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

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 XII Dec 2025- Available at www.ijraset.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,academicperformance,anddigitalsafety.Mainstream platformslackinstitution-specificcontrols,thusallowingharmful 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 mech- anism and an admin moderation dashboard. The platform usesa 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:*** ***Cyberbullyingdetection,BERT,HateBERT,Se- cure Social Media, NLP, Academic Communication, Moderation Systems.***

## I. INTRODUCTION

Socialmediaplaysasignificantroleinmoderncommunica- tion. It enables large-scale information sharing, collaboration, and interaction among users. However, these sites have like- wise become environments where cyberbullying, harassment, hate speech and abusive behavior manifest. Students, in par- ticular, are more susceptible to these dangers given the con- stant online presence, and peer pressure-driven interactions. Researchstudieshavedemonstratedthatcyberbullyingdirectly 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 institu- tions. They lack institution-level authentication and real-time moderation for student interactions and administrative over- sight. 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 inspiredbyTwitter,butwithcriticalsafetyfeaturesincluding:

- AI-poweredcyberbullyingdetection.
- Domain-basedauthenticationforstudentsandfaculty.
- Ademerit-basedviolationtrackingsystem.
- Adminmoderationdashboard.
- Securestorageandrole-basedaccesscontrol.

## II. RELATED WORK

### *A. Cyberbullying Detectionon Social Media*

There is a large body of research on detecting abusive or harmful content across platforms such as Twitter, YouTube, andInstagram.Existingapproachesrangefromtraditionalma- chinelearning to deeplearning and transformer-basedmodels.

1) *Machine Learning Models:* Early cyberbullying detec- tion systems employed traditional machine learning models such as Support Vector Machines (SVM), Na¨ıve Bayes, and LogisticRegression.Theseapproachesdependonhandcrafted features including:

- n-grams
- TF–IDFrepresentations
- sentimentpolarityfeatures

©IJRASET: All Rights are Reserved 

- profanityandhate-speechlexicons

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) *DeepLearningApproaches:*Recentdeeplearningarchi- tectureshavedemonstratedbetterperformanceforcyberbully- ing detection, especially on short and noisy texts:

- ConvolutionalNeuralNetworks(CNNs)forlocaltext feature extraction.
- LSTMs/GRUsformodelingsequentialcontextinuser posts.
- BidirectionalLSTMswithattentionmechanismstoem- phasize important words and phrases.

Thesemethodsarebetteratcapturingsemanticinformation compared to purely feature-engineered models.

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

- BERT
- RoBERTa
- DistilBERT
- HateBERT(fine-tunedfortoxicandabusivecontent)

Transformersprovidedeepcontextualrepresentationsandhave improved accuracy in classifying bullying, detecting toxic speech,andrecognizinghatespeech.Ourprojectisclosely aligned with this line of research by using AI-powered mod- eration through large language models and transformer-based APIs.

### *B. AI-AssistedContentModeration*

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:

- Highfalsepositivesduetolackofcontext.
- Easy evasion through spelling variations, code words, or sarcasm.

2) *Context-Aware AI Moderation:* Modern systems incor- porate:

- sentimentanalysisandtoxicityscoring,
- contextualembeddingsfromdeepneuralmodels,
- conversation-level modeling of reply chains and threads. LargeplatformssuchasTwitterandRedditemployautomated flaggingsystems,butthesearetypicallynottransparentand arenottailoredtospecificinstitutions.RNSTweetscontributes byapplyingAImoderationinaconstrainedacademicenviron- ment rather than an open global social network.

### *C. Digital Safety in Academic Environments*

Several studies emphasize the need for safe digital commu- nication 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 mecha- nisms with built-in safety controls and monitoring.
- ExistingLMStoolssuchasMoodle,Canvas,andGoogle Classroom are primarily focused on teaching workflows, not peer-to-peer social interaction.
- Cyberbullying on college campuses often goes unre- ported, leading to long-term psychological and academic impact.

Our work addresses this gap by introducing a campus- exclusivemicrobloggingsystemwithintegratedcyberbullying detection and moderation targeted at academic communities.

### *D. Institution-Restricted Authentication*

Relatedworkinsecurecommunicationplatformshighlights the importance of controlled digital spaces with strict identity verification. Examplesinclude:

- “edu-only”communitiesforcoursediscussions.
- Internalcorporatemessagingsystems.
- Domain-restricted logins using SSO, SAML, or OAuth protocols.

