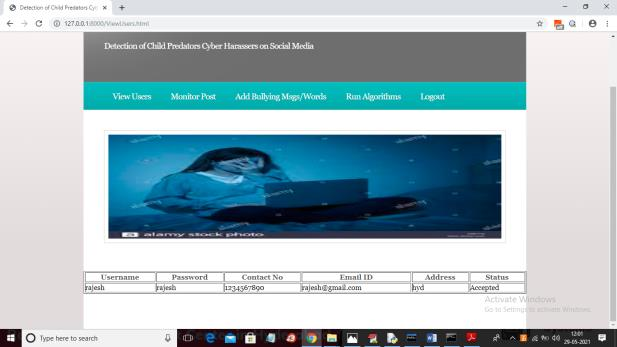
After completing the sign-up process on the aforementioned screen, select the "Administrator" link and log in as the administrator to view the new user details.



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.

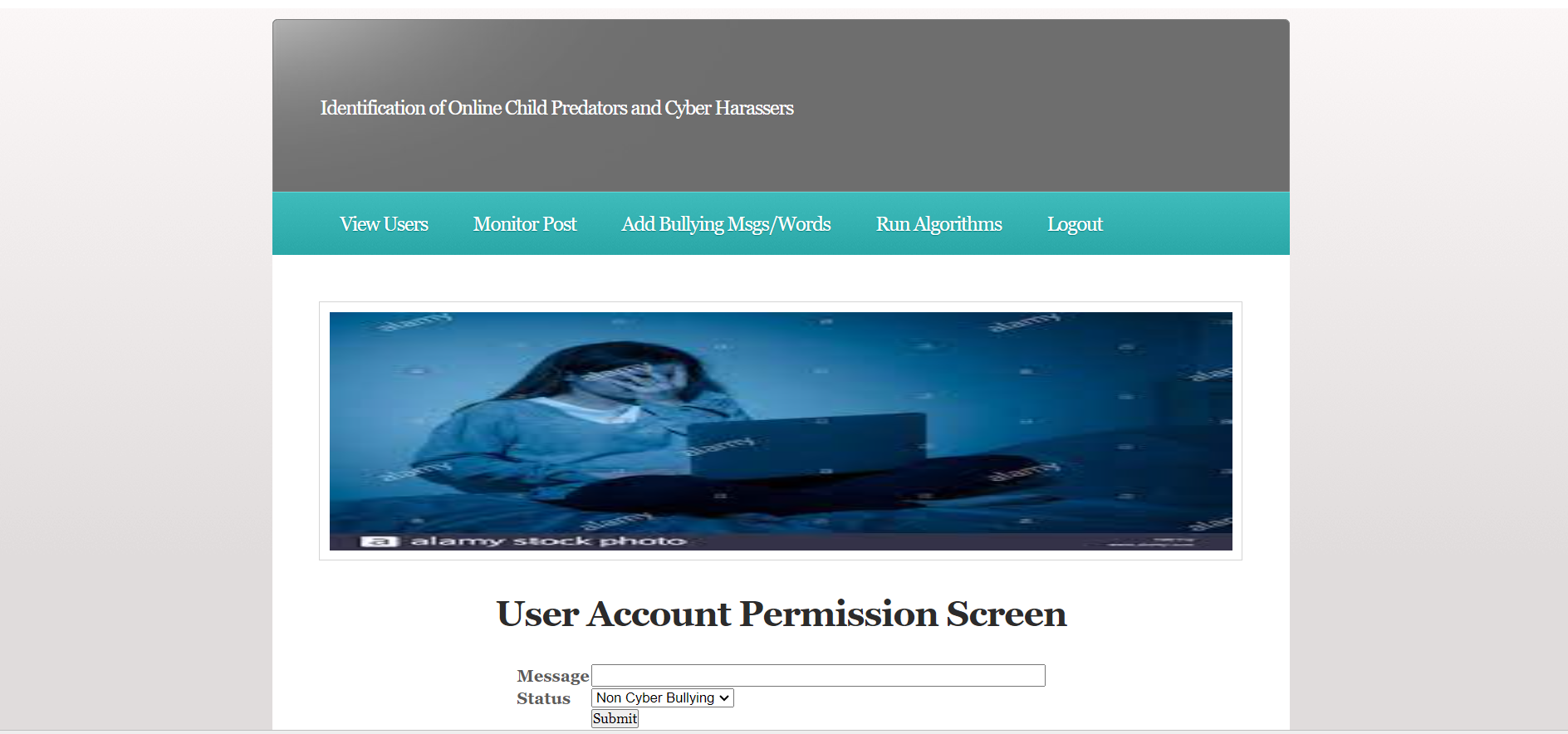


Now admin can click on ‘View Users’ link to view all users list

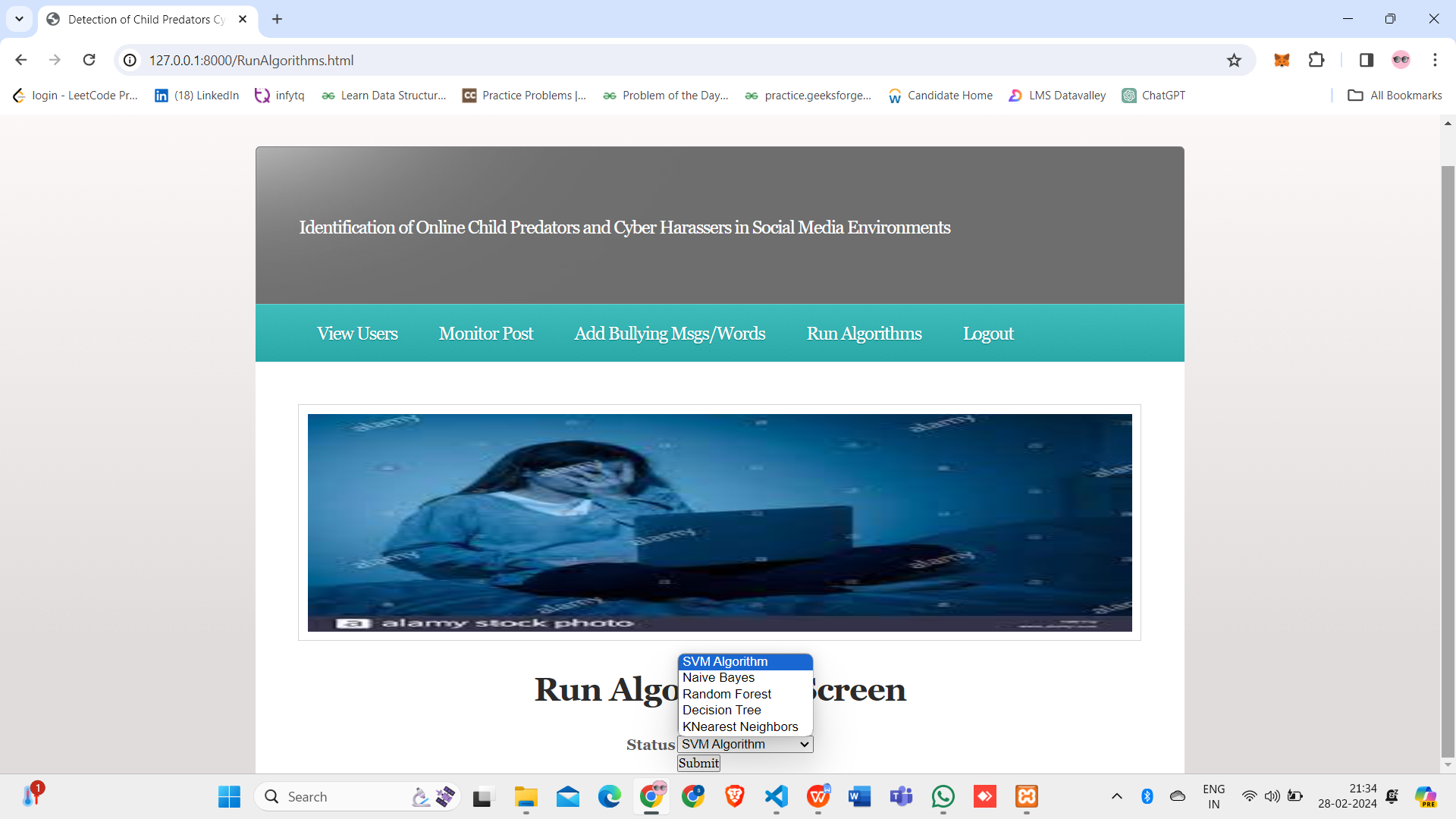


In the screen above, the creation of the "Rajesh"

