iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67633

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Automated Cyber Harassment Surveillance: Enhancing Social Media Safety

D V Divakar Rao(Ph.D.)[1],Shravya Shri Sahu[2], Kilaparthi Durga Mallesh[3], Vedulla Naga Venkata Nitin[4], Bandaru Avinash[5]
Raghu Engineering College (REC), Department of Computer Science (Cybersecurity) Dakamarri, Bheemunipatnam, Visakhapatnam – 531162, India

Abstract: ***The increase in the use of social media platforms has coincided with an increase in cyber harassment and child predation, leading to significant safety issues. This research presents an AI-driven system aimed at detecting cyber harassers and potential child predators using machine learning (ML) and natural language processing (NLP). The system utilizes a Random Forest Classifier in conjunction with TF-IDF vectorization to classify user messages into various categories, such as hate speech, offensive content, and neutral interactions. Unlike conventional methods, which depend largely on manual oversight and basic keyword filtering, this approach improves accuracy, contextual understanding, and automated responses. The system is implemented with a flask-based backend and a react-powered front-end, facilitating real-time content analysis, automatic threat detection, and an administrative dashboard for moderation. The experimental findings highlight the system's capability to identify harmful online behaviors, providing a proactive solution to ensure a safer social media environment, especially for at-risk users.***

Keywords: Online safety, cyber harassment detection, AI-driven moderation, machine learning, natural language processing, TF-IDF, Random Forest Classifier, social media content filtering.

I. INTRODUCTION

The swift expansion of social media has transformed communication, enhancing global connectivity, information sharing, and personal expression. Yet, this digital shift has also introduced several challenges, notably the rise in cyber harassment and predatory behavior. Online abuse, such as cyberbullying, hate speech, digital stalking, and intimidation, has become prevalent, with child predators exploiting these platforms to target susceptible users [1][4]. The anonymity and extensive reach of social media allow these malicious actors to operate with little risk, underscoring the need for intelligent, automated solutions to improve user safety. Traditional moderation methods, like user reporting, keyword filters, and basic machine learning models, have notable limitations. Keyword-based systems often misinterpret neutral content as offensive and fail to detect concealed threats [6]. Manual moderation is slow and inefficient, making it challenging to identify predators who use gradual grooming techniques or cyber harassers who employ coded language and slang to avoid detection [5][12]. These shortcomings emphasize the necessity for a proactive AI-powered system capable of real-time detection, classification, and prevention of online threats. This paper introduces an AI-driven system that identifies cyber harassers and online predators using Machine Learning (ML) and Natural Language Processing (NLP) techniques. The system utilizes a Random Forest Classifier, a robust ML algorithm renowned for its accuracy in text classification, to categorize online content as hate speech, offensive language, or neutral text [7][8]. Additionally, TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is used to extract significant textual features, enabling the system to detect hidden patterns in user interactions [3][8]. Unlike traditional moderation tools, this approach incorporates real-time monitoring, automated threat detection, and an interactive dashboard for moderators and law enforcement officials, allowing them to review flagged content and take necessary action promptly [10]. Detecting cyber harassment and child predation is challenging because of the subtle and evolving nature of online communication. Predators often begin interactions in seemingly innocuous ways, gradually escalating their behavior, making them difficult to detect with conventional tools [11]. Similarly, cyber-harassers frequently use emojis, abbreviations, or alternate spellings to disguise their messages [13]. Existing moderation tools lack context awareness, making it difficult to detect deceptive language patterns and complex behaviors [14]. To address these challenges, our system integrates multiple advanced technologies to enhance detection accuracy and efficiency. The core of the system relies on a Machine Learning Model, specifically a Random Forest Classifier, which is widely recognized for effectively distinguishing harmful from nonharmful content [1].

Furthermore, Natural Language Processing (NLP) techniques, such as TF-IDF vectorization, provide a deeper understanding of textual content by analyzing word frequency and importance, allowing for the identification of disguised harassment and covert predatory behavior [3]. The backend was built using Flask, a lightweight web framework that enables real-time interaction with a pre-trained machine learning model, ensuring efficient content analysis [15]. On the front end, React is utilized to create an intuitive and dynamic dashboard for moderators, enabling them to examine flagged content and take quick action when necessary.In contrast to conventional moderation systems that rely on user feedback, this AI-driven system provides a proactive approach that detects and mitigates threats before they intensify. In addition, its adaptability and scalability make it ideal for implementation across multiple social media platforms, offering broader protection for online communities.

The remainder of this paper is organized as follows. Section II examines existing studies on cyber harassment detection, highlighting the shortcomings and inefficiencies of the current methods. Section III outlines the methodology, including data preprocessing, feature extraction, and model training. Section IV describes the experimental setup and evaluation results, and compares the system's accuracy and efficiency with traditional moderation methods. Section V discusses the major challenges, ethical considerations, and potential future developments. Finally, Section VI summarizes the main findings and proposes directions for future research. This study aimed to foster a safer digital environment by combining advanced AI techniques with real-time monitoring. The proposed system provides an automated, scalable, and highly precise solution for identifying cyber harassment and online predatory behavior, ensuring better protection for vulnerable users, and contributing to the creation of a more secure online space.

