



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: <https://doi.org/10.22214/ijraset.2025.75989>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

RNSTweets: An AI-Driven Secure Social Media Platform for Cyberbullying Prevention in Academic Institution

Nayana CP, Manish P, Manish PM, Nagarjun PV, Pavan Kumar H

Department of Information Science and Engineering RNS Institute of Technology Bengaluru, India

Abstract: Cyberbullying has become a significant concern across social media platforms and has affected students' mental well-being, academic performance, and digital safety. Mainstream platforms lack institution-specific controls, thus allowing harmful content to spread unbounded. RNSTweets is an AI-powered, closed-community microblogging platform designed for the RNS Institute of Technology. It integrates transformer-based NLP models such as BERT and HateBERT for real-time abusive language detection, supported by a demerit-based penalty mechanism and an admin moderation dashboard. The platform uses a modular architecture with a React and Next.js frontend and a Next.js and MongoDB backend for secure student data storage. By combining automated moderation with secure authentication and controlled community access, RNSTweets offers a secure, collaborative environment for academic communication.

Index Terms: Cyberbullying detection, BERT, HateBERT, Secure Social Media, NLP, Academic Communication, Moderation Systems.

I. INTRODUCTION

Social media plays a significant role in modern communication. It enables large-scale information sharing, collaboration, and interaction among users. However, these sites have likewise become environments where cyberbullying, harassment, hate speech and abusive behavior manifest. Students, in particular, are more susceptible to these dangers given the constant online presence, and peer pressure-driven interactions. Research studies have demonstrated that cyberbullying directly influences a student's mental health, academic performance, and self-esteem. Mainstream social media platforms such as Twitter and Instagram are not tailored for academic institutions. They lack institution-level authentication and real-time moderation for student interactions and administrative oversight. Schools and colleges increasingly require safer digital spaces for communication that are monitored, moderated, and accountable.

To address this need, we introduce RNSTweets, a secure microblogging platform designed exclusively for RNSIT. It is inspired by Twitter, but with critical safety features including:

- AI-powered cyberbullying detection.
- Domain-based authentication for students and faculty.
- A demerit-based violation tracking system.
- Admin moderation dashboard.
- Secure storage and role-based access control.

II. RELATED WORK

A. Cyberbullying Detection on Social Media

There is a large body of research on detecting abusive or harmful content across platforms such as Twitter, YouTube, and Instagram. Existing approaches range from traditional machine learning to deep learning and transformer-based models.

1) *Machine Learning Models:* Early cyberbullying detection systems employed traditional machine learning models such as Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression. These approaches depend on handcrafted features including:

- n-grams
- TF-IDF representations
- sentiment polarity features

- profanity and hate speech lexicons

While effective to a certain extent, these models struggle with capturing context, sarcasm, and multi-lingual slang, which are common in social media discourse.

2) *Deep Learning Approaches*: Recent deep learning architectures have demonstrated better performance for cyberbullying detection, especially on short and noisy texts:

- Convolutional Neural Networks (CNNs) for local text feature extraction.
- LSTMs/GRUs for modeling sequential context in user posts.
- Bidirectional LSTMs with attention mechanisms to emphasize important words and phrases.

These methods are better at capturing semantic information compared to purely feature-engineered models.

3) *Transformer-Based Models*: State-of-the-art work increasingly employs transformer-based language models such as:

- BERT
- RoBERTa
- DistilBERT
- HateBERT (fine-tuned for toxic and abusive content)

Transformers provide deep contextual representations and have improved accuracy in classifying bullying, detecting toxic speech, and recognizing hate speech. Our project is closely aligned with this line of research by using AI-powered moderation through large language models and transformer-based APIs.

B. AI-Assisted Content Moderation

Recent studies explore AI-assisted moderation mechanisms for large-scale social networks.

1) *Rule-Based Filtering*: Early moderation systems mainly relied on keyword and profanity lists. Although simple to implement, these rule-based filters suffer from:

- High false positives due to lack of context.
- Easy evasion through spelling variations, code words, or sarcasm.

2) *Context-Aware AI Moderation*: Modern systems incorporate:

- sentiment analysis and toxicity scoring,
- contextual embeddings from deep neural models,
- conversation-level modeling of reply chains and threads. Large platformssuchasTwitterandRedditemployautomated flaggingsystems, but these are typically not transparent and are not tailored to specific institutions. RNS Tweets contributes by applying AI moderation in a constrained academic environment rather than an open global social network.

C. Digital Safety in Academic Environments

Several studies emphasize the need for safe digital communication systems for students. Key findings include:

- Students are highly susceptible to online harassment due to peer dynamics and power imbalances.
- Most institutions lack dedicated communication mechanisms with built-in safety controls and monitoring.
- Existing LMS tools such as Moodle, Canvas, and Google Classroom are primarily focused on teaching workflows, not peer-to-peer social interaction.
- Cyberbullying on college campuses often goes unreported, leading to long-term psychological and academic impact.

