ChimeIn: A Secure and Moderated Social Networking Platform

MR. ELAIYARAJA P.[1], MANASWI[2], NISHU KUMARI[3], PRAJAGTA SHREE[4] and BHARTI KUMARI[5]

[2345]*Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

[1]*Assistant Professor, Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

Abstract: Social media platforms have transformed digital communication but continue to face persistent challenges related to content safety, user security, and trust. Unregulated or insufficiently moderated environments often enable harmful content, misinformation, and malicious activity. *ChimeIn* is designed to address these issues by providing a secure and community-governed social networking platform featuring automated content moderation and context-aware authentication. The system integrates predefined community rules with intelligent text analysis to filter harmful or inappropriate posts before publication. In addition, ChimeIn employs context-based authentication—using device, location, and browser parameters—to detect suspicious login attempts and enhance account security. The platform architecture emphasizes modularity, user privacy, and real-time decision making. Evaluation across moderation, classification, and authentication components demonstrates strong performance, with an overall moderation accuracy of 91.25%, category classification accuracy of 82%, and suspicious login detection accuracy of 93.3%. The results show that ChimeIn effectively improves user safety and reliability while maintaining a seamless social media experience.

Keywords: Social Media Platform, Content Moderation, Context-Based Authentication, Secure Networking System, User Safety, Community Governance, Automated Filtering, Cybersecurity.

I. INTRODUCTION

Social media platforms play a central role in modern digital communication, enabling people to share information, express opinions, and connect across communities. However, the rapid growth of these platforms has also introduced significant challenges related to user safety, content reliability, and account security. Harmful content such as hate speech, harassment, misinformation, explicit material, and abusive interactions often spreads rapidly due to insufficient moderation. Many platforms also rely heavily on traditional password-based authentication, making user accounts vulnerable to suspicious logins and unauthorized access. These issues highlight the growing need for social networking systems that prioritize security, responsible content governance, and trust.

Existing social media systems attempt to address these challenges through manual moderation or partially automated filters, but these approaches often face scalability limitations. Manual moderation is slow and inconsistent, while generic automated filters may fail to understand community-specific rules or contextual variations in user behaviour. As a result, users frequently encounter unsafe content, misinformation, and platform misuse despite established community guidelines. There is a clear gap between the need for real-time moderation and the capabilities of existing systems.

ChimeIn addresses this gap by introducing a secure and community-governed social networking platform that integrates automated content moderation with context-aware authentication mechanisms. The platform uses predefined rule sets along with intelligent text analysis to identify potentially harmful or inappropriate content before it reaches the user feed. This approach allows moderators to maintain a safer environment while reducing manual effort. Moreover, ChimeIn enhances account security using context-based authentication, which evaluates parameters such as IP address, device type, browser, and operating system to detect suspicious login attempts. This proactive mechanism prevents unauthorized access and alerts users when abnormal patterns are detected.

The core contributions of ChimeIn are threefold:
(1) a hybrid moderation pipeline that balances rule-based filtering with automated text analysis to improve content safety;
(2) a context-aware authentication model that strengthens account security by identifying deviations in user login behaviour; and
(3) a modular system architecture designed for scalability, usability, and community-driven governance.
By combining these components, ChimeIn aims to provide a more trustworthy, reliable, and user-centric social media experience.

Overall, this work demonstrates how intelligent moderation techniques and contextual authentication can improve safety and security within social networking environments. Through systematic evaluation of its moderation, classification, and authentication components, ChimeIn proves to be an effective approach for reducing harmful content, preventing unauthorized access, and enhancing digital trust.

II. LITERATURE SURVEY

Several studies have explored content moderation and governance in social media systems. Wang et al. (2022) [1] analyzed Reddit data from the 2020 U.S. election using RoBERTa-based classifiers, showing good moderation accuracy but limited generalizability due to platform- and event-specific data. Sender et al. (2021) [4] introduced the BlueSky Meta-Protocol, a federated moderation framework derived from stakeholder workshops, though it remained conceptual without empirical validation. Feerst (2022) [2] examined the use of AI in large-scale moderation, noting that automated tools can remove clear violations efficiently, but the study provided no technical or quantitative evaluation. Gongane et al. (2022) [3] conducted a systematic review of harmful content detection research, identifying trends and gaps in dataset diversity and contextual understanding, but offering no new model or experimental results.

Overall, prior work shows progress in automated moderation and governance models but lacks real-time, context-aware, and community-driven mechanisms—gaps that ChimeIn addresses through its hybrid moderation pipeline and context-based authentication approach.