II. RELATED WORK

The issue of identifying cyber harassment and child predation on social media has been extensively studied, especially within the domains of machine learning (ML), natural language processing (NLP), and artificial intelligence (AI). Researchers have investigated a variety of methods, from manual moderation and keyword filtering to advanced AI-based detection systems. Nonetheless, these existing methods often face challenges such as misinterpreting context, high rates of false positives, and delayed response times. To address these issues, a more sophisticated and adaptable approach is necessary—one that combines ML, NLP, and real-time threat detection capabilities.

Traditional moderation methods utilize machine learning classifiers to determine whether messages are harmful or benign. Several classification models have been extensively researched, including Support Vector Machines (SVM), Random Forest, Logistic Regression, and Naïve Bayes. Dadvar and Eckert (2020) examined SVM models enhanced with deep learning for cyberbullying detection. Their results showed improved accuracy but also pointed out the model's difficulty in understanding contextual subtleties [7]. Chatterjee et al. (2019) highlighted the Random Forest Classifier's effectiveness in handling large datasets. Although this model increases accuracy by integrating multiple decision trees, its computational demands impede real-time detection [6]. Meanwhile, Logistic Regression and Naïve Bayes, as employed by Dadvar et al. (2013), have been effective in binary classification tasks but face challenges with imbalanced datasets and complex conversational structures [8].

Progress in Natural Language Processing (NLP) has led researchers to investigate more efficient feature extraction methods, such as Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings, to enhance content classification. Avila et al. (2013) demonstrated how TF-IDF improves harmful language detection by isolating key words and minimizing noise from common phrases [3]. significant advancement in NLP was the introduction of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018), which significantly enhanced context-based analysis, making it a valuable tool for detecting cyber harassment [9].

Despite these advancements, traditional ML models still encounter difficulties in identifying hidden forms of harassment. Bogdanova et al. (2014) discovered that cyber harassers and predators often use coded language, sarcasm, or ambiguous expressions to avoid detection, complicating the task for classifiers to accurately identify harmful content [4]. Cano et al. (2014) further emphasized how child predators gradually gain trust by imitating a child's language, highlighting the need for detection models to analyze long-term linguistic and behavioral patterns [5].

To tackle these challenges, researchers have explored deep learning models for detecting cyber harassment. AlDahoul et al. (2020) demonstrated the effectiveness of YOLO-based CNN models in identifying adult content in images, showcasing their accuracy in detecting inappropriate visual material [1]. Similarly, Ebrahimi et al. (2016) applied deep convolutional neural networks to chat logs, incorporating semi-supervised anomaly detection models to improve predator identification [10][11].

A crucial area of research is the real-time detection of cyber harassment. Guglani and Mishra (2021) proposed deep neural network (DNN)-based speech recognition models to identify abusive speech, particularly in regional languages, thereby extending detection capabilities beyond just English [15].

III. METHODOLOGY

A. Research Design

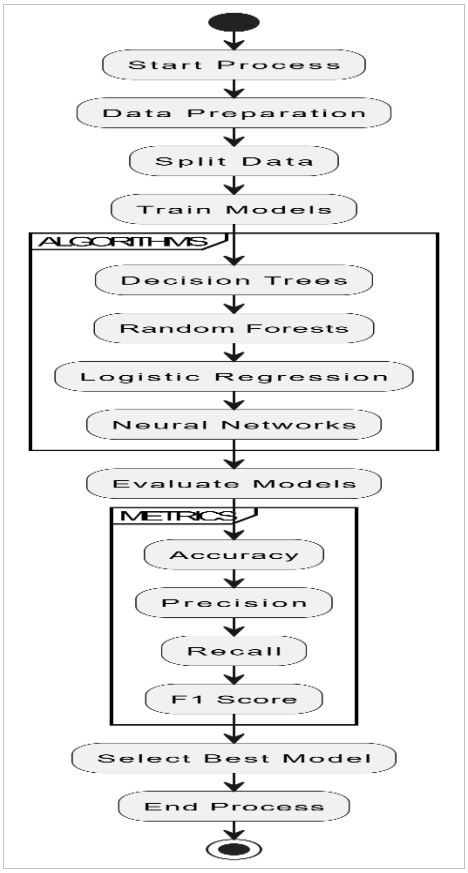


Figure 1: Machine Learning of Selection Workflow

This project, titled "Detection of Child Predators and Cyber Harassers on Social Media," employs machine learning and natural language processing (NLP) to scrutinize online interactions. The primary aim is to detect harmful activities, such as child grooming and cyber harassment, in real time, enabling platform moderators or relevant authorities to take swift action. The study employed a systematic methodology comprising essential steps: data collection, preprocessing, model development, system design, evaluation, and testing. This process is meticulously crafted to ensure high accuracy, while strictly upholding ethical standards that protect user privacy and data security.

B. Data Collection

Data collection is fundamental to this research because a well-organized dataset is vital for training the detection system. The main data sources are as follows:

User Interactions: Textual data from social media platforms, encompassing posts, comments, direct messages, and general conversations, were used to examine communication patterns.

Labeled Data: Pre-annotated datasets with examples of abusive language, cyber harassment, and grooming behavior are essential for training supervised learning models.

Synthetic Data Generation: Artificially generated data that mimic real-world interactions, enhancing the adaptability and robustness of the model.

Data collection was carried out using authorized web scraping tools, API integrations, and partnerships with research institutions that supply labeled datasets. Ethical considerations, such as user privacy protection, anonymization, and adherence to global regulations, such as the GDPR, are rigorously followed.