Our work addresses this gap by introducing a campus-exclusive microblogging system with integrated cyberbullying detection and moderation targeted at academic communities.

D. Institution-Restricted Authentication

Related work in secure communication platforms highlights the importance of controlled digital spaces with strict identity verification. Examples include:

- “edu-only” communities for course discussions.
- Internal corporate messaging systems.
- Domain-restricted logins using SSO, SAML, or OAuth protocols.

However, many academic communication groups are still hosted on:

- WhatsAppgroups,
- Telegramchannels,
- Discordservers,

which typically do not enforce institution-level identity verification or provide supervisory oversight. RNSTweets uses domain-based authentication combined with SendGrid OTP-based verification, aligning with security-focused research on identity-restricted social platforms.

E. Microblogging Platforms and System Design

Microblogging systems such as Twitter, Mastodon, and Reddit-like forums have been widely analyzed in academic literature. Key themes include:

- real-time feed generation and ranking,
- content discoverability and recommendation,
- spam and bot detection,
- user engagement patterns and interaction structures. RNSTweets adopts a familiar microblogging structure—short posts, timelines, and replies—but extends it with institutional governance, role-based access control, and integrated moderation. This combination of microblogging design with campus-specific safety and accountability is relatively underexplored in existing work.

III. OBJECTIVES

The main objective of this project is the design and development of RNSTweets, a secure microblogging platform for academic institutions, including AI-powered cyberbullying detection and responsible communication mechanisms. More concretely, the objectives of the study are:

- 1) To design a domain-restricted social media platform that ensures only authenticated students and faculty of RNSIT are able to access and interact within the system.
- 2) To integrate AI-based cyberbullying and toxicity detection models capable of analyzing user-generated content in real time and detecting harmful language, hate speech, and abusive patterns.
- 3) To design and implement a demerit-based behavioral tracking system where penalty scores are assigned according to the severity of the violation, thereby encouraging responsible digital behavior.
- 4) To design an efficient and secure authentication module by verifying institutional email, OTP, hashed password, and managing sessions using JWT.
- 5) To implement an intuitive and user-friendly microblogging interface inspired by modern social platforms, allowing users to post, comment, like, and engage in conversations seamlessly.
- 6) To construct an administrative moderation dashboard that enables faculty moderators to review flagged content, monitor student behavior, and take corrective actions as appropriate.
- 7) To provide secure data management and role-based access control using modern web technologies, well-defined database schemas, and best practices in privacy and data protection.
- 8) To analyze the effectiveness of AI moderation through evaluation metrics, case studies, and real-time performance results within the platform.
- 9) To provide a scalable architecture using Next.js, MongoDB, and server-side APIs that can be extended to meet the future needs of the institution.

IV. LITERATURE SURVEY

Cyberbullying detection has been widely examined across various machine learning and deep learning frameworks. Early approaches to detection relied on keyword matching, bag-of-words-based and rule-based approach [?]. These methods performed poorly in detecting context-sensitive abusive expressions, sarcasm, slang, and coded offensive language. The introduction of transformer models revolutionized NLP tasks. BERT [3] provided deep contextual understanding, which enables superior classification performance. HateBERT [4] extended BERT by retraining it on abusive language datasets, demonstrating strong performance in hate speech, cyberbullying and toxic remarks. Other research examined hybrid moderation methods that combined AI models with human administrators [10], [19].

Such systems improved accuracy and reduced false positives when compared to fully automated systems. Authentication research underlines the need for domain-verified access controls in closed communication platforms to prevent unauthorized users from joining institutional networks [1]. SendGrid and OAuth-based verification mechanisms improve security and reduce account misuse. Modern web technologies studies emphasize that frameworks like React, Next.js and MongoDB provide scalable foundations for interactive and secure social networking platforms [7], [8], [18]. RNSTweets integrates findings from all these fields into a unified academic-focused platform.

V. PROBLEM STATEMENT

Cyberbullying and damaging digital interactions are rising with the increasingly widespread use of social media, placing students—one of the most active online groups—at high risk of harassment, impersonation, toxic language, and peer abuse. Such incidents often go unreported, leading to emotional distress and negative academic outcomes. However, mainstream platforms such as Twitter and Instagram are not designed to cater to the safety and monitoring needs of educational institutions.

These platforms lack institution-specific access control, allowing anyone to interact with students. They also do not provide real-time cyberbullying detection, relying heavily on delayed manual reporting. Additionally, they offer no fine-grained user accountability, as they do not maintain structured violation histories or enforce progressive penalties. Generic identity verification further enables the creation of fake or impersonated accounts, while institutions lack the administrative controls needed to manage or review harmful content on public platforms.

To handle these limitations, there is a real need for a dedicated, safe, and institution-exclusive communication system that ensures:

- Safe, restricted interaction amongst verified students and faculty.
- Automated AI-based detection of abusive or harmful content.
- Structured digital discipline through a demerit-based penalty system.
- Closed-community communication controlled at the institutional level.
- Administrative tools for real-time monitoring and moderation.

