WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Safe conn: AI Enhanced Secure Messaging and Communication

[1] Rayadurgam Yuvath Kumar, [2] Keerthipati Jahnavi, [3] B Sashank, [4] Gamani Sreedhar ,[5] Konduru Anantha Padmanabam,[6] MS. S Shilpa

[1] Student, Dept. of Computer Science and Engineering(AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India. yuvathrayadurgam@gmail.com

[2] Student, Dept. of Computer Science and Engineering(AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India. jahnavikeerthipati94@gmail.com

[3] Student, Dept. of Computer Science and Engineering(AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India. Shashankravijay1234@gmail.com

[4] Student, Dept. of Computer Science and Engineering(AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India. gamanisreedhar@gmail.com

[5] Student, Dept. of Computer Science and Engineering(AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India. ananthkonduru@gmail.com

[6] Guided by., M.E., Assistant Professor, Dept. of Computer Science and Engineering, Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India

ABSTRACT :

The increasing prevalence of harmful content on social communication platforms necessitates advanced moderation techniques beyond traditional manual approaches. This project introduces SafeConn, an AI-enhanced secure messaging and communication platform that ensures a safer digital environment through real-time content moderation. By integrating state-of-the-art AI tools, ChatGPT and Gemini, the system effectively detects and filters offensive language, cyberbullying, illegal content, and privacy violations across text, images, videos, and audio files. The platform leverages ReactJS for a responsive user interface and Firebase for scalable and secure backend services, ensuring robust performance and reliability. Additionally, end-to- end encryption guarantees user privacy, making SafeConn an efficient and secure alternative to conventional messaging platforms. Preliminary results indicate improved moderation accuracy, user safety, and a seamless communication experience.

Keywords: Artificial intelligence, content moderation, end-to-end encryption, Firebase, ReactJS.

INTRODUCTION

Social media platforms significantly influence modern social interactions, enabling real-time sharing and community engagement. Despite their advantages, the rise of harmful content such as bullying, misinformation, hate speech, and privacy violations necessitates more advanced moderation methods beyond manual reporting systems, often resulting in delays and inconsistencies [3].The project addresses this gap by integrating advanced AI tools like ChatGPT and Gemini to automate real-time content moderation, creating a safer digital communication environment [5]. The system employs ReactJS for the user interface and Firebase for backend management, emphasizing robust privacy through end-to-end encryption, ensuring optimal user security [9].

LITERATURE SURVEY

Several studies underline the transition from manual moderation to AI-driven approaches for addressing digital safety challenges:
Livingstone (2017) emphasizes the limitations of manual moderation, including delayed reactions and inconsistent enforcement, advocating for proactive, automated moderation systems [3]. Binns (2018) discusses supervised learning and NLP methods, highlighting their effectiveness but pointing out challenges like dataset bias and privacy concerns [5]. Abreuetal. (2019) illustrate the benefits of deep learning methods such as CNNs and RNNs in multimedia moderation, significantly reducing operational costs and enhancing accuracy [6]. Mohanty et al. (2016) demonstrate CNN effectiveness in real-time moderation of visual content [7], and Singh et al. (2016) explore supervised and unsupervised learning for handling extensive datasets effectively [10]. Ferentinos (2018) critically assesses deep neural networks' effectiveness for automated moderation, underlining computational demands and privacy issues but acknowledging their superior capability in real-time moderation [8].This literature survey reveals a clear evolution towards AI-driven moderation, emphasizing the importance of addressing challenges related to data quality, biases, computational resources, and privacy.

PROPOSED SYSTEM

We propose a platform integrating advanced AI moderation using ChatGPT and Gemini, ensuring instant and accurate filtering of harmful content. ReactJS ensures a responsive user interface, while Firebase provides scalable and reliable backend services. End-to-end encryption significantly enhances user privacy and security.

Key features include:

- Real-time AI-driven content moderation.
- Robust end-to-end encryption.
- Scalable infrastructure using ReactJS and Firebase.