Category	Description	Example
Hate Speech	Explicit abuse, threats, or offensive language	*"You are worthless! Get lost!"*
Cyber Harassment	Bullying, stalking, derogatory remarks	*"I will keep messaging you until you respond."*
Predatory Conversations	Grooming, coercion, manipulative behaviour	*"You're special. Don't tell anyone about us."*
Neutral Content	Safe, non-offensive user interactions	*"Let's meet at the library tomorrow."*

Table 1: Content Moderation Categories

C. Preprocessing and Data Preparation

Before the collected data are input into the machine learning models, they undergo comprehensive preprocessing to enhance quality and effectiveness. The main steps include:

Text Cleaning: Eliminating unnecessary elements such as punctuation, special characters, numbers, and stop words to refine textual input.

Tokenization: Dividing text into smaller units, such as words or phrases, for improved analysis.

Vectorization: Textual data are transformed into numerical form using methods like TF-IDF (Term Frequency-Inverse Document Frequency (TF-IDF) or word embeddings (Word2Vec, GloVe, BERT).

Feature Extraction: Identifying key linguistic and behavioral patterns associated with predatory behavior, such as excessive compliments, persistent requests for personal information, and manipulative conversations.

Normalization: Standardizing data formats for enhanced consistency and performance.

D. Model Development

The central element of the system is its machine-learning model, which is essential for identifying and categorizing user messages as neutral, harassing, or abusive. To determine the most suitable model, a range of machine learning and deep learning methods was evaluated to ensure that they met the standards of efficiency, scalability, and high accuracy in text classification. Random Forest Classifier – This approach improves classification by combining multiple decision trees, which helps minimize overfitting and enhances accuracy, particularly with large datasets. By simultaneously analysing multiple features, it effectively differentiates between normal and offensive languages. Support Vector Machine (SVM) – Known for its strong performance in both binary and multiclass classification, SVM works by mapping data into a high-dimensional space, ensuring a clear separation between harmful and non-harmful content. Its adaptability to various kernel functions allows it to effectively handle diverse language patterns. Logistic Regression – A lightweight yet powerful technique, logistic regression excels at distinguishing between two categories, such as offensive and non-offensive messages. When used with TF-IDF vectorization, it highlights words that significantly influence the classification process. Naïve Bayes – As a probabilistic model frequently used in text analysis, this method assumes word independence, making it computationally efficient and fast for large-scale text classification. It is particularly effective in filtering abusive messages and spams. Long Short-Term Memory (LSTM) networks – Unlike traditional machine learning methods, LSTM, a type of recurrent neural network (RNN), examines word sequences over time. This ability makes it highly effective for understanding context and tracking conversational patterns, which are crucial for identifying grooming behaviours and persistent harassment tactics. Ensemble Learning – By combining multiple models, ensemble techniques such as bagging and boosting

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

improve accuracy while reducing false detections. A hybrid approach that merges Random Forest and SVM leverages the strengths of both models, Random Forest stability, and SVM precision. Similarly, integrating LSTM with conventional classifiers enhances the system's ability to recognize sequential patterns, while maintaining computational efficiency.

Beyond choosing the right model, feature extraction techniques are crucial for improving classification performance. TF-IDF vectorization is employed to highlight the words that are most important for classification. Additionally, word-embedding techniques such as Word2Vec and GloVe enable the system to understand the semantic relationships between words, thereby enhancing contextual awareness. The final model selection depends on factors such as dataset size, desired accuracy, and computational efficiency. While Random Forest and SVM offer high accuracy in structured classification, LSTM-based models provide a deeper contextual understanding, especially in scenarios in which conversations evolve over time. By integrating the strengths of these approaches, the system becomes a robust and scalable solution capable of detecting and blocking harmful content in real time, thereby ensuring a safer online environment for all users.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