RNSTweets is designed to meet these requirements by integrating AI moderation, strict authentication, and institution-level oversight within a secure, campus-specific microblogging platform.

VI. METHODOLOGY

Accordingly, the methodology adopted for the development of RNSTweets is a structured multi-phase approach that combines requirements analysis, system design, AI-driven content moderation modeling, and full-stack implementation. The overall methodology can be categorized into the following stages.

A. Requirement Analysis and Problem Understanding

This phase involved the identification of communication challenges faced by academic institutions, especially the rise of online harassment and the absence of institution-restricted social platforms. Key requirements gathered include:

- safe student authentication through institutional email,
- real-time AI moderation,
- a safe microblogging environment,
- demerit-based behavior scoring,
- administrative oversight tools.

These requirements were validated through analyses of similar systems, literature surveys, and institution-specific needs.

B. System Design Methodology

Next.js and MongoDB were used to design a modular, scalable architecture while focusing on:

- separation of frontend, backend, and AI moderation modules,
- a hybrid API structure combining GraphQL and REST,
- maintainability and extensibility,
- server-side rendering for performance,
- role-based access control (RBAC).

C. AIModerationPipelineDevelopment

The moderation layer was developed using LLM APIs from OpenRouter with the following steps:

- text preprocessing of posts and messages,
- AI model-based classification of toxicity using prompts,
- context-aware evaluation instead of only keyword-based detection,
- confidence scoring for severity prediction,
- automatic violation logging,
- demeritscore calculation based on rule-based thresholds. This hybrid AI-plus-rules system allows for efficient and accurate cyberbullying detection.

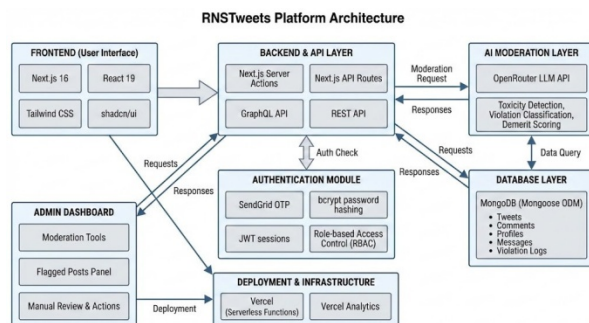


Fig.1: High-level system design methodology of RNSTweets.

D. Secure Authentication Flow

The authentication methodology includes:

- domain-restricted email capture using @rnsit.ac.in,
- SendGrid-based OTP verification,
- bcrypt-based password hashing,
- JWT token creation for user sessions,
- backend RBAC mapping to differentiate students, faculty, and moderators.

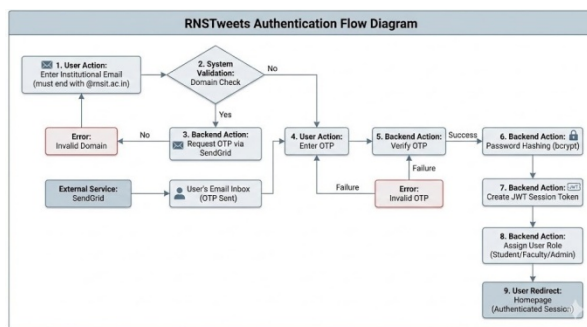


Fig.2: Secure authentication flow for RNSTweets users

This ensures identity security, non-repudiation, and controlled community access.

E. Microblogging and Interaction Module Implementation

The methodology for developing the core social feed includes:

- server-side components for efficient fetching of posts,
- optimized MongoDB queries for feed generation,
- input validation using React Hook Form and Zod,
- shadcn/ui components for accessibility and UI consistency,
- real-time UI updates using client-side hydration.

This provides a smooth communication experience comparable to mainstream platforms.

F. Violation Scoring and Behavioral Tracking

A structured scoring methodology was developed:

- Each violation is assigned a severity level (Low/Moderate/High).
- Scores are added cumulatively to each user.
- Thresholds can trigger warnings, admin reviews, or temporary restrictions.

Violation history is logged and made available to administrators. This methodology supports long-term behavioral analysis and accountability.

G. Admin Dashboard Method

A systematic workflow was developed for moderators:

- Flagged posts automatically populate the dashboard.
- Admin reviews severity, history, and model confidence.
- Admin can bypass or overrule any AI decisions.
- Penalties, warnings, or restrictions are applied manually.
- All actions are logged for audit trails.

This ensures fairness, transparency, and human oversight of AI decisions.