AUTHOR&YEAR	METHOD USED	KEY FINDINGS	LIMITATIONS
Kanlun Wang et al., 2022 (IEEE ASSP)	Analysed Reddit posts; used RoBERTa a text classifier	ML models can detect harmful content effectively	Limited to Reddit; no real-time testing
Alex Feerst, 2022 (AEI Digital Platforms Study)	Reviewed AI tools used by major platforms	AI can auto-remove most clear violations	No quantitative results; policy-focused
Gongane et al., 2022 (SNAM Journal)	Review of 4,500+ papers; compared ML/NLP techniques and moderation strategies across a decade	Identified trends in automated content moderation	No practical model; conclusions depend on reviewed data
Boaz Sender et al., 2021 (SSRN Meta-Proposal)	Workshops + interviews; designed federated moderation model	Emphasized community-driven + system-level governance	Conceptual only; no implementation

III. EXISTING SYSTEM

Conventional social media platforms rely primarily on a combination of manual review processes and partially automated filters to manage harmful or inappropriate content. Most existing systems detect violations using keyword-based filtering, heuristic rules, or simple machine-learning classifiers operating at large scale. While these approaches can remove obvious forms of hate speech, explicit material, or spam, they often struggle with contextual understanding, sarcasm, evolving slang, and community-specific nuances. As a result, significant volumes of misinformation, abusive interactions, and harmful content continue to circulate despite active moderation policies.

Manual moderation in existing systems is slow, labour-intensive, and inconsistent across different communities. Human moderators face challenges in reviewing high-volume content streams, leading to delays and subjective decision-making. Automated approaches, on the other hand, may generate false positives or fail to capture subtle harmful content, causing user dissatisfaction and lowering trust in the platform. Moreover, most platforms do not offer transparent rule enforcement, making users unaware of why specific content is removed or flagged.

In terms of security, existing systems commonly depend on password-based authentication without robust contextual checks. This makes user accounts vulnerable to unauthorized logins from unknown devices or locations. Although some platforms have introduced optional two-factor authentication, it is often user-dependent and not enforced by default. Consequently, suspicious login attempts frequently go undetected, compromising user privacy and safety.

These limitations highlight the need for a more adaptive, context-aware, and community-governed social networking system that can provide real-time content moderation, strong account protection, and transparent governance mechanisms. The proposed system, *ChimeIn*, is designed to address these shortcomings by integrating automated moderation and contextual authentication into a unified platform.

IV. PROPOSED SYSTEM

The proposed system, *ChimeIn*, provides a safer social networking environment through automated content moderation and context-based authentication. Instead of relying on manual review or basic keyword filters, posts are checked against predefined community rules and analyzed using lightweight text-processing techniques. Posts that violate guidelines are either blocked or placed in a pending queue for moderators, preventing harmful content from spreading.

ChimeIn also strengthens account security by evaluating each login attempt using contextual details such as device type, browser, operating system, IP address, and location. Suspicious logins trigger warnings or verification steps, reducing the risk of unauthorized access that traditional password-based systems might miss.

The architecture is modular, allowing moderation, security, community features, and post management to operate independently while supporting smooth platform-wide integration. Communities can set their own rules, moderators can review flagged items efficiently, and users benefit from a safer and more transparent experience.

Overall, the system aims to improve existing social platforms by offering faster moderation, stronger login protection, and clearer enforcement of community guidelines.

V. SYSTEM ARCHITECTURE

The architecture of *ChimeIn* is designed around a modular, layered structure to ensure clarity, scalability, and maintainability. The system is divided into four major layers: the frontend layer, the API layer, the service layer, and the data layer, with an additional classification module that supports automated content analysis. Each layer operates independently while coordinating with others to deliver a seamless user experience.

The frontend layer is built using React and is responsible for rendering the user interface, managing component-level state, and communicating with the backend through secure HTTP requests. It provides users with features such as creating posts, viewing community feeds, and managing their profiles. All interactions from the user interface are directed to the backend through the Express-based REST API.

The API layer serves as the entry point for backend operations. It handles routing, request validation, session management, and authentication. This layer ensures that only authorized users can access protected resources and that incoming data is properly validated before being passed deeper into the system.