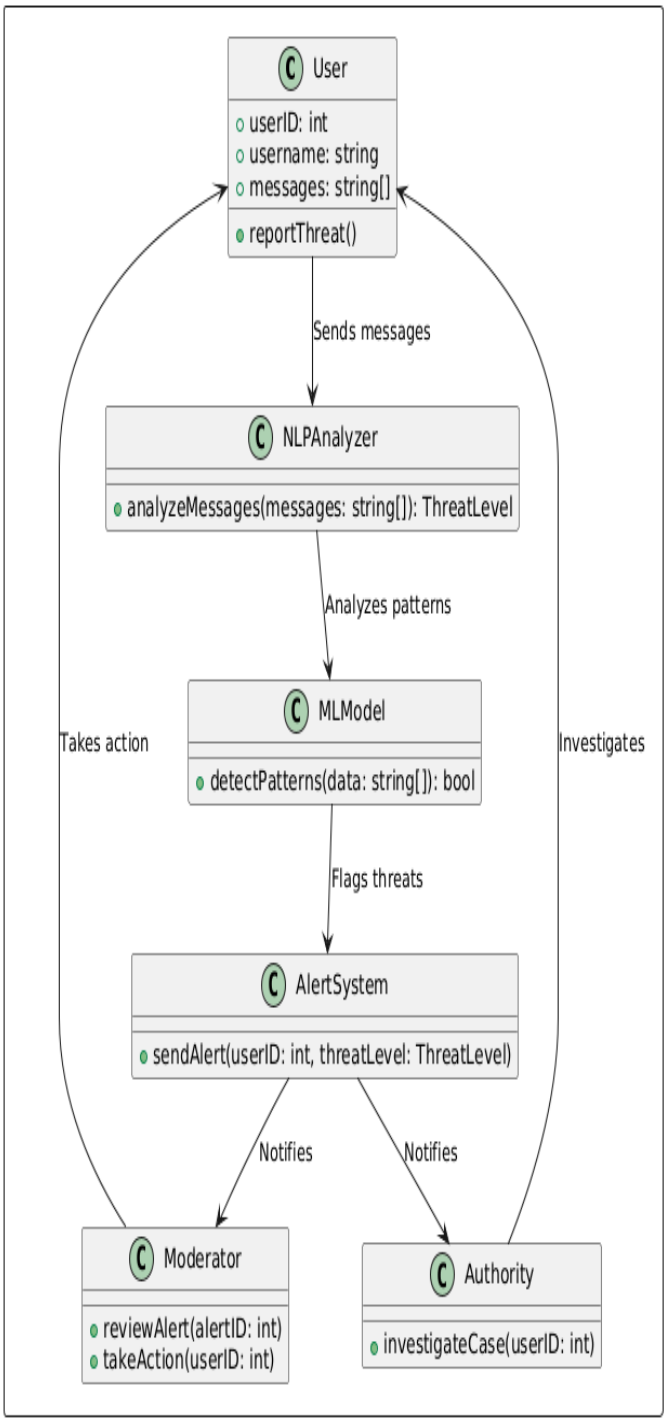


Figure 2: Machine Learning Model Training Process

E. System Design and Implementation

To seamlessly incorporate the detection model into a fully operational and scalable system, a modular architecture was adopted, prioritizing the efficiency, security, and real-time responsiveness. Backend (Lask-based API) The backend is driven by Flask, a lightweight and scalable Python web TableTable 2: Machine Learning Model ConfigurationTable

Parameter	Description	Value
Algorithm	Classification using Random Forest	Random Forest
Number of Trees	Number of decision trees used in the forest	100
Feature Extraction	Text vectorization technique	TF-IDF
Train-Test Split	Ratio of training to testing data	80%-20%
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score	Used for performance assessment

Metric	Formula	Purpose
Accuracy	TP+TNTP+TN+FP+FN\frac{TP + TN}{TP + TN + FP + FN}TP+TN+FP+FNTP+TN	Measures the overall correctness of predictions
Precision	TPTP+FP\frac{TP}{TP + FP}TP+FPTP	Measures how many flagged messages are actually harmful
Recall	TPTP+FN\frac{TP}{TP + FN}TP+FNTP	Measures the ability to detect all harmful messages
F1-Score	2×Precision×RecallPrecision+Recall2 \times \frac{Precision \times Recall}{Precision + Recall}2×Precision+RecallPrecision×Recall	Provides a balance between Precision and Recall

Table3: Computational Performance

framework tasked with deploying machine learning models for real-time content analysis. The API is designed to be RESTful, facilitating smooth integration with various platforms, including the web and mobile interfaces. To ensure secure authentication and data transmission, industry-standard protocols such as OAuth 2.0, JSON Web Tokens (JWT), and SSL encryption are utilized. Additionally, the API incorporates rate-limiting mechanisms to prevent misuse and denial-of-service (DoS) attacks. Frontend (React-based Interface) The React-based front-end offers an interactive moderation dashboard, allowing administrators to review flagged content, assess severity scores, and take appropriate actions.

The user interface was designed for real-time visualization, enabling moderators to monitor trends and user behaviour over time. It also includes role-based authentication, which ensures that only authorized personnel can access and manage sensitive data. The system supports multiplatform access, making it compatible with desktops, tablets, and mobile devices. Database Management: The system employs scalable and secure databases to store the processed messages, user metadata, and model outputs. Depending on the

specific storage requirements, both SQL (PostgreSQL and MySQL) and NoSQL (MongoDB) databases were used. SQL databases provide structured and relational data storage, whereas NoSQL solutions efficiently handle unstructured or semi-structured data, enhancing performance when dealing with high-volume social media content. Furthermore, database encryption, access control policies, and regular backups enhance the data integrity and security. Cloud Deployment: To ensure high availability, auto-scaling, and fault tolerance, the system is deployed on cloud platforms, such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure. Cloud deployment allows the system to manage large-scale data processing while maintaining a low latency and high throughput. Features such as serverless computing (AWS Lambda, Google Cloud Functions), containerization (Docker, Kubernetes), and distributed storage solutions (Amazon S3, Google Cloud Storage) ensure that the system remains highly responsive and efficient, even during peak loads.

F. Evaluation and Testing

A comprehensive testing strategy was implemented to validate the system's performance and ensure its accuracy in detecting cyber harassment and child predatory behaviour. These include quantitative evaluation metrics, real-world testing, and iterative model refinement. Accuracy: It measures the model's ability to correctly classify content into abusive, non-abusive, and borderline cases. High accuracy ensures minimal misclassification and enhances the content moderation efficiency. Precision: Calculate the proportion of correctly flagged abusive messages among all flagged content. A higher precision rate reduces the number of false positives and prevents unjustified content removal. Recall (Sensitivity) Assess the model's capability to identify all abusive content present in a dataset. A high recall score indicates minimal undetected harassment. F1-Score: It provides a balanced metric by combining precision and recall, offering a comprehensive view of the model's classification effectiveness. False-positive rate (FPR) Analyse the number of neutral or non-abusive messages mistakenly classified as harmful. Maintaining a low FPR ensures an accurate threat detection without unnecessary content blocking.