However,manyacademiccommunicationgroupsarestill hosted on:

- WhatsAppgroups,
- Telegramchannels,
- Discordservers,

which typically do not enforce institution-level identity ver- ification 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. MicrobloggingPlatformsandSystemDesign*

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

- real-timefeedgenerationandranking,
- contentdiscoverabilityandrecommendation,
- spamandbotdetection,
- user engagement patterns and interaction structures. RNSTweetsadoptsafamiliarmicrobloggingstructure–short posts,timelines,andreplies—butextendsitwithinstitutionalgovernance,role-basedaccesscontrol,andintegrated modera-tion. 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 de- velopment of RNSTweets, a secure microblogging platform foracademicinstitutions,includingAI-poweredcyberbullying detection and responsible communication mechanisms. More concretely, the objectives of the study are:

*1)* To design a domain-restricted social media platformthat ensures only authenticated students and faculty of RNSITareabletoaccessandinteractwithinthesystem.
*2)* To integrate AI-based cyberbullying and toxicity detec- tionmodelscapableofanalyzinguser-generatedcontent inrealtimeanddetectingharmfullanguage,hatespeech, and abusive patterns.
*3)* To design and implement a demerit-based behavioral tracking system where penalty scores are assigned ac- cording to the severity of the violation, thereby encour- aging 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 microblog- ging 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 con- tent, monitor student behavior, and take corrective ac- tions 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 perfor- mance results within the platform.
*9)* To provide a scalable architecture using Next.js, Mon- goDB,andserver-sideAPIsthatcanbeextendedtomeet the future needs of the institution.

## IV. LITERATURE SURVEY

Cyberbullying detection has been widely examined across variousmachinelearninganddeeplearningframeworks.Early approaches to detection relied on keyword matching, bagof- word-based and rule-based approach [?]. These methods per- formed poorly in detecting context-sensitive abusive expres- sions. sarcasm, slang, and coded offensive language. The introductionoftransformermodelsrevolutionizedNLP.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, cyberbul- lying and toxic remarks. Other research examined hybrid moderation methods that combined AI models with human administrators [10],[19].

Such systems improved accuracyand reduced false positives when compared to fully auto- mated systems. Authentication research underlines the need for domain-verified access controls in closed communication platforms to prevent unauthorized users from joining institu- tional networks [1]. SendGrid and OAuth-based verification mechanisms improve security and reduce account misuse. Modern web technologies studies emphasize that frameworks likeReact,Next.jsandMongoDBprovidescalablefoundations forinteractiveandsecuresocialnetworkingplatforms[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—oneofthemostactiveonlinegroups—athigherrisk of harassment, impersonation, toxic language, and peer abuse. Such incidents often go unreported, leading to emotional dis- tress and negative academic outcomes. However, mainstream platformssuchasTwitterandInstagramarenotdesigned to cater to the safety and monitoring needs of educational institutions.

These platforms lack institution-specific access control, al- lowing 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 donot maintain structured violation histories or enforce progressive penalties. Generic identity verification further enables the creation of fake or im- personated accounts, while institutions lack the administrative controlsneededtomanageorreviewharmfulcontentonpublic platforms.

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

- Safe, restricted interaction amongst verified students and faculty.
- AutomatedAI-baseddetectionofabusiveorharmful content.
- Structureddigitaldisciplinethroughademerit-based penalty system.
- Closed-communitycommunicationcontrolledatthein- stitutional level.
- Administrativetoolsforreal-timemonitoringandmod- eration.

RNSTweets is designed to meet these requirements by in- tegrating AI moderation, strict authentication, and institution-leveloversightwithinasecure,campus-specificmicroblogging platform.

## VI. METHODOLOGY

Accordingly, the methodology adopted for the develop- mentofRNSTweetsisastructuredmulti-phaseapproach thatcombinesrequirementsanalysis,systemdesign,AI-driven content moderation modeling, and full-stack implementation. Theoverallmethodologycanbecategorizedintothefollowing stages.

### *A. RequirementAnalysisandProblemUnderstanding*

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:

- safestudentauthenticationthroughinstitutionalemail,
- real-timeAImoderation,
- asafemicrobloggingenvironment,
- demerit-basedbehaviorscoring,
- administrativeoversighttools.

Theserequirementswerevalidatedthroughanalysesofsimilar systems, literature surveys, and institution-specific needs.