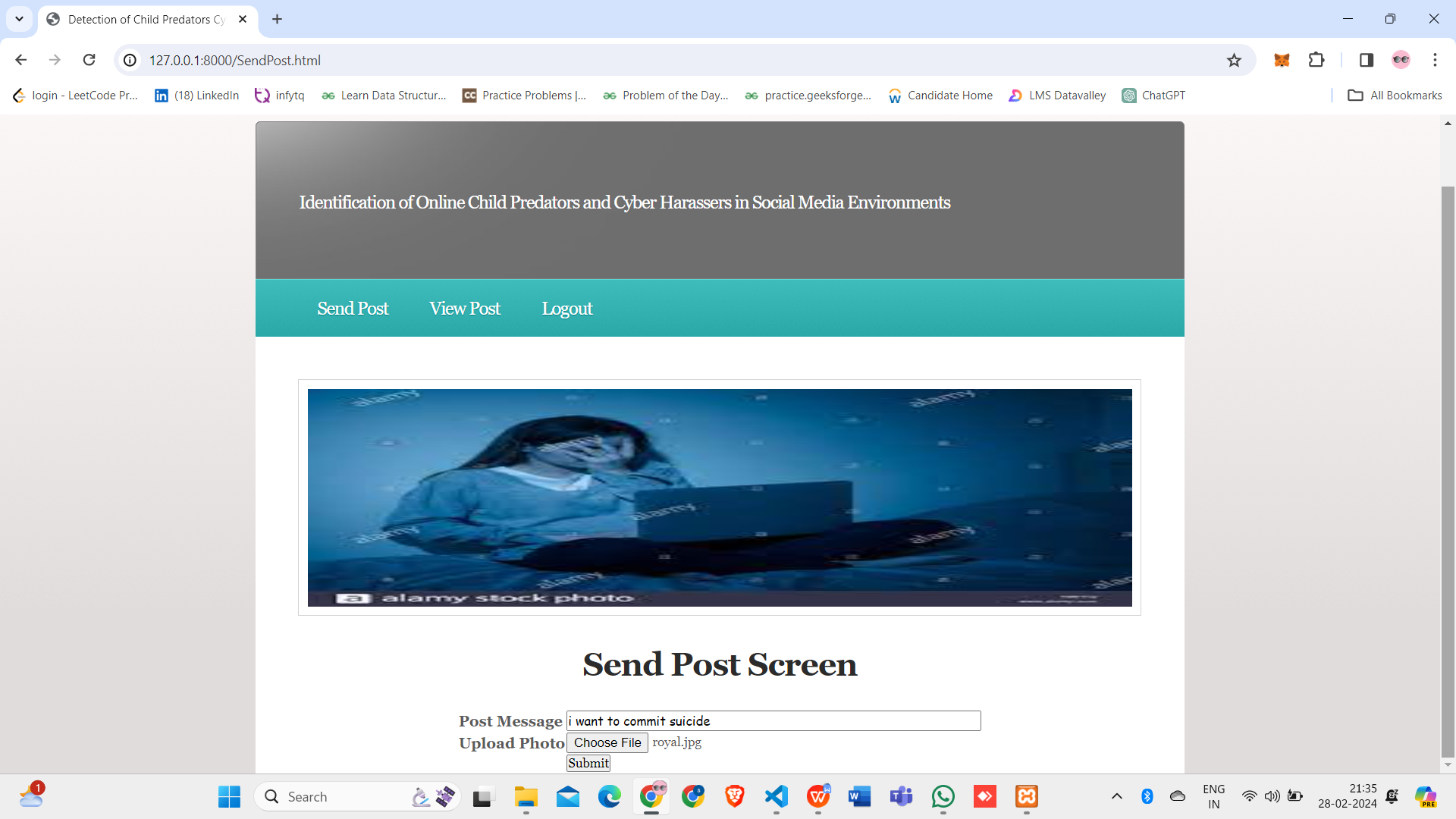
account is demonstrated. The administrator can gain access to a history of posts made by users by clicking the "Monitor Posts" button.



