Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278 - 5140
Volume 14 Issue 01,2025

SafeSpace AI: Intelligent Content Moderation Platform

[1]Prof. Anmol Budhewar, [2]Tushar Mohan Agrawal, [3]Rohit Pravin Shinde, [4]Reeshoo Rameshpratap Yadav, [5]Roshan Kashinath Deore
[1]Prof, Computer Engineering,Sandip Institute Of Technology and Research Center, Nashik (SITRC)
[2 3 4 5]Student, Computer Engineering,Sandip Institute Of Technology and Research Center, Nashik (SITRC)
Email: anmolsbudhewar@gmail.com, tusharagrawal271@gmail.com, rohitshinde7922@gmail.com, reeshooyadav4@gmail.com, roshandeore702@gmail.com

Peer Review Information

Submission: 1 Sept 2025

Revision: 28 Sept 2025

Acceptance: 12 Oct 2025

Keywords

Artificial Intelligence, Content Moderation, Natural Language Processing, Computer Vision, Reputation System, Social Media Safety, Web Application Development

Abstract

Social media has transformed into a primary medium for communication and content sharing, yet its open nature has simultaneously created avenues for harmful material to spread unchecked. Instances of abusive language, offensive imagery, political misinformation, sexually explicit posts, and targeted harassment have become increasingly common, often posing serious risks to vulnerable users. Existing moderation systems are limited in scope — many lack transparency, fail to support multilingual content, and do not provide effective mechanisms for repeated violators.
This project proposes the development of an AI-based content moderation system that combines natural language processing, image analysis, and reputation scoring to ensure a safe and reliable digital environment. The system is designed to analyze both text and images in real time, identify harmful content, and provide clear explanations when a post is blocked or flagged. An integrated reputation model assigns scores to users based on their behavior, rewarding positive contributions while progressively restricting harmful accounts. Administrators are equipped with a dedicated dashboard to review flagged content, monitor user activity, and configure moderation rules.
To demonstrate the effectiveness of the solution, a prototype social media platform is implemented where posting, commenting, and interactions are actively moderated. The system operates on a lightweight technical stack (React.js, Node.js, Python Flask, and SQLite) to ensure compatibility with student laptops in an offline setup. This ensures accessibility for academic evaluation while aligning with real-world needs for safer online spaces.
In practical terms, the project addresses both academic significance—showcasing the application of AI/ML and web technologies—and societal impact by promoting digital safety and responsible online behavior.

Introduction

1. Background and Motivation

Over the past decade, social media has grown from being a simple communication tool into one of the most powerful mediums for information exchange, networking, and public engagement. Platforms such as Facebook, Instagram, Twitter, and numerous regional networks have created digital communities that connect billions of people in real time. This rapid expansion has not only transformed how individuals interact but has also influenced politics, commerce, education, and culture on a global scale.
However, this growth has also resulted in the widespread circulation of harmful digital content. Abusive language, political propaganda,

sexually explicit material, cyberbullying, deepfake videos, and misinformation are now common challenges faced by online communities. Such harmful content spreads faster than traditional monitoring mechanisms can handle, creating risks for individuals and undermining trust in online platforms. In particular, young users often become both consumers and producers of such content without fully realizing the consequences.

1.2 Importance of Digital Safety and Responsible Online Spaces

The increasing dependence on digital platforms highlights the urgent need for safe and responsible online environments. Digital safety is not merely about removing offensive posts but also about creating a system where users feel protected, informed, and respected. A secure platform ensures that communities can thrive without exposure to psychological harm, exploitation, or manipulation.

In practical terms, digital safety involves three critical elements:

- **Prevention** of harmful content before it reaches a wide audience.
- **Transparency** in how moderation decisions are made and communicated to users.
- **Accountability** by discouraging repeat offenders and promoting positive behavior.

Responsible online spaces also play a key role in shaping social attitudes. When platforms implement clear rules and intelligent moderation, they encourage respectful communication, reduce harassment, and safeguard vulnerable users such as teenagers. By striking a balance between free expression and community protection, content moderation systems can ensure that social media remains a constructive tool rather than a source of harm.

The motivation behind this project arises from these realities: while social media continues to grow in influence, its unchecked misuse demands the creation of intelligent moderation systems. This project seeks to address that need by providing a transparent, AI-driven platform that improves user trust and ensures digital safety for all.

Literature Survey

Introduction

A literature survey provides the foundation for understanding prior work in a specific research area. For this project, the focus is on **content moderation systems, natural language processing, image analysis, and user reputation management**. Reviewing existing research and tools highlights both the advancements made in the field and the gaps that remain unaddressed. This section discusses academic studies, industrial approaches, and the identified research gaps that motivate the proposed system.

Review of Related Work

1. Text-Based Moderation Studies

Several research efforts have attempted to address harmful text detection using machine learning. Early studies relied on keyword filtering, which, while simple, suffered from high false positives and an inability to understand context. More recent studies introduced **deep**

2. learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for hate speech detection. A 2019 study demonstrated that **Bidirectional LSTM models** performed better than traditional classifiers in detecting cyberbullying on social media. However, these models required large datasets and significant computational resources, limiting their applicability in lightweight systems such as student projects.

3. Multilingual and Regional Language Moderation

In multilingual societies, moderation systems must go beyond English. Research using models like **MuRIL (Multilingual Representations for Indian Languages)** and **IndicBERT** has shown promise in handling Indian languages. A 2020 study on multilingual abuse detection indicated that transformer-based models significantly outperformed rule-based approaches. Still, these models require fine-tuning on domain-specific datasets, which are not always available for regional slang and mixed languages like "Hinglish." This limitation reveals the need for practical, rule-based + lightweight AI hybrid solutions for academic contexts.

4. Image and Multimedia Moderation

On the visual side, several tools have been developed for detecting explicit imagery. Open-source libraries such as **NSFW.js