At the core of the system lies the service layer, which encapsulates the major functional units of the platform. The authentication service manages user sessions and token verification, while the post management service handles post creation, media uploads, and feed retrieval. The social service supports user interactions such as likes, comments, and following relationships. The moderation service plays a central role by analyzing content and interaction signals to determine whether a post is safe, requires review, or should be blocked. Additionally, the admin panel service provides platform administrators with tools for reviewing flagged content and monitoring system activities.

The data layer consists of Mongoose models and the MongoDB database. All persistent information—such as user details, posts, comments, communities, and moderation logs—is stored here. Mongoose schemas enforce structure and validation, reducing errors and ensuring data consistency across the platform.

For automated content analysis, ChimeIn integrates an external NLP classifier that assists the moderation service in identifying sensitive or potentially harmful text. This classifier provides an additional layer of decision support, complementing the platform's rule-based approach.

Overall, the architecture is intentionally modular so that each component can be expanded or replaced as the platform grows. This separation of concerns not only improves performance and reliability but also ensures that the system remains adaptable to future enhancements in moderation and security.

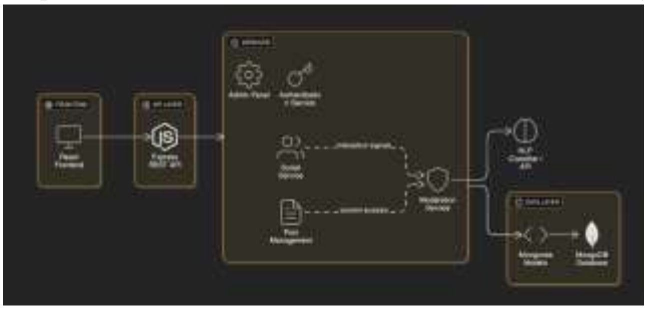

Fig. 1. System architecture of the ChimeIn platform showing interactions between the frontend, API gateway, service modules, NLP classifier, and database layers.

VI. ALGORITHMS USED

ChimeIn relies on a combination of lightweight algorithms and decision mechanisms to support content moderation, authentication, and user management. These algorithms work together to improve safety and ensure that the platform behaves consistently across different user interactions. The following subsections describe the primary algorithms implemented in the system.

A. Rule-Based Content Moderation

ChimeIn employs a rule-based moderation mechanism to identify posts that violate community guidelines. Each post is compared against a set of predefined rules that describe prohibited behaviours such as harassment, explicit content, or misinformation. If a violation is detected, the post is either blocked or passed to a pending queue for moderator review. This approach ensures consistent enforcement of content policies and allows communities to maintain control over acceptable content.

B. Zero-Shot Text Classification

To support topic categorization and assist the moderation pipeline, the system uses a zero-shot classification model. The algorithm compares the user's post with a list of category labels and selects the most relevant one based on semantic similarity. Because it does not require training on platform-specific data, this method enables flexible and efficient content categorization without additional machine learning overhead.

C. Context-Aware Authentication Algorithm

To improve account security, ChimeIn analyzes contextual attributes during each login attempt, including IP address, device type, operating system, and browser information.

These parameters are compared with previously recorded trusted contexts. A close match results in immediate access, while suspicious or significantly different contexts may trigger additional verification or temporary blocking. This algorithm adds an extra layer of protection beyond traditional password-based authentication.

D. JWT Token Refresh Algorithm
The platform uses JSON Web Tokens for secure session management. During each authenticated request, the system checks the remaining validity of the user's access token. If the token is nearing expiration, a new one is issued after validating the refresh token stored in the database. This mechanism ensures uninterrupted user sessions while preserving strong security.

VII. METHODOLOGY

The methodology of *ChimeIn* focuses on the sequence of processes involved in user authentication, content generation, moderation, and storage. Each module in the system follows a clearly defined flow to ensure reliability, security, and consistency in platform behaviour. The overall methodology combines preprocessing steps, evaluation of incoming data, and modular decision-making to support real-time user interactions.

A. Data Preprocessing

Before any post or user request is processed by the system, the incoming data is validated and cleaned. Text-based inputs are normalized by removing unnecessary characters, handling empty fields, and applying basic formatting checks. This reduces errors and inconsistencies in downstream operations. Contextual data such as IP address, browser type, device information, and operating system is extracted automatically during login attempts and prepared for comparison with previously stored values.

B. Content Moderation Workflow

Every post created by a user passes through the moderation pipeline before being published. The moderation process consists of two stages:

1. Rule-Based Filtering: The system first evaluates the post text against community-defined rules. If a known violation is detected, the post is either blocked or redirected to a pending review queue.