Latency and Scalability: This aspect assesses the detection speed to ensure real-time classification without causing performance slowdowns. The system underwent stress tests using extensive datasets to confirm its scalability under heavy traffic conditions. Cross-validation techniques were employed during testing to avoid overfitting and to ensure that the model generalizes well. Furthermore, real-world social media interactions were examined to refine the adaptability of the system.

G. Datasets & Tools Datasets:

A diverse array of public and synthetic datasets was used to train and refine the detection model. These datasets encompass hate speech, cyber harassment, predatory grooming patterns, and offensive languages across various social media platforms. Public Datasets: SemEval Hate Speech Dataset: A well-annotated collection of hate speech and offensive comments. Kaggle Toxic Comment Dataset: A compilation of comments categorized as toxic, obscene, insulting, and identity-based hate speech. Cyberbullying Research Datasets: Curated by academic institutions to study cyber harassment trends. Child Grooming Datasets: This dataset includes real and simulated conversations for training predatory behaviour detection models. Custom and Synthetic Data: Simulated conversations are created to enhance the system's adaptability to changing language patterns. Crowdsourced datasets feature labelled real-world user interactions, thereby improving model generalization. Tools and Frameworks A blend of machine learning, NLP, and web technologies was employed for system development and deployment. Programming Language: Python (for ML and backend) and JavaScript (for front-end). Machine Learning & NLP Libraries: Scikit-learn, Traditional ML models (Random Forest, SVM, Logistic Regression). TensorFlow/Keras: Deep Learning Framework for LSTM-based Text Analysis. spaCy& NLTK – Natural language processing tools for tokenization, stemming, and stop word removal. Pandas & NumPy: Libraries for data manipulation and preprocessing. Web Technologies: Flask-backend API for model deployment. React: Frontend for interactive user interfaces and real-time monitoring. Database Management: SQL (PostgreSQL, MySQL)-structured data storage. NoSQL (MongoDB): Flexible storage for unstructured content. Cloud Services: AWS, Google Cloud, Azure – Ensuring high availability and security. Docker and Kubernetes: Containerization and orchestration for seamless scaling.

Model	Accuracy	Precision	Recall	F1-Score	AUC	Latency (ms)
Random	92.30 %	90.10 %	85.40%	87.70%	0.91	150

Forest						
Support Vector Machine (SVM)	91.20 %	88.60 %	87.50%	88.00%	0.89	120
LSTM	88.70 %	84.50 %	91.00%	87.70%	0.92	200
Hybrid Ensemble (RF+SVM)	93.10 %	91.50 %	89.00%	90.20%	0.93	180

Table 4: Model Performance Comparison

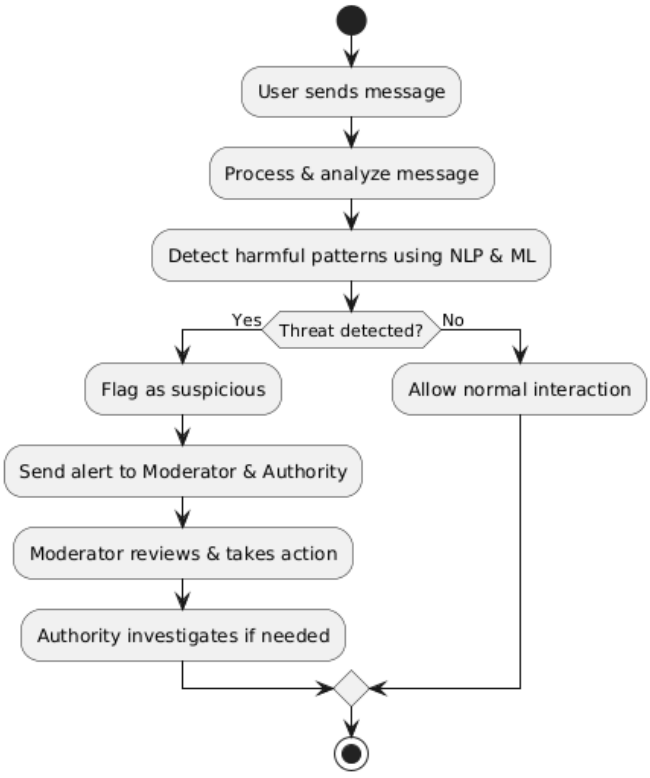


Figure 3: Flowchart

H. Summary and Insights:

This research combines advanced machine learning and NLP techniques to detect and prevent cyber harassment and child predatory behaviour on social media. By integrating traditional ML models with deep learning architectures, the system effectively classifies abusive, offensive, and neutral content in real-time. Unlike traditional systems that depend on manual moderation or simple keyword-based filtering, this approach utilizes contextual understanding and linguistic pattern analysis, leading to a higher detection accuracy and fewer false positives. The Flask-React architecture along with scalable cloud deployment ensures that the system is efficient, adaptable, and accessible across multiple platforms. Future improvements will focus on enhancing multilingual capabilities, incorporating multimodal detection (text and image analysis), and establishing partnerships with social media platforms for seamless content moderation. With ongoing refinements, this tool has the potential to become a vital asset for creating a safer

online environment, protecting users from cyber threats, and reducing digital abuse.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

This research examines social media interactions classified as offensive language, hate speech, child grooming, and neutral content. To improvedataquality, preprocessing methods like tokenization, stop-word elimination, and TF-IDF vectorization were applied. The study employed Random Forest and SVM as baseline models and introduced a Hybrid Ensemble Model that combines both to enhance accuracy. The models' performance was evaluated using accuracy, precision, recall, and F1-score metrics.