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 used for training machine learning models to detect such content.

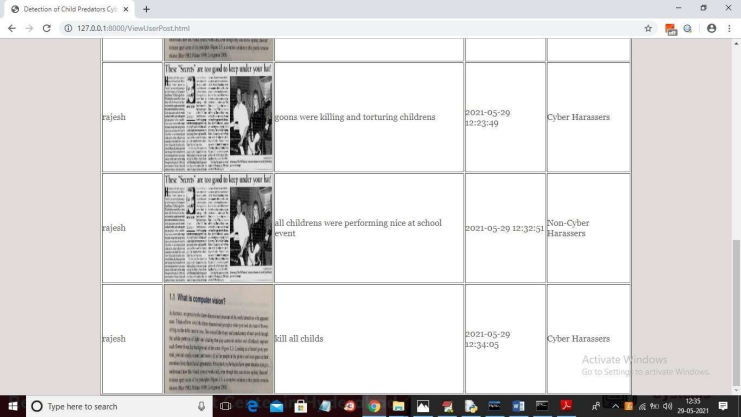


In Above Figure admin must select each algorithm and click on ‘Submit’ button to train model and the accuracy will be shown for each algorithm. Admin must repeat this step whenever first time he starts the server or upon adding new bulling messages. He must run at least one algorithm to perform automatic detection of harasser’s or non-harassers.



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.



View post page for viewing all the messages with uploaded photos posted by users. All user postings are displayed in the Figure , and it is noted that the suggested system can identify if a message is from a cyber or non-cyber harasser by using machine learning. Based on the dataset records, machine learning models are used here to forecast the harasser or non-harasser term. Thus, using the admin's "add words" module, as previously said, admin may add all potential harasser and non-harasser terms to the dataset.   
 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

The backend logic of a web application with an emphasis on cyberbullying detection and user management is presented in this study. It represents a thorough strategy for addressing cyberbullying-related problems and improving user experiences. The application's main selling point is its powerful user management features. Users are able to check their profiles, log in, and signup. A MySQL database is used to carefully store user data, such as usernames, passwords (which should be further protected using methods like hashing and salting), contact details, and status. This feature serves as the cornerstone for both the user experience and the administration of user interactions. The application's ability to identify cyberbullying is one of its most notable features.From these text messages, the program extracts pertinent characteristics, which it then keeps in a dataset for training and inference. Users may submit posts with sender names, messages, timestamps, statuses, and filenames (for attached photos) using this application.   
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. Wachs S, Wolf KD, Pan C. Cyber grooming: Risk factors, coping strategies and associations with cyberbullying. Psicothema 2012;24(4):628–33.
2. [2] KELLER, N.B.a.M.H. video games and online chats are ‘‘hunting grounds” for sexual predators. Available from https://www.nytimes.com/interactive/2019/ 12/07/us/video-games-child-sexabuse.html [Accessed DEC. 7, 2019]
3. Amer, N. Arabic-sexual-harassment-dataset. Available from https:// github.com/Nooramer8/Arabic-sexual-harassment-dataset. [Accessed 09-10- 2023].
4. 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.
5. 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.
6. 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.
7. 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
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. N. Pendar, “Toward spotting the pedophile telling victim from predator in text chats,” in International Conference on Semantic Computing (ICSC 2007), Sep. 2007, pp. 235–241.
14. I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski, “Learning to identify internet sexual predation,” International Journal of Electronic Commerce, vol. 15, no. 3, pp. 103– 122, 2011.
15. G. Inches and F. Crestani, “Overview of the international sexual predator identification competition at PAN-2012,” in CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012, 2012.
16. E. Villatoro-Tello, A. Juarez-Gonz ´ alez, H. J. Escalante, M. Montes-y- ´ Gomez, and L. V. Pineda, “A two-step approach for effective detection ´ of misbehaving users in chats,” in CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012, 2012.
17. G. Eriksson and J. Karlgren, “Features for modelling characteristics of conversations,” in CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012, 2012
18. Muhammad Ali Fauzi, Patric Bours, “Ensemble Method for Sexual Predator Identification”. IEEE, 2020,25 June 2020
19. M. C. Seto, "Pedophilia and sexual offenses against children", Annual Review of Sex Research, vol. 15, no. 1, pp. 321-361, 2004.