### *B. SystemDesignMethodology*

Next.jsandMongoDBwereusedtodesignamodular, scalable architecture while focusing on:

- separationoffrontend,backend,andAImoderationmod- ules,
- ahybridAPIstructurecombiningGraphQLandREST,
- maintainabilityandextensibility,
- server-siderenderingforperformance,
- role-basedaccesscontrol(RBAC).

### *C. AIModerationPipelineDevelopment*

ThemoderationlayerwasdevelopedusingLLMAPIsfrom OpenRouter with the following steps:

- textpreprocessingofpostsandmessages,
- AImodel-basedclassificationoftoxicityusingprompts,
- context-awareevaluationinsteadofonlykeyword-based detection,
- confidencescoringforseverityprediction,
- automaticviolationlogging,
- demeritscorecalculationbasedonrule-basedthresholds. This hybrid AI-plus-rules system allows for efficient and accurate cyberbullying detection.

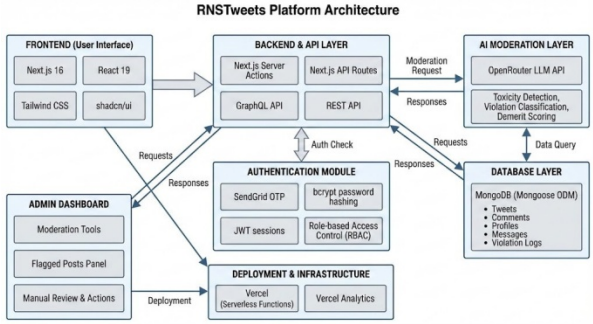



Fig.1:High-levelsystemdesignmethodologyofRNSTweets.

### *D. Secure Authentication Flow*

Theauthenticationmethodologyincludes:

- domain-restrictedemailcaptureusing@rnsit.ac.in,
- SendGrid-basedOTPverification,
- bcrypt-basedpasswordhashing,
- JWTtokencreationforusersessions,
- backendRBACmappingtodifferentiatestudents,faculty, and moderators.

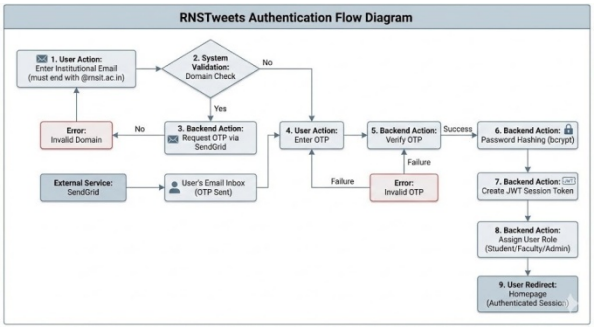



Fig.2:SecureauthenticationflowforRNSTweetsusers

Thisensuresidentitysecurity,non-repudiation,andcon- trolled community access.

### *E. MicrobloggingandInteractionModuleImplementation*

Themethodologyfordevelopingthecoresocialfeedin- cluded:

- server-sidecomponentsforefficientfetchingofposts,
- optimizedMongoDBqueriesforfeedgeneration,
- inputvalidationusingReactHookFormandZod,
- shadcn/uicomponentsforaccessibilityandUIconsis- tency,
- real-timeUIupdatesusingclient-sidehydration.

Thisprovidesasmoothcommunicationexperiencecomparable to mainstream platforms.

*F. ViolationScoringandBehavioralTracking*

Astructuredscoringmethodologywasdeveloped:

- Eachviolationisassignedaseveritylevel (Low/Moderate/High).
- Scoresareaddedcumulativelytoeachuser.
- Thresholds can trigger warnings, admin reviews, or tem- porary restrictions.

Violation history is logged and made available to administra- tors.Thismethodologysupportslong-termbehavioralanalysis and accountability.

*G. AdminDashboardMethod*

Asystematicworkflowwasdevelopedformoderators:

- Flaggedpostsautomaticallypopulatethedashboard.
- Adminreviewsseverity,history,andmodelconfidence.
- Admincanbypassoroverruleany AIdecisions.
- Penalties,warnings,orrestrictionsareapplied manually.
- Allactionsareloggedforaudittrails.

Thisensuresfairness,transparency,andhumanoversightof AI decisions.