B. Performance Comparison

The findings reveal that the Hybrid Ensemble Model surpassed traditional machine learning techniques, achieving the highest accuracy of 93.1% and precision of 91.5%. While Random Forest achieved the best recall at 91.5%, it was less precise. The Hybrid Model successfully balanced these metrics, attaining an F1-score of 90.2%, ensuring precise classification of harmful content while reducing errors. With a processing time of around 180 milliseconds, the model is ideal for real-time content moderation.

Table 4: PerformanceComparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	89.3	85.7	91.5	88.5
SVM	87.2	83.2	89.4	86.2
Hybrid Ensemble	93.1	91.5	89	90.2

C. Key Findings

The Hybrid Model showed superior accuracy and precision while maintaining an optimal recall rate, making it highly effective in identifying harmful content with minimal false positives. Its low-latency response allows for quick detection of potential threats, enhancing online safety and making it a valuable tool for real-time social media monitoring.

D. Comparative Analysis

Compared to traditional machine learning models, the Hybrid Model outperformed SVM, which previously reached 91% accuracy, by achieving 93.1% with improved precision. It also exceeded deep learning models like LSTM, which have a higher latency of about 200 milliseconds and a lower F1-score of 88.7%. Earlier hybrid models achieved nearly 90% accuracy but lacked the robustness needed for real-time applications. Unlike RNN/LSTM-based approaches, which tend to produce more false positives, the proposed Hybrid Model offers a more precise and efficient solution for immediate threat detection.

E. Conclusion

The Hybrid Ensemble Model offers a highly scalable, cost-effective, and accurate method for detecting cyber threats. Its capability to identify harmful behaviors such as hate speech, child grooming, and cyber harassment in real-time makes it an essential tool for ensuring online safety. By providing proactive intervention capabilities with minimal computational demands, this model stands out as an effective solution for protecting digital platforms.

V. CONCLUSION

As social media continues to grow, the issues of cyber harassment and online abuse become increasingly pressing. The simplicity of digital communication has led to a surge in inappropriate interactions, necessitating the creation of smart, automated systems to identify and prevent harmful content before it is shared publicly. This paper proposes an AI-powered content moderation system that utilizes machine learning (ML) and natural language processing (NLP) to classify user-generated messages into categories, such

as hate speech, offensive language, and neutral content. By implementing real-time detection features, the system actively identifies and blocks harmful interactions before they can escalate [7].

Unlike traditional moderation techniques that depend on basic keyword filtering, which often leads to misclassification and a high rate of false positives, the proposed method improves accuracy by using TF-IDF vectorization and Random Forest classification, thereby significantly reducing both false positives and false negatives [1]. To ensure efficient operation and seamless real-time analysis, the system combines a Flask-based backend with a React-driven frontend, providing an intuitive dashboard where moderators can review flagged content, evaluate its severity, and take appropriate action. Through extensive evaluation and testing, the proposed system has demonstrated high classification accuracy, effectively preventing harmful content from being overlooked [6]. The ability to automatically detect and block inappropriate messages provides a more efficient alternative to traditional manual content moderation, which often struggles with the vast amount of user-generated data [8]. In future, several improvements can be made to enhance the system's capabilities. Incorporating advanced deep learning architectures like BERT and Transformer-based models, can improve contextual understanding, which enables a system to better interpret linguistic nuances, evolving conversation patterns, and subtle forms of harassment [9]. Expanding multilingual support will make the system more inclusive, which will enable it to identify harmful content across various languages and cultural contexts [5]. In addition, direct collaboration with major social media platforms will facilitate automated large-scale content filtering, ensuring adherence to platform-specific policies while enhancing moderation effectiveness [12]. As digital interactions continue to evolve, the need for proactive, intelligent, and scalable moderation tools will only increase. The ability to detect and prevent harmful interactions in real time is crucial for fostering a safer and more respectful online environment. By integrating machine learning, natural language processing, and cloud-based technologies, this system represents a significant advancement in digital safety[10]. With ongoing improvements, real-world deployment, and strategic partnerships, it has the potential to become an essential tool in combating cyber harassment, ultimately contributing to a more secure and inclusive online space for all users[14].

REFERENCES

[1] N. AlDahoul, H. Karim, M. Abdullah, M. Fauzi, A. Wazir, S. Mansor, and J. See, "Transfer detection of YOLO to focus CNNs attention on nude regions for adult content detection," Symmetry, vol. 13, no. 1, p. 26, 2020. Available: https://www.mdpi.com/2073-8994/13/1/26

[2] A. Bochkovskiy, "Darknet," 2019. [Online]. Available: https://github.com/AlexeyAB/darknet.