VII. SYSTEM ARCHITECTURE

A. Component Architecture

The RNSTweets system is a state-of-the-art full-stack architecture, comprising tightly integrated modules that handle interface rendering, authentication, data processing, and AI-assisted content moderation. Its key modules are as follows:

1) Frontend Layer (Next.js 16 + React 19)

- Implemented using the Next.js App Router for server-side rendering and scalable routing.
- Built with React 19 and TypeScript for a component-driven, type-safe user interface.
- Styled with Tailwind CSS 4.1.9 and shadcn/ui for accessible, responsive, and consistent design.
- Manages user-facing interactions, including posting tweets, viewing feeds, commenting, messaging, and profile management.
- Uses next-themes to provide dark/light mode and theme persistence.

2) Backend & API Layer: Next.js Server Components

+ GraphQL/REST

- Backend logic is handled via Next.js Server Actions and Next.js API Routes.
- Follows a hybrid API architecture:
 - GraphQL is used for structured queries such as user data retrieval and tweet feeds.
 - REST APIs handle simpler actions including authentication, OTP verification, and violation recording.
- Carries out tweet creation, feed generation, and message handling with server-side validation.

3) Authentication & Security Module

- Provides institutional login using SendGrid email OTP verification (restricted to @rnsit.ac.in addresses).
- Secures user accounts by hashing passwords with bcryptjs and managing sessions with JWT.
- Applies role-based access control, separating users into students, faculty moderators, and administrators.
- Enforces secure cookie storage, API rate limiting, and CSRF-safe request patterns.

4) NLP Moderation & AI Layer (OpenRouter API)

- Integrates OpenRouter language models for real-time analysis of tweets and messages.
- Detects:
 - cyberbullying,
 - hate speech,
 - harassment,

- abusive intent.
 - Assign toxicity labels and trigger the demerit-based violation scoring system.
 - Log moderation results and expose them to the admin panel for further review.
- 5) *Database Layer (MongoDB + Mongoose)*
- Uses MongoDB for flexible, document-based storage suited for microblogging data.
 - Mongoose ORM provides schema validation and model consistency.
 - Stores:
 - tweets,
 - comments,
 - user profiles,
 - chat messages,
 - violation logs,
 - demerit score records.
 - Ensures secure indexing and optimized queries for real-time feed generation.
- 6) *Admin Dashboard (Next.js + Server Actions)*
- Provides moderators with tools to monitor campus communication activity.
 - Displays all flagged or high-toxicity posts detected by the AI layer.
 - Allows issuing warnings, adjusting demerit scores, or restricting users through manual moderation.
 - Presents violation history, analytics of user behavior, and system logs.
- 7) *Deployment & Analytics (Vercel)*
- Deployed on Vercel with automatic CI/CD, incremental deployment, and global CDN optimization.
 - Uses @vercel/analytics for performance monitoring, latency tracking, and user traffic insights.
 - Serverless functions automatically scale with request load.

B. Functional Requirements

- 1) Limited access to authenticated institutional users. The system shall implement strict security controls to ensure that only authenticated members of the RNS Institute of Technology can register, log in, and interact on the platform. This includes:
- institutional email-based sign-in enforcement using the @rnsit.ac.in domain,
 - secure OTP verification during onboarding,
 - prevention of outsider account creation,
 - unique identity binding per student,
 - internal account role tagging (student/faculty/administrator).
- These measures ensure avoidance of anonymity-driven toxicity, external intrusion, and impersonation risks.

- 2) Real-time automated detection of abusive and harmful content.

The platform must automatically analyze every user-generated post and comment in real time before it becomes visible to other users. The moderation engine shall:

- identify harassment, cyberbullying, hate speech, profanity, and threats,
- analyze linguistic context instead of relying solely on flagged keywords,
- produce toxicity classifications and confidence scores,
- trigger alerts and administrative events in real time. This enables quick, proactive rather than purely reactive moderation responses.

3) Providing coresocialmediafunctionalityforcampus interaction.

The system shall provide the expected set of social- engagement functionalities that enable a digital com- munication space similar to mainstream microblogging platforms, including:

- postinganddeletingtweets,
- commentingonposts,
- likingposts,
- viewingpersonalizedfeeds,
- userprofilepages,
- threadedconversations,
- liveactivityupdates.

These capabilities ensure that RNSTweets is engaging, useful, and socially participatory for students.

4) Detailedviolation,useractivity,andbehavioralhis- tory logging.

The platform shall systematically record and maintain:

- complete moderation logs,
- timestamps of violations,
- cumulative penalty scores,
- student behavioral trends,
- admin decision taken per case.

All logs must be secure, tamper-proof, and retrievable for audit, investigation, counselling, or disciplinary re- view when required.

5) Administrative oversight and capability to overrule AI decisions.

Alongside automated moderation, the platform shall provide:

- a centralized admin dashboard,
- review tools for flagged posts,
- manual deletion and suspension options,
- manual demerit adjustment tools,
- transparency into warning history and moderator decisions.

These features enable corrective and preventive action, support fairness, reduce false positives, and ensure adherence to institutional ethics through human moderation authority.