## VII.SYSTEM ARCHITECTURE

*A. Component Architecture*

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

*1) FrontendLayer(Next.js16+React19)*

- 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,viewingfeeds,commenting,messaging,and profile management.
- Uses next-themesto provide dark/light mode and theme persistence.

*2) Backend&APILayer:Next.jsServerComponents*

+GraphQL/REST

- BackendlogicishandledviaNext.jsServerActions and Next.js API Routes.
- FollowsahybridAPIarchitecture:
  - 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&SecurityModule*

- Provides institutional login using SendGrid email OTP verification (restricted to @rnsit.ac.in ad- dresses).
- Securesuseraccountsbyhashingpasswordswithbcryptjsandmanagingsessionswith JWT.
- Applies role-based access control, separating users into students, faculty moderators, and administra- tors.
- Enforcessecurecookiestorage,APIratelimiting, and CSRF-safe request patterns.

*4) NLPModeration&AILayer(OpenRouterAPI)*

- IntegratesOpenRouterlanguagemodelsforreal- time analysis of tweets and messages.
- Detects:
  - cyberbullying,
  - hatespeech,
  - harassment,

- abusive intent.
- Assignstoxicitylabelsandtriggersthedemerit- based violation scoring system.
- Logsmoderationresultsandexposesthemtothe admin panel for further review.

*5) DatabaseLayer(MongoDB+Mongoose)*

- UsesMongoDBforflexible,document-basedstor- age suited for microblogging data.
- MongooseORMprovidesschemavalidationand model consistency.
- Stores:
  - tweets,
  - comments,
  - userprofiles,
  - chatmessages,
  - violationlogs,
  - demeritscorerecords.
- Ensures secure indexing and optimized queries for real-time feed generation.

*6) AdminDashboard(Next.js+ServerActions)*

- Provides moderators with tools to monitor campus communication activity.
- Displays all flagged or high-toxicity posts detected by the AI layer.
- Allowsissuingwarnings, adjustingdemerit scores, or restricting users through manual moderation.
- Presentsviolationhistory,analyticsofuserbehavior, and system logs.

*7) Deployment&Analytics(Vercel)*

- Deployed on Vercel with automatic CI/CD, incre- mental deployment, and global CDN optimization.
- Uses@vercel/analytics forperformance monitoring, latency tracking, and user traffic in- sights.
- Serverless functions automatically scale with re- quest load.

*B. Functional Requirements*

*1)* Limited access to authenticated institutional users. Thesystemshallimplementstrictsecuritycontrolsto ensurethatonlyauthenticatedmembersoftheRNS 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,
- secureOTPverificationduringonboarding,
- preventionofoutsideraccountcreation,
- uniqueidentitybindingperstudent,
- internal account role tagging (student/faculty/administrator).

Thesemeasuresensureavoidanceofanonymity-driven toxicity, external intrusion, and impersonation risks.

*2)* Real-time automated detection of abusive and harm- ful 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:

- identifyharassment,cyberbullying,hatespeech, profanity, and threats,
- analyzelinguisticcontextinsteadofrelyingsolely on flagged keywords,
- producetoxicityclassificationsandconfidence scores,
- trigger alerts and administrative events in real time. This enables quick, proactive rather than purely reactive moderation responses.

*3)* Providingcoresocialmediafunctionalityforcampus 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.

ThesecapabilitiesensurethatRNSTweetsisengaging, useful, and socially participatory for students.

*4)* Detailedviolation,useractivity,andbehavioralhis- tory logging.

Theplatformshallsystematicallyrecordandmaintain:

- completemoderationlogs,
- timestampsofviolations,
- cumulativepenaltyscores,
- studentbehavioraltrends,
- admindecisionstakenpercase.

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:

- acentralizedadmindashboard,
- reviewtoolsforflaggedposts,
- manualdeletionandsuspensionoptions,
- manualdemeritadjustmenttools,
- transparency into warning history and moderator decisions.

These features enable corrective and preventive action, support fairness, reduce false positives, and ensure ad- herencetoinstitutionalethicsthroughhumanmoderation authority.

### *C. SystemArchitectureDiagram*

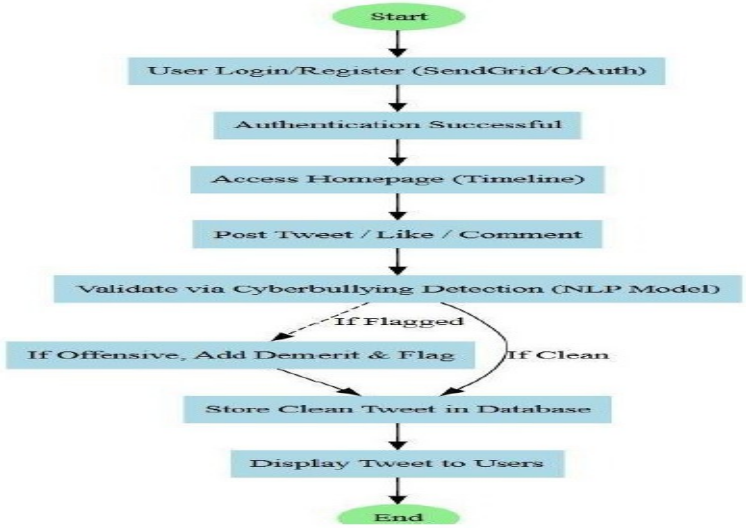



Fig.3:SystemArchitectureofRNSTweets

## 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:

- Safecontentisstoredanddisplayedonthefeed.
- 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.




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

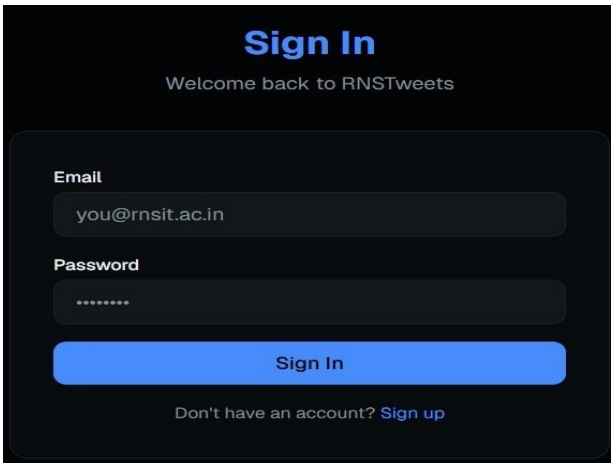



Fig.5:End-to-endmoderationandresponselatencyunder varying load conditions.

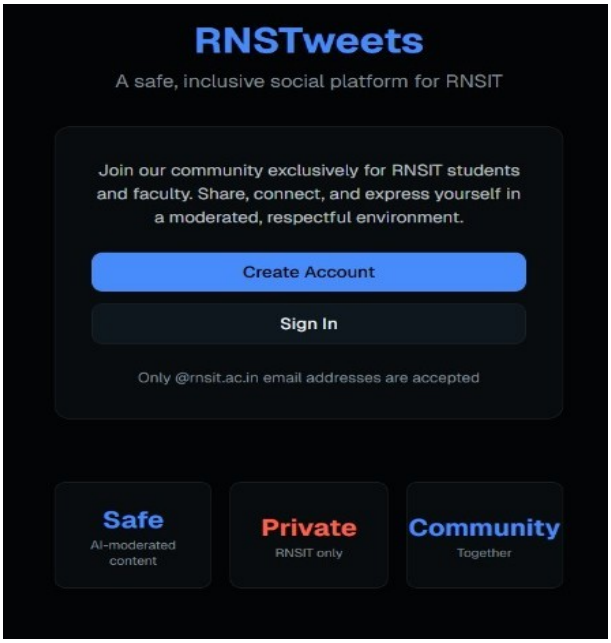



Fig. 6: System throughput and scalability characteristics with concurrent users.

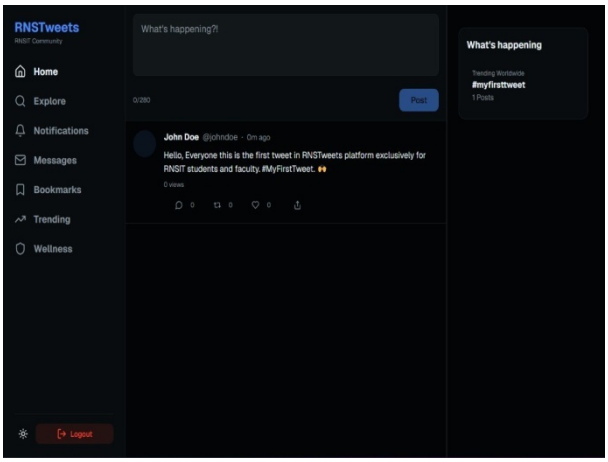



Fig. 7: Demerit-based penalty workflow and escalation levels for repeat violations.

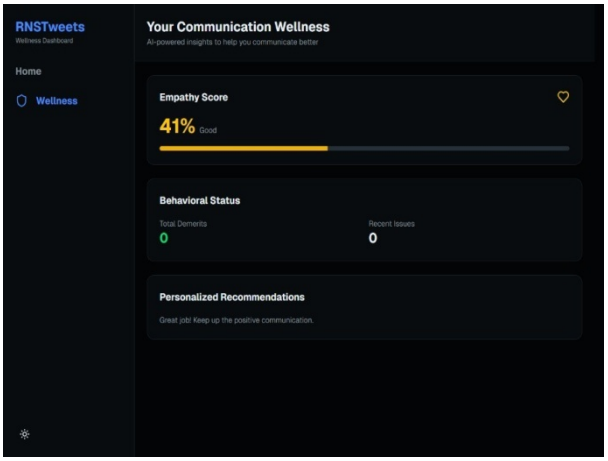



Fig. 8: Admin dashboard view showing violation history and moderation decisions.

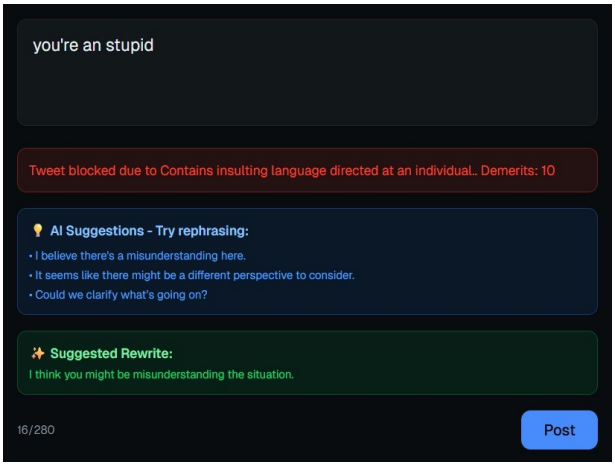



Fig.9:Studentfeedbackonperceiveddigitalsafetyandplatform usability.

## X. LIMITATIONS

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

### A. DependencyonThird-PartyAIModerationAPIs

ThemoderationengineusestheOpenRouterLLMAPI.This introduces several constraints:

- Internetconnectivityisrequiredforeachmoderation request.
- APIlatencymaydelayreal-timepostinginhigh-traffic scenarios.
- Modelperformancemayvarydependingonupdatesby third-party providers.
- Duetoexternalhostingofthemodel,fulltransparency of AI decision-making is limited.

### B. FalsePositiveandFalseNegativeOutcomes

AI-basedtoxicitydetectioncannotbeperfectlyaccurate despite advanced contextual analysis:

- Falsepositivesoccurwhenharmlesspostsareincorrectly flagged as abusive.
- Falsenegativesoccurwhensubtlebullying,sarcasm,or veiled speech bypasses detection.
- Ambiguousormultilingualpostsmayreduceclassifica- tion accuracy.

Therefore,humanmoderatorsremainessentialtoensureim- partiality, fairness, and correct handling of edge cases.

### C. LimitedtoInstitutionalEmailEcosystem

Thesystemislimitedtoemailaddressesat@rnsit.ac.in, which:

- preventsexternalusersfromjoining(intentionalforsafety, but limiting),
- makestheplatformunsuitableforgeneral-purposepublic usage,
- requiresinstitutionstokeepemailserversandverification infrastructure active.

Thisconstrainsdeploymentoutsideacademiccircles.

### D. ScalabilityLimitationswithServerlessArchitecture

AlthoughVercelserverlessfunctionsprovidegoodperfor- mance and automatic scaling, they have certain limitations:

- Coldstartscanoccurundersuddenloads.
- Highconcurrencycanincreaseoperationalcosts.
- Long-runningorheavybackgroundprocessesarenot ideal on serverless infrastructure.

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

### E. MongoDBLimitationsinReal-TimeAnalytics

MongoDBefficientlystorestweets,violations,andlogs. However:

- additional indexing may be necessary for real-time ana- lytics such as trending topics and behavioral graphs,
- aggregationpipelinescanbecomeexpensiveundervery large datasets,
- heavymoderationlogsrequirecarefuloptimizationand archiving strategies.

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

### F. NoNativeMobileApplication

Currently,RNSTweetsisoptimizedprimarilyfortheweb:

- thereisnonativeAndroidoriOSapplication,
- pushnotificationsarelimited,
- someUIcomponentsmaynotbefullymobile-friendly on low-end devices.

Thisreducesaccessibilityforstudentswhorelyexclusivelyon smartphones.

### G. AdminModerationRequiresManualOversight

AlthoughAIassistswithflaggingharmfulcontent:

- admins still have to manually review critical or ambigu- ous cases,
- ahighvolumeofviolationscanoverwhelmmoderators,
- complex disputes require human judgment beyond auto- mated system logic.

Thislimitsthedegreeoffullautomationthatcanbesafely achieved [19].

### *H. PreliminaryWellnessandBehavioralIndicators*

Wellnessandbehavioralindicators,includingdemerit scores:

- arebasedonsimplerule-basedclassification,
- do not yet consider deeper psychological context, repeti- tion of patterns, or long-term chat history graphs,
- havenotbeenclinicallyvalidated.

Theythereforeprovideindicativeguidanceonlyandshould not be treated as authoritative psychological assessment.

## XI. CONCLUSION

RNSTweets proves that an institution-restricted microblog- ging platform, supported by transformer-based NLP models, can meaningfully improve digital safety in an academic envi- ronment. By embedding BERT and HateBERT for real-time identification of toxic language, the system provides imme- diate intervention capabilities, reducing reliance on delayed manualreportingandenablinghealthieronlinecommunication among students. The closed-community access model reduces impersonation risks and prevents unauthorized participation, directly addressing limitations observed in public social sys- tems.

Accountabilityisensuredthroughsecureauthenticationlay- ers, role-based access mechanisms, and a structured demerit-basedpenaltymodelthatencouragesresponsibleonlinebehav- ior. The violation tracking pipeline and moderation dashboard ensure appropriate escalation of repeated infractions and pro- vide 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 performancein concurrent usage through its modular full-stack design, responsivefrontend,andoptimizeddatabaseoperations.These results demonstrate that the platform is both technically feasi- ble 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 studentwell-being,accountability,andconstructiveinteraction [20].

## XII.FUTURE IMPROVEMENTS

AlthoughRNSTweetshasdemonstratedstrongfoundational capabilities, several enhancements can further increase its effectivenessandimpactinacademicenvironments.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 supportwouldenablemorecomprehensiveanalysisofregional languagesandmixed-codeexpressions,buildingonpriorwork in multilingual abusive language detection [11]–[13]. In addi- tion, 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 behav- ioral trend analysis could support campus leaders in identi- fying emerging patterns, such as community-level stress or harassment cycles. Other platform-level improvements might include automated workflows for false-positive reviews, im- proved accessibility, and tighter integrations with student- support or counselling channels to extend intervention capa- bilities beyond detection.

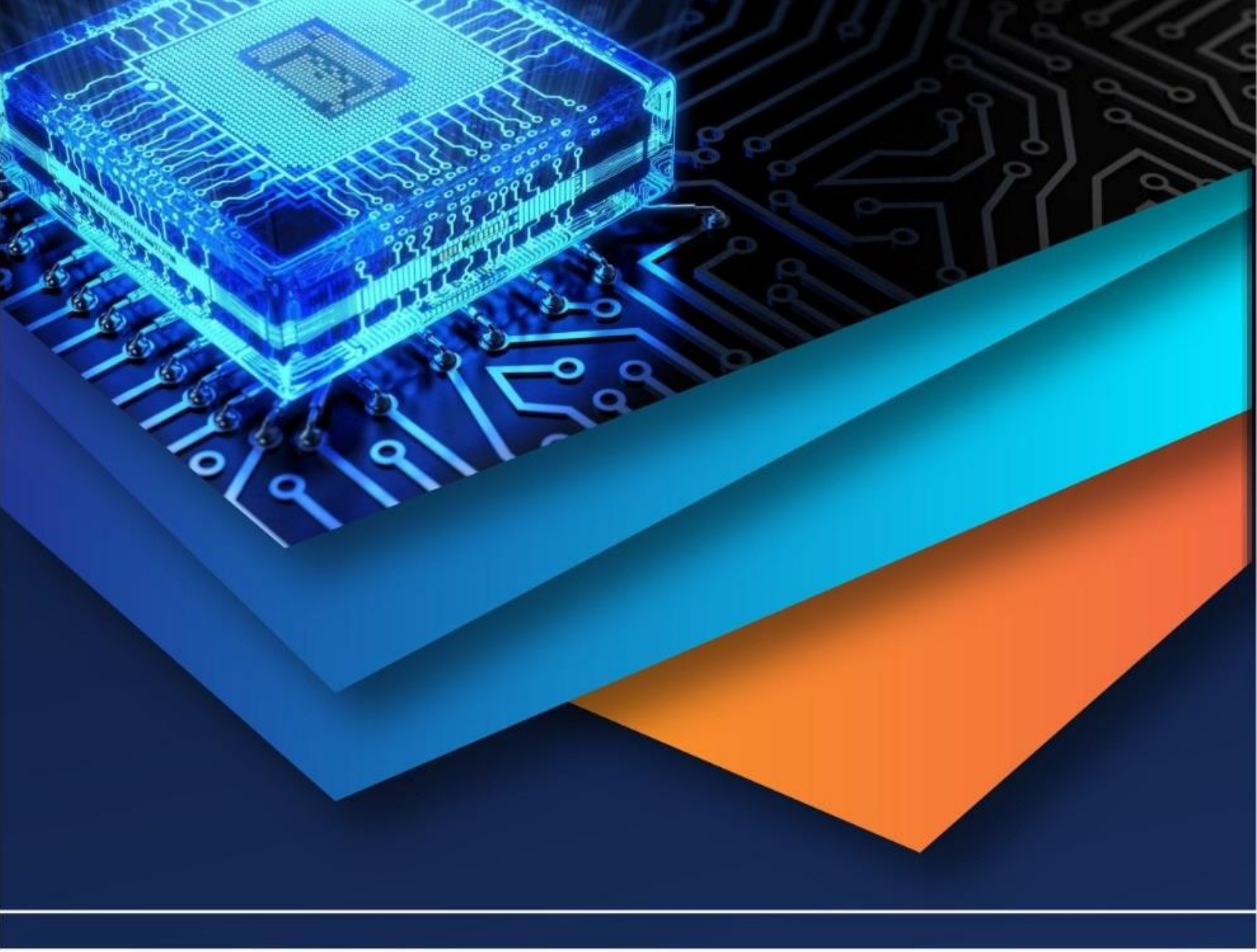
Additionally, feature expansion opportunities include event feeds, student-organisation channels, faculty announcement layers, and focused course or academic collaboration spaces that could turn RNSTweets into a comprehensive digital cam-pusecosystem.Combined,theseenhancementswouldimprove scalability,makethesystemmoreinclusive,andreinforce the long-term aim of creating a safer, resilient, institution- governedcommunicationenvironmentforstudentsandfaculty alike [15],[20].

## REFERENCES

[1] T. Gao et al., “Using Social Media to Automate the AuthenticationCeremony in Secure Messaging,” 2023.
[2] A.Vaswanietal.,“AttentionIsAllYouNeed,”inProc.NeurIPS,2017.
[3] J.Devlin,M.-W.Chang,K.Lee,andK.Toutanova,“BERT:Pre-trainingof Deep Bidirectional Transformers for Language Understanding,” inProc. NAACL-HLT, 2019.
[4] V. Caselli et al., “HateBERT: Retraining BERT for Abusive LanguageDetection,” arXiv:2010.12472, 2020.
[5] Y. Xu et al., “Cyberbullying Detection Using Machine Learning Tech-niques,” IEEE Access, 2021.
[6] N. Chandra and R. Kaur, “AI-based Moderation Systems for SocialMedia Platforms,” International Journal of Computer Applications,2022.
[7] Google,“FirebaseDocumentation,”2024.

[8] MetaPlatformsInc.,"ReactDocumentation,"2024.
[9] TwilioSendGrid,"SendGridDocumentation,"2024.
[10] R.Kumaretal.,"Human-in-the-LoopModerationSystems,"ACMDigital Library, 2022.
[11] P.FortunaandS.Nunes,"ASurveyonAutomaticDetectionofHateSpeech in Text," ACM Computing Surveys, 2018.
[12] Z. Zhang et al., "Detecting Cyberbullying on Social Media Using NLPTechniques: 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.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



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)