System implementation

System Architecture: The proposed system follows a three-tier architecture:

- **Frontend:** ReactJS, offering intuitive interaction.
- **Backend:** Firebase infrastructure manages authentication, database operations, and real-time synchronization.
- **AI Integration:** ChatGPT and Gemini APIs for content moderation.

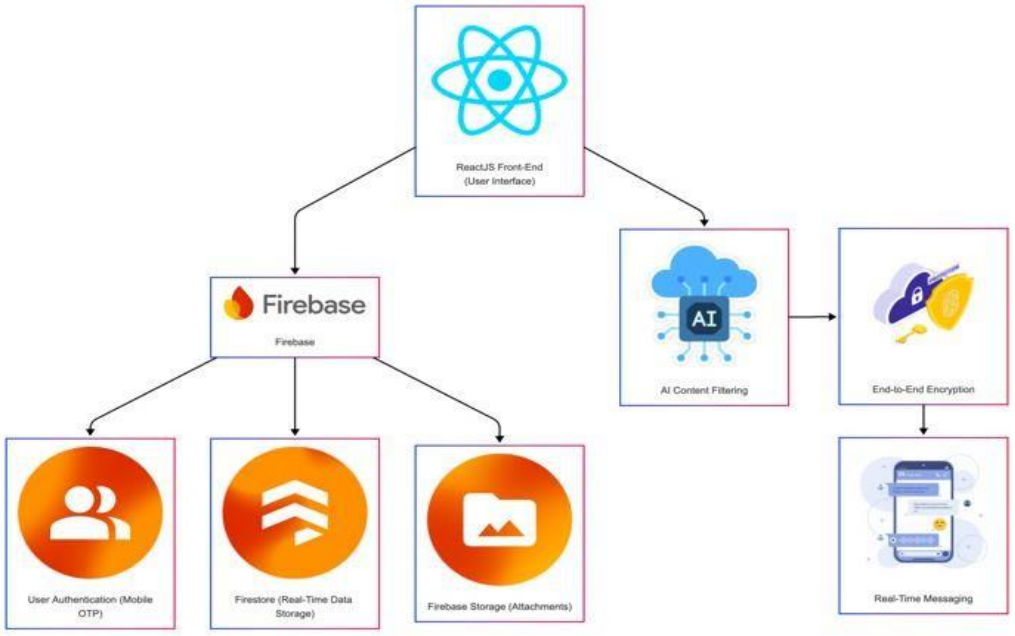

Fig 1. System Architecture

METHODOLOGY

The AI moderation engine employs advanced Natural Language Processing (NLP) and deep learning algorithms to detect harmful content instantly. The system is designed to process text, images, videos, and audio using state-of-the-art AI models like ChatGPT and Gemini, ensuring comprehensive moderation.

The data processing and content filtering phase involves multiple techniques for identifying inappropriate material. Text processing is performed using tokenization, lemmatization, and semantic analysis, which help detect hate speech, offensive language, and misinformation. Meanwhile, image and video moderation is achieved using Convolutional Neural Networks (CNNs) and object detection models, which analyze visual content to flag explicit, violent, or illegal material. Additionally, audio content analysis is facilitated through Automatic Speech Recognition (ASR), converting speech to text for real-time filtering of harmful or inappropriate speech.

For AI model integration, SafeConn employs ChatGPT and Gemini, which enhance content moderation accuracy by providing context-aware filtering. Furthermore, computer vision techniques are utilized for analyzing image and video content, ensuring protection against misleading or inappropriate media. The system also integrates reinforcement learning, enabling continuous AI model improvements based on user feedback and flagged content, ensuring adaptability and effectiveness over time.