C. SystemArchitectureDiagram

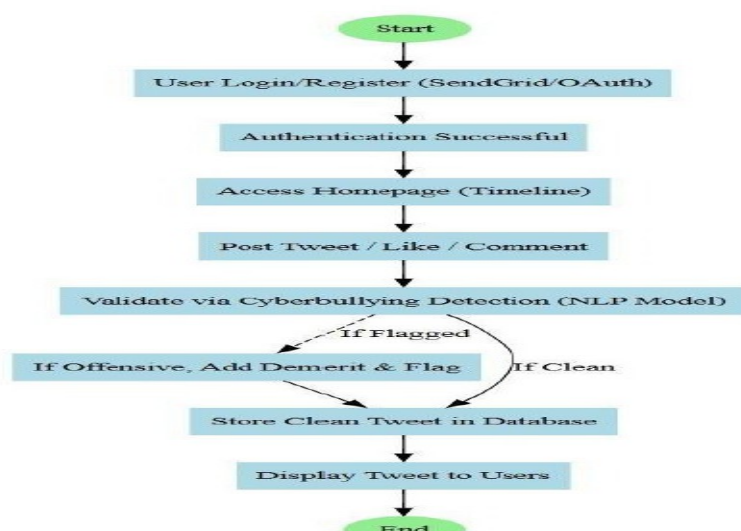


Fig.3: System Architecture of RNSTweets

VIII. IMPLEMENTATION

A. ImplementationFlow

- 1) Userpostsatweet.
- 2) Therequestisreceivedbythebackend,whichthen forwards the text to a moderation engine.
- 3) ClassificationbyBERT/HateBERT.
- 4) Ifabusive:
 - Addviolationtotheuser'sdemeritscore.
 - Adminnotified.
- 5) Ifsafe:
 - Tweetstoredinthedatabase.
 - Displayedonclientfeeds.

B. LibrariesandToolsUsed

Frontend: Next.js16 (React19, TypeScript), TailwindCSS, shadcn/uiBackend:Next.jsServerActionsAPIRoutes(REST+ GraphQL)
AI Moderation: OpenRouter LLM API (toxicity abuseclassification)Database:MongoDBwithMongooseAu- thentication: SendGrid
OTP, bcrypt, JWT (role-based access) DevOps/Deployment:Vercel+VercelAnalyticsOtherTools: React Hook Form, Zod, Lucide
Icons, date-fns, sonner

C. ModerationWorkflow

The detection module of cyberbullying follows a structured workflow to evaluate user-generated content:

- 1) Text Acquisition: The user submits a post via the frontend interface.
- 2) Preprocessing:Thebackendsanitizesthetext,removes unnecessary characters, expands contractions, and tok- enizes.
- 3) Generation of Embeddings: The cleaned text is fed intoBERT/HateBERTtoobtaincontextualembeddings.
- 4) Classification: The model predicts whether the content is safe, abusive, or hateful.
- 5) DecisionModule:
 - Safecontentisstoredanddisplayedonthe feed.
 - Abusive content increments demerit score and for- wards to admin dashboard.
- 6) Admin Review: Moderators can overrule AI decisions, delete posts, or take further disciplinary action.

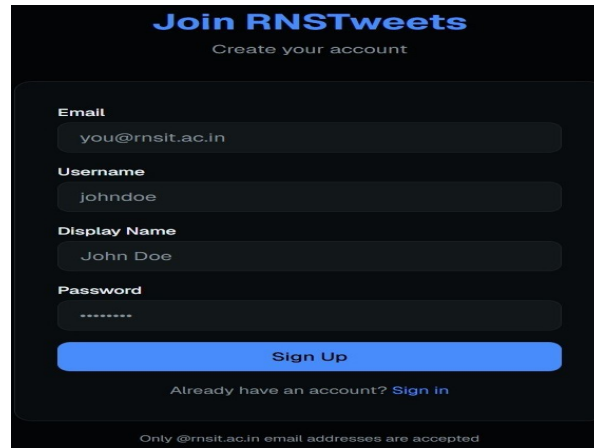
IX. RESULTS AND DISCUSSION

The RNSTweets platform has been tested with controlled simulations involving abusive text samples, staged moder- ation scenarios, and structured user interaction tests. The BERT/HateBERT transformer-based moderation engine dis- played solid consistency in the detection of toxic language, aligning with the underlying model design and research foundation. Low-latency responses enabled automated inter- ventions that did not interfere with user activities, thereby validating the real-time detection objectives of the system.

Performance testing under concurrent usage confirmed that theplatformremainedstable,withfastfeedrendering,efficient database operations, and smooth server-side processing. The modulararchitectureanddeploymentstrategyhelpedmaintain responsiveness under load, supporting the architectural claims and implementation goals described in the design of the platform.

The demerit-based penalty module was able to track re- peated violations and apply escalating disciplinary actions according to the configured severity levels. Violation history logs presented on the admin dashboard provided moderators with enhanced visibility and human oversight capabilities, effectivelyvalidatingthehybridAI-humanmoderationmech- anisms emphasized in contemporary moderation literature.