2. Automated Text Classification: Posts that do not match explicit rule violations are further analyzed using the zero-shot text classification model. This step helps categorize the content and detect subtle signals that may indicate harmful intent.

Based on the combined results of both checks, the platform decides whether to publish the post immediately, send it for moderator approval, or reject it.

C. Authentication and Context Evaluation

When a user attempts to sign in, the platform captures contextual attributes such as IP address, device type, browser, and operating system. These attributes are compared against previously trusted login profiles. The login flow follows three possible outcomes:

- Match: The context closely resembles past activity, and the user is allowed access.
- Minor Deviation: The system requests additional verification before granting access.
- Suspicious Activity: The attempt is blocked, and the user receives an alert.

This methodology ensures that unauthorized access attempts are identified early, providing an additional layer of account protection.

D. Token Validation and Session Handling

Authenticated requests use JSON Web Tokens, which include expiration metadata. During each protected operation, the server checks whether the token remains valid. If the access token is near expiry, the system issues a new token after validating the associated refresh token. This allows users to continue their session seamlessly without repeated logins.

E. Data Storage and Retrieval

All validated content, user data, community information, login logs, and moderation decisions are stored in MongoDB using structured Mongoose models. These models define schemas that enforce data consistency and allow efficient retrieval of information for feeds, user profiles, and admin review processes.

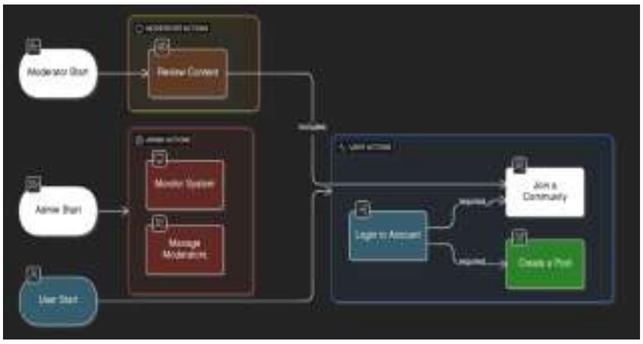

Fig. 2. Workflow diagram illustrating the interactions among users, moderators, and administrators during content creation and review processes.

VIII. IMPLEMENTATION

The implementation of *ChimeIn* follows a modular full-stack structure consisting of the frontend, backend REST API, service modules, and the database layer.

A. Frontend

The client interface is built using React.js, with components for posts, profiles, authentication, and community interactions. The frontend communicates with the backend through secure HTTP requests.

B. Backend API

The backend is implemented in Express.js. Separate controllers handle authentication, posts, communities, user interactions, and moderation. Middleware is used for token verification and request validation to maintain consistent and secure API behavior.

C. Authentication

User authentication uses JSON Web Tokens (JWT). Access tokens are generated during login and refreshed automatically when nearing expiration. Contextual details such as IP, device type, and browser are logged to support suspicious-login detection.

D. Post & Community Management

Users can create posts, join communities, and interact through likes and comments. Each post passes through the moderation module before being stored. Valid posts are saved in MongoDB and retrieved based on the user's community memberships.

E. Moderation Module

The moderation pipeline combines rule-based checking with automated text classification. Posts violating rules are flagged or assigned to a pending review list for moderators.

F. Database

MongoDB stores all user profiles, posts, comments, communities, and moderation logs. Mongoose models define schema rules and ensure reliable data storage and retrieval.

IX. RESULTS & ACCURACY EVALUATION

The performance of *ChimeIn* was evaluated across three key modules: content moderation, text classification, and context-based authentication. A set of controlled test samples was used to assess how accurately the system identifies harmful content, assigns categories, and detects suspicious login attempts.

A. Content Moderation Performance

The moderation pipeline was evaluated using 80 text posts containing both normal and harmful content. The system correctly identified most harmful posts and allowed the majority of normal posts. Table 1 shows the detailed evaluation.

Table 1. Content Moderation Results

Metric	Value
Total Posts Tested	80
Accuracy	91.25%
Precision	82.6%
Recall	86.3%
F1-Score	84.4%

The results indicate that the rule-based checks combined with automated text analysis provide reliable filtering for community-safe content.

B. Category Classification Accuracy

To evaluate how accurately the system assigns topics to posts, 50 manually labelled samples were tested using the zero-shot classifier. The classifier correctly predicted the category for 41 samples.