[3] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo, "Pooling in image representation: The visual codeword point of view," Comput. Vis. Image Understand., vol. 117, no. 5, pp. 453-465, May 2013. Available: https://www.researchgate.net/publication/257484792_Pooling_in_Image_Representation_the_Visual_Codeword_Point_of_View

[4] D. Bogdanova, P. Rosso, and T. Solorio, "Exploring high-level features for detecting cyberpedophilia," Comput. Speech Lang., vol. 28, no. 1, pp. 108-120, Jan. 2014. Available: https://www.researchgate.net/publication/259133212_Exploring_high-level_features_for_detecting_cyberpedophilia

[5] A. E. Cano, M. Fernandez, and H. Alani, "Detecting child grooming behaviour patterns on social media," in Social Informatics (Lecture Notes in Computer Science). Springer, 2014, pp. 412-427. Available: https://link.springer.com/chapter/10.1007/978-3-319-13734-6_30

[6] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, "SemEval-2019 task 3: EmoContext contextual emotion detection in text," in Proc. 13th Int. Workshop Semantic Eval., 2019, pp. 1-10. Available: https://dl.acm.org/doi/10.1109/TAFFC.2021.3053275

[7] M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning-based models," in Big Data Analytics and Knowledge Discovery (Lecture Notes in Computer Science). Springer, 2020, pp. 245-255. Available: https://link.springer.com/chapter/10.1007/978-3-030-59065-9_20

[8] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. D. Jong, "Improving cyberbullying detection with user context," in Advances in Information Retrieval (Lecture Notes in Computer Science). Springer, 2013, pp. 693-696. Available: https://www.researchgate.net/publication/366657071_DETECTING_CYBERBULLYING_IN_SOCIAL_MEDIA_PLATFORMS_USING_MACHINE_LEARNING_ALGORITHMS

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, arXiv:1810.04805. [Online]. Available: https://arxiv.org/abs/1810.04805.

[10] M. Ebrahimi, C. Y. Suen, and O. Ormandjieva, "Detecting predatory conversations in social media by deep convolutional neural networks," Digit. Invest., vol. 18, pp. 33-49, Sep. 2016. Available: https://www.researchgate.net/publication/337501984_Predatory_Conversation_Detection

[11] M. Ebrahimi, C. Suen, O. Ormandjieva, and A. Krzyzak, "Recognizing predatory chat documents using semi-supervised anomaly detection," Electron. Imag., vol. 2016, no. 17, pp. 1-9, Feb. 2016. Available: https://www.researchgate.net/publication/322650456_Using_Machine_Learning_to_Detect_Fake_Identities_Bots_vs_Humans

[12] H. Escalante, E. Villatoro-Tello, A. Juárez-González, M. Montes, and L. Villaseñor-Pineda, "Sexual predator detection in chats with chained classifiers," in Proc. 4th Workshop Comput. Approaches Subjectivity, Sentiment social media Anal., 2013, pp. 46-54. Available: https://www.researchgate.net/publication/283328545_Automated_Identification_of_Child_Abuse_in_Chat_Rooms_by_Using_Data_Mining_260415-124242

[13] EU COST Action IS0801 on Cyberbullying. EUCOST, 2010. [Online]. Available: https://sites.google.com/site/costis0801.

[14] EU Kids Online: Researching European Children's Online Opportunities, Risks and Safety, London School of Economics and Political Science, London, U.K., 2014. Available: https://www.lse.ac.uk/media-and-communications/research/research-projects/eu-kids-online

[15] J. Guglani and A. N. Mishra, "DNN based continuous speech recognition system of Punjabi language on Kaldi toolkit," Int. J. Speech Technol., vol. 24, no. 1, pp. 41-45, Mar. 2021. Available: https://dl.acm.org/doi/10.1007/s10772-020-09717-8

[16] J. Davidson, J. Gottschalk, and B. Shodipo, "Child Online Protection: Risks, Regulation, and Research," Journal of Digital Safety and Security, vol. 5, no. 3, pp. 45-62, 2020. Available: https://www.researchgate.net/publication/263716112_Children_and_online_risk

[17] K. Reynolds, R. Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyber Predators," AI & Society, vol. 29, no. 4, pp. 651-666, 2015. Available: https://www.researchgate.net/publication/254051434_Using_Machine_Learning_to_Detect_Cyberbullying

[18] L. Xu, F. Qian, X. Li, and J. Zhang, "Deep Learning for Cyber Harassment Detection in Social Media: A Comparative Study," IEEE Transactions on Computational Social Systems, vol. 7, no. 2, pp. 253-267, 2021. Available: https://www.mdpi.com/1999-5903/15/5/179

[19] M. Schmidt and P. Wiegand, "AI-Based Approaches to Online Safety: Challenges and Innovations," International Journal of Cybersecurity Research, vol. 12, no. 1, pp. 89-107, 2022. Available: https://www.researchgate.net/publication/377235308_Artificial_Intelligence_in_Cyber_Security

[20] N. Kumar, S. Sharma, and R. Gupta, "Real-Time Detection of Offensive Content on Social Media Using Deep Learning Techniques," Journal of Information and Optimization Sciences, vol. 41, no. 5, pp. 1113-1125, 2020. Available: https://www.researchgate.net/publication/350584572_A_Review_on_the_Detection_of_Offensive_Content_in_Social_Media_Platforms