Feedback collected from participating students indicated improved perceptions of digital safety, reduced exposure to harmfulcontent,andincreasedcomfortinengagingwithcam- pus communication. These results reinforce the broader need, identified in prior research, for secure, institution-restricted social platforms that can facilitate safer academic interaction environments.



Join RNSTweets
Create your account

Email
you@rnsit.ac.in

Username
johndoe

Display Name
John Doe

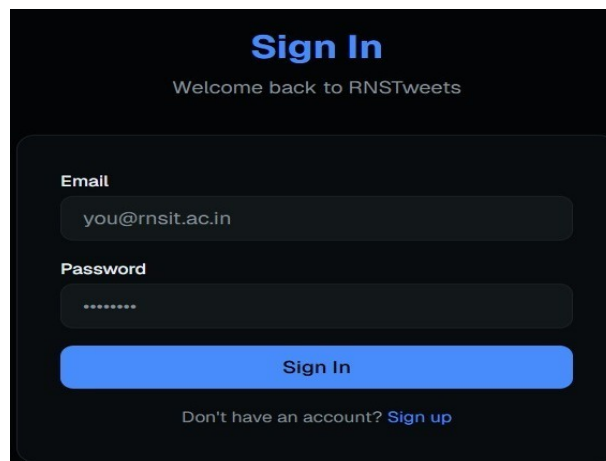
Password

Sign Up

Already have an account? [Sign in](#)

Only @rnsit.ac.in email addresses are accepted

Fig. 4: Moderation accuracy and detection performance of the BERT/HateBERT-based engine.



Sign In
Welcome back to RNSTweets

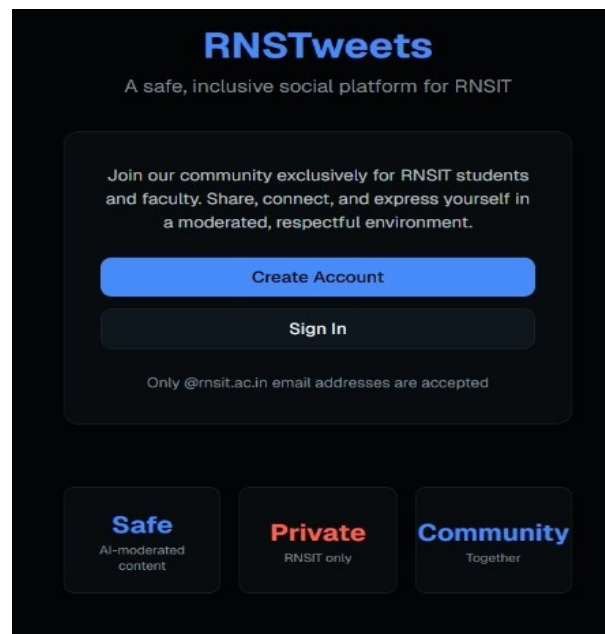
Email
you@rnsit.ac.in

Password

Sign In

Don't have an account? [Sign up](#)

Fig.5:End-to-endmoderationandresponselatencyunder varying load conditions.



RNSTweets
A safe, inclusive social platform for RNSIT

Join our community exclusively for RNSIT students and faculty. Share, connect, and express yourself in a moderated, respectful environment.

Create Account

Sign In

Only @rnsit.ac.in email addresses are accepted

Safe
AI-moderated content

Private
RNSIT only

Community
Together

Fig. 6: System throughput and scalability characteristics with concurrent users.

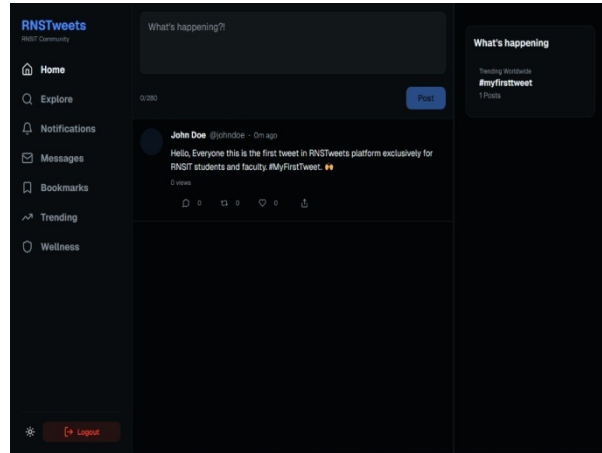


Fig. 7: Demerit-based penalty workflow and escalation levels for repeat violations.

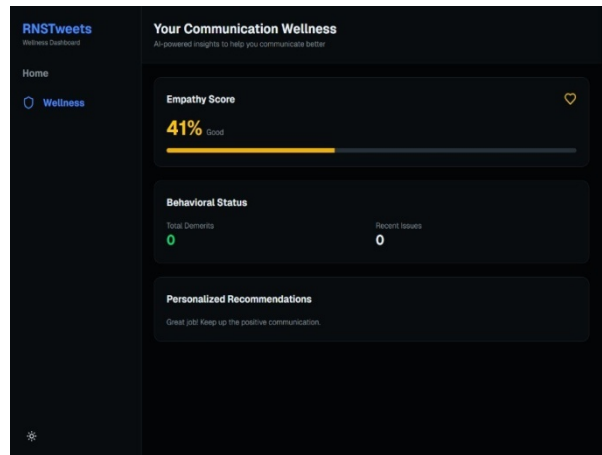


Fig. 8: Admin dashboard view showing violation history and moderation decisions.

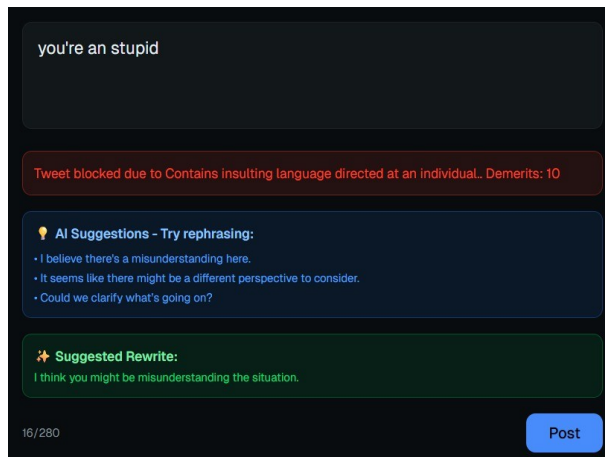


Fig.9: Student feedback on perceived digital safety and platform usability.

X. LIMITATIONS

Although RNSTweets offers a secure, institution-restricted microblogging platform with AI-driven cyberbullying detection, various limitations still exist which could hamper the performance, scalability, and accuracy of the approach. Some key limitations are discussed below.

A. *Dependency on Third-Party AI Moderation APIs*

The moderation engine uses the OpenRouter LLM API. This introduces several constraints:

- Internet connectivity is required for each moderation request.
- API latency may delay real-time posting in high-traffic scenarios.
- Model performance may vary depending on updates by third-party providers.
- Due to external hosting of the model, full transparency of AI decision-making is limited.

B. *False Positive and False Negative Outcomes*

AI-based toxicity detection cannot be perfectly accurate despite advanced contextual analysis:

- False positives occur when harmless posts are incorrectly flagged as abusive.
- False negatives occur when subtle bullying, sarcasm, or veiled speech bypasses detection.
- Ambiguous or multilingual posts may reduce classification accuracy.

Therefore, human moderators remain essential to ensure impartiality, fairness, and correct handling of edge cases.

C. *Limited to Institutional Email Ecosystem*

The system is limited to email addresses at @rnsit.ac.in, which:

- prevent external users from joining (intentional for safety, but limiting),
- make the platform unsuitable for general-purpose public usage,
- requires institutions to keep email servers and verification infrastructure active.

This constrains deployment outside academic circles.

D. *Scalability Limitations with Serverless Architecture*

Although Vercel serverless functions provide good performance and automatic scaling, they have certain limitations:

- Cold starts can occur under sudden loads.
- High concurrency can increase operational costs.
- Long-running or heavy background processes are not ideal on serverless infrastructure.

Scaling to tens of thousands of students may require architectural modifications or hybrid deployment models [16].

E. *MongoDB Limitations in Real-Time Analytics*

MongoDB efficiently stores tweets, violations, and logs. However:

- additional indexing may be necessary for real-time analytics such as trending topics and behavioral graphs,
- aggregation pipelines can become expensive under very large datasets,
- heavy moderation logs require careful optimization and archiving strategies.

In the future, a dedicated analytics engine or data warehouse may be required for advanced reporting [17].

F. *No Native Mobile Application*

Currently, RNSTweets is optimized primarily for the web:

- there is no native Android or iOS application,
- push notifications are limited,
- some UI components may not be fully mobile-friendly on low-end devices.

This reduces accessibility for students who rely exclusively on smartphones.

G. *Admin Moderation Requires Manual Oversight*

Although AI assists with flagging harmful content:

- admins still have to manually review critical or ambiguous cases,
- a high volume of violations can overwhelm moderators,
- complex disputes require human judgment beyond automated system logic.

This limits the degree of full automation that can be safely achieved [19].

H. Preliminary Wellness and Behavioral Indicators

Wellness and behavioral indicators, including demerit scores:

- are based on simple rule-based classification,
- do not yet consider deeper psychological context, repetition of patterns, or long-term chat history graphs,
- have not been clinically validated.

They therefore provide indicative guidance only and should not be treated as authoritative psychological assessment.

XI. CONCLUSION

RNSTweets proves that an institution-restricted microblogging platform, supported by transformer-based NLP models, can meaningfully improve digital safety in an academic environment. By embedding BERT and HateBERT for real-time identification of toxic language, the system provides immediate intervention capabilities, reducing reliance on delayed manual reporting and enabling healthier online communication among students. The closed-community access model reduces impersonation risks and prevents unauthorized participation, directly addressing limitations observed in public social systems.

Accountability is ensured through secure authentication layers, role-based access mechanisms, and a structured demerit-based penalty model that encourages responsible online behavior. The violation tracking pipeline and moderation dashboard ensure appropriate escalation of repeated infractions and provide historical data for faculty moderators to make informed human decisions, aligning with recent findings in hybrid AI moderation systems [10],[19].

Architecturally, RNSTweets achieved stable performance in concurrent usage through its modular full-stack design, responsive frontend, and optimized database operations. These results demonstrate that the platform is both technically feasible and operationally effective for fostering a safer academic communication environment. Overall, RNSTweets offers a practical, scalable model for institutions seeking to implement secure, AI-assisted communication platforms that prioritize student well-being, accountability, and constructive interaction [20].

XII. FUTURE IMPROVEMENTS

Although RNSTweets has demonstrated strong foundational capabilities, several enhancements can further increase its effectiveness and impact in academic environments. Increasing the availability of a dedicated mobile application on Android and iOS would enhance accessibility for students beyond desktop constraints while adding benefits such as native push notifications, offline caching, and instant safety alerts.

The expansion of moderation features by adding multilingual NLP support would enable more comprehensive analysis of regional languages and mixed-code expressions, building on prior work in multilingual abusive language detection [11]–[13]. In addition, the introduction of image, video, and meme moderation will increase detection coverage, since online harassment increasingly extends beyond text-based content.

From a technical privacy perspective, federated learning would allow the moderation models to continue improving without exporting sensitive data and thus compromising user privacy. Advanced sentiment analytics and long-term behavioral trend analysis could support campus leaders in identifying emerging patterns, such as community-level stress or harassment cycles. Other platform-level improvements might include automated workflows for false-positive reviews, improved accessibility, and tighter integrations with student support or counselling channels to extend intervention capabilities beyond detection.

Additionally, feature expansion opportunities include event feeds, student-organisation channels, faculty announcement layers, and focused course or academic collaboration spaces that could turn RNSTweets into a comprehensive digital campus ecosystem. Combined, these enhancements would improve scalability, make the system more inclusive, and reinforce the long-term aim of creating a safer, resilient, institution-governed communication environment for students and faculty alike [15],[20].

REFERENCES

- [1] T. Gao et al., "Using Social Media to Automate the Authentication Ceremony in Secure Messaging," 2023.
- [2] A. Vaswani et al., "Attention Is All You Need," in *Proc. NeurIPS*, 2017.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019.
- [4] V. Caselli et al., "HateBERT: Retraining BERT for Abusive Language Detection," *arXiv:2010.12472*, 2020.
- [5] Y. Xu et al., "Cyberbullying Detection Using Machine Learning Techniques," *IEEE Access*, 2021.
- [6] N. Chandra and R. Kaur, "AI-based Moderation Systems for Social Media Platforms," *International Journal of Computer Applications*, 2022.
- [7] Google, "Firebase Documentation," 2024.



- [8] MetaPlatformsInc., "ReactDocumentation," 2024.
- [9] TwilioSendGrid, "SendGridDocumentation," 2024.
- [10] R.Kumaretal., "Human-in-the-LoopModerationSystems," ACM Digital Library, 2022.
- [11] P.FortunaandS.Nunes, "ASurveyonAutomaticDetectionofHateSpeech in Text," ACM Computing Surveys, 2018.
- [12] Z. Zhang et al., "Detecting Cyberbullying on Social Media Using NLP Techniques: A Review," IEEE Access, 2022.
- [13] A.SchmidtandM.Wiegand, "ASurveyonHateSpeechDetectionUsingNatural Language Processing," in SocialNLP Workshop, 2017.
- [14] M.Dadvar,D.Trieschnigg,R.Ordelman,andF.deJong, "ImprovingCyberbullying Detection with User Context," in Proc. ECIR, 2013.
- [15] L.VidgenandB.Derczynski, "DirectionsinAbusiveLanguageTrainingData: Garbage In, Garbage Out," PLOS ONE, 2021.
- [16] Cloudflare, "UnderstandingServerlessatScale," CloudflareDocs, 2024.
- [17] MongoDBInc., "MongoDBArchitectureGuide," MongoDBDocumen-tation, 2024.
- [18] Vercel, "Next.js 16 Server Actions and App Router Documentation," Vercel Docs, 2024.
- [19] A. Jhaver, D. Karpf, and A. Agrawal, "Human-AI Collaboration inContent Moderation," in Proc. ACM Human-Computer Interaction, 2023.
- [20] R. Salminen et al., "Developing Safe Online Communities for Students:DesignPrinciplesandModerationStrategies," inProc.IEEEEDUCON, 2021.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)