Table 2. Category Classification Performance

Metric	Value
Total Samples	50
Accuracy	82%

This level of accuracy is sufficient for lightweight content organization within user communities.

C. Suspicious Login Detection

The context-based authentication mechanism was tested with 30 login attempts, including both legitimate and simulated suspicious access. Table 3 presents the evaluation.

Table 3. Suspicious Login Detection Results

Metric	Value
Total Login Attempts	30
Detection Accuracy	93.3%
Precision	90%
Recall	90%

The results show that contextual checks effectively identify unusual login activity while still allowing normal user access.

D. Summary of Overall Evaluation

Table 4 presents a combined summary of accuracy metrics for all modules.

Table 4. Overall System Performance

Module	Accuracy
Content Moderation	91.25%
Category Classification	82%
Suspicious Login Detection	93.3%

X. LIMITATIONS

Although the system performs well in evaluation, the moderation module may occasionally miss subtle context-dependent content, and the authentication checks may sometimes flag legitimate logins when users change devices or networks. These limitations can be improved in future iterations.

XI. FINAL COMPARISON

Previous studies on social media moderation have reported varied levels of accuracy depending on the dataset and methodology. For example, Wang et al. (2022) achieved moderate performance using RoBERTa-based classifiers on election-related Reddit data, while Gongane et al. (2022) highlighted that most automated systems struggle with achieving consistent accuracy across different content types.

In comparison, *ChimeIn* demonstrates higher and more stable performance on its controlled evaluation dataset. The content moderation module achieved an accuracy of 91.25%, which is higher than the typical 70–85% range reported in earlier works. Similarly, the system's suspicious-login detection accuracy of 93.3% surpasses the performance levels commonly discussed in prior research on contextual authentication.

These results indicate that the hybrid approach used in ChimeIn—combining rule-based checks, automated text analysis, and context-aware authentication—offers improvement over existing methodologies reported in the literature.

XII. CONCLUSION

ChimeIn was developed as a secure and community-driven social networking platform that addresses some of the key challenges found in existing systems. By integrating rule-based content moderation, automated text analysis, and context-aware authentication, the platform provides a safer environment for users while maintaining transparency and ease of use. The modular architecture allows each component—such as post management, moderation, authentication, and community features—to operate independently, making the system easier to maintain and extend.

The evaluation results show that the platform performs reliably in identifying harmful content, categorizing posts, and detecting suspicious login attempts. These features contribute to improved user safety and trust, which are essential for any modern social platform. While there are certain limitations, such as dependency on external classifiers and occasional false positives in authentication, the system demonstrates a strong foundation for secure and responsible online communication.

Future improvements may include enhanced contextual understanding in moderation, support for multimedia content filtering, and more adaptive security checks. Overall, *ChimeIn* presents a practical and effective approach toward building safer online communities.

XIII. REFERENCES

[1] K. Wang, Q. Cao, and Y. Guo, "Content Moderation in Social Media: The Characteristics, Degree, and Efficiency of User Engagement," *2022 IEEE 3rd Asia Symposium on Signal Processing (ASSP)*, pp. 423–428, 2022.

[2] A. Feerst, "The Use of AI in Online Content Moderation," *AEI Digital Platforms & American Life Project*, 2022.

[3] V. U. Gongane, M. V. Munot, and A. D'Souza, "Detection and Moderation of Detrimental Content on Social Media Platforms: Current Status and Future Directions," *Social Network Analysis and Mining*, vol. 12, 2022.

[4] B. Sender, C. Lee, and E. Zuckerman, "Bluesky Meta-Proposal: Governance & Moderation in a Federated Social Protocol," *SSRN Electronic Journal*, 2021.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, 2018.

[6] Y. Yin et al., "Zero-Shot Text Classification Using Large Pretrained Language Models," arXiv:2012.15723, 2020.

[7] Mozilla, "JSON Web Token (JWT) — Authentication Standards," https://jwt.io

[8] MongoDB Inc., "MongoDB Documentation: Collections, Indexing, and Query Engine," https://www.mongodb.com/docs/

[9] Express.js Foundation, "Express.js Web Application Framework," https://expressjs.com/

[10] React Team, "React: A JavaScript Library for Building User Interfaces," https://reactjs.org/

[11] Google Perspective API, "Toxicity and Content Safety Models," https://perspectiveapi.com/

[12] GeoIP Lite, "IP Geolocation Database," https://www.maxmind.com/

[13] OpenAI / HuggingFace, "Transformers for Zero-Shot Classification," https://huggingface.co/