Security and privacy are fundamental aspects of SafeConn's design. The system implements End-to-End Encryption (E2EE) to ensure that user conversations remain private while still allowing AI-driven moderation. Anonymized data processing techniques are employed, ensuring that the system filters content without storing any personally identifiable information. Additionally, secure authentication mechanisms are integrated using Firebase authentication, which safeguards user identities and prevents unauthorized access to the platform.
To optimize performance, SafeConn is designed for real-time detection, using low-latency APIs to ensure instant content filtering with minimal delay. The platform is built on Firebase's cloud infrastructure, offering high availability and fault tolerance, making it highly scalable. Moreover, adaptive filtering techniques allow AI models to dynamically adjust sensitivity
levels based on user preferences and regulatory compliance, providing a more personalized and reliable moderation system.

Looking ahead, SafeConn aims to introduce several future enhancements to further improve its moderation capabilities. One of the key areas of expansion includes multilingual moderation, where NLP models will be optimized to support multiple languages, ensuring inclusivity for a global user base. To address concerns of algorithmic bias, fairness-aware AI models will be implemented, reducing the risk of biased moderation outcomes. Additionally, blockchain technology is being explored as a potential solution for decentralized verification of content authenticity and moderation logs, adding an extra layer of transparency and security to the system.
Through these methodologies, SafeConn ensures an advanced, AI-driven content moderation platform that not only enhances digital safety but also maintains user privacy and a seamless communication experience.

RESULT AND DISCUSSION

Preliminary evaluations indicate significant improvements in moderation speed, accuracy, and user satisfaction compared to manual and traditional automated moderation methods. Real-time filtering effectively reduced exposure to harmful content.
Pilot testing results demonstrate

- **Accuracy of AI Content Filtering:** High accuracy in detecting offensive language and harmful content in text messages.
- **Real-Time Performance:** Instant message filtering with minimal delay, ensuring a smooth and responsive experience for users.

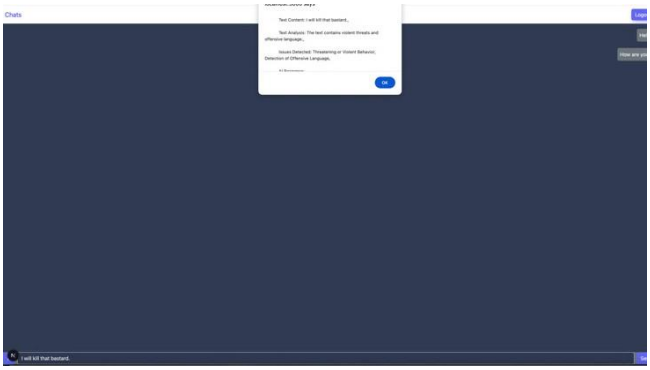

Fig 2. Chat Screen AI Response

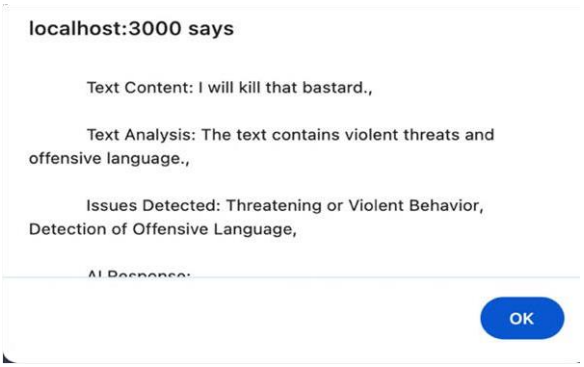

Fig 3. AI Response for a Message

ADVANTAGES OF THE PROPOSED SYSTEM

- Instantaneous detection and removal of harmful content.
- Enhanced user privacy via end-to-end encryption.
- Scalability and reliability through ReactJS and Firebase integration.
- Superior moderation accuracy leveraging AI algorithms

CRITERIA OF SAFE COMMUNICATION

1. Detection of Offensive Language
2. Hate Speech and Bullying
3. Illegal or Inappropriate Content
4. Misleading or False Information
5. Age-Inappropriate Content (18+)
6. Spam and Unwanted Solicitations
7. Threatening or Violent Behavior
8. Privacy Violations
9. Impersonation or Deceptive Profiles
10. Scams and Fraudulent Activities

11. Inappropriate Use of AI Filters
12. Context-Aware Filtering
13. Cultural Sensitivity
14. Self-Harm or Suicide Promotion
15. Terrorism and Extremist Content
16. Deceptive or Manipulated Media
17. Harassment and Stalking
18. Implied or Hidden Harmful Content
19. Encouraging Unsafe Behavior
20. Banned Substances and Items
21. Grooming or Exploitation
22. Gender, Sexuality, and Identity-Based Discrimination
23. Subtle Forms of Cyberbullying
24. Encouraging Harmful Ideologies
25. Obscured or Coded Language
26. Detection of Deepfakes in Images, Videos, and Audio

CONCLUSION

In conclusion, the proposed AI-enhanced moderation platform addresses major shortcomings of traditional moderation approaches, offering real-time, automated moderation with advanced AI integration. It promises a safer and more engaging user environment, significantly improving user experience, safety, and privacy on social communication platforms.

Index Terms: Artificial Intelligence, ChatGPT, Content Moderation, End-to-End Encryption, Firebase, Gemini, ReactJS, Social Communication Platform.

REFERENCES

[1] J. Zhang, L. Sui, L. Zhuo, Z. Li and Y. Yang, "An approach of bag-of-words based on visual attention model for pornographic images recognition in the compressed domain" in Neurocomputing, Elsevier BV, vol. 110, pp. 145-152, Jun. 2013.
[2] M. Fleck, D. Forsyth and C. Bregler, "Finding naked people", Computer Vision ECCV, pp. 593-602, 1996.
[3] "Most popular mobile messaging apps worldwide as of January 2017 b." in Most popular messaging apps 2017 | Statista, 2017.
[4] Sri Nishant Reddy Lakkireddy, Aaron A Thomas, T. Shreya Shree, and Talakoti Mamatha (2023). Web-based Application for Real-Time Chatting using Firebase.
[5] C. Schmidt and J. Hofmann, "Automated Hate Speech Detection Using Deep Learning Techniques," in IEEE Transactions on Computational Social Systems, vol. 6, no. 3,
pp. 512-523, 2020.
[6] M. Abreu, L. Silva, and R. A. Santos, "Deep Learning Approaches for Detecting Harmful Content in Social Media Platforms," in International Conference on Artificial Intelligence and Security, Springer, pp. 120-132, 2019.
[7] Mohanty, B. Singh, and K. Sharma, "Real-time Image and Video Moderation using Convolutional Neural Networks," in Proceedings of the ACM Multimedia Conference,
pp. 211-220, 2016.
[8] D. Ferentinos, "Deep Learning Models for Cyberbullying Detection in Online Communications," in Neural Networks Journal, Elsevier, vol. 105, pp. 178-187, 2018.
[9] H. Kim and S. Lee, "Privacy-Preserving AI Moderation in Encrypted Messaging Services," in IEEE Access, vol. 9, pp. 34029-34041, 2021.
[10] R. Singh, P. Kumar, and V. Gupta, "Scalable AI-based Moderation for Online Messaging Applications," in Journal of Web Intelligence, vol. 27, pp. 198-210, 2022.
[11] B. T. Johnson and M. K. Patel, "Comparative Study of NLP Techniques for Detecting Offensive Language in Chat Applications," i n International Journal of Artificial Intelligence & Data Science, vol. 15, no. 2, pp. 87-102, 2021.
[12] Z. Ahmed and L. Wang, "End-to-End Encryption and AI-Powered Moderation: A Dual-Layer Approach," in Cybersecurity and Privacy Journal, vol. 5, no. 3, pp. 233- 247, 2022.
[13] Y. Nakamura, "The Role of AI in Combating Misinformation and Fake News on Social Media," in International Conference on Social Computing and AI, pp. 45-57, 2020.