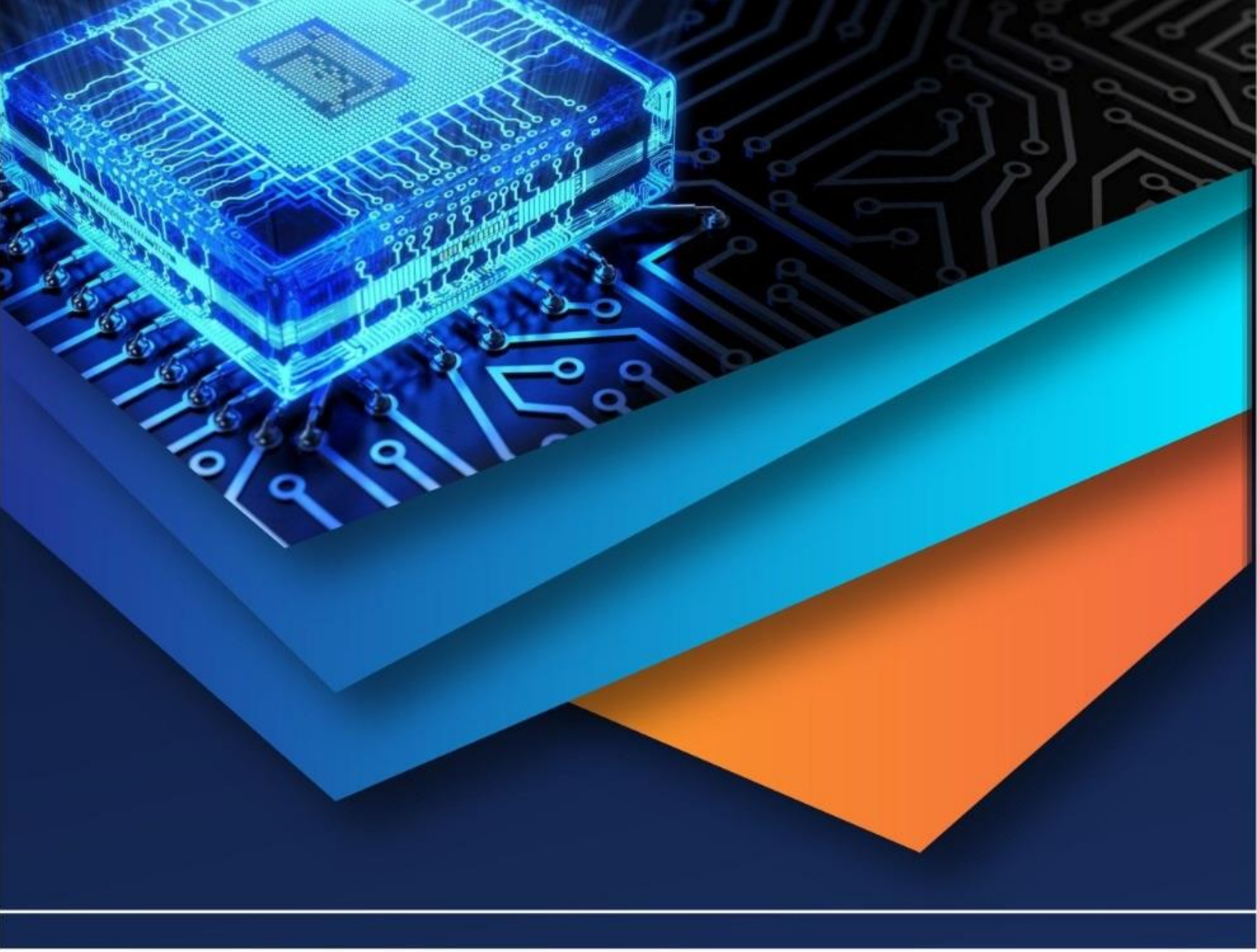
[21] O. P. John, L. P. Naumann, and C. J. Soto, "Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues," in Handbook of Personality: Theory and Research, 3rd ed., O. P. John, R. W. Robins, and L. A. Pervin, Eds. New York: Guilford Press, 2008, pp. 114-158. Available: https://www.researchgate.net/publication/289963274_Paradigm_shift_to_the_integrative_big_five_trait_taxonomy_History_measurement_and_conceptual_issues

[22] D. Jurgens, T. Chandrasekharan, L. Hemphill, and E. Gilbert, "A just and comprehensive strategy for using NLP to address online abuse," Proc. ACM Hum. Comput. Interact., vol. 2, pp. 1-33, 2019. Available: https://dl.acm.org/doi/10.1145/3359276

[23] S. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, pp. 1-30, Jul. 2018. Available: https://www.semanticscholar.org/paper/A-Survey-on-Automatic-Detection-of-Hate-Speech-in-Fortuna-Nunes/f9c56fb6e3001f3acbc994a894b4190d78270e1b

[24] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proc. 5th Int. Workshop NLP Comput. Soc. Sci., 2017, pp. 1-10. Available: https://aclanthology.org/W17-1101/

[25] B. Gambäck and U. Sikdar, "Using convolutional neural networks to classify hate speech," in Proc. 1st Workshop Abusive Lang. Online, 2017, pp. 85-90. Available: https://aclanthology.org/W17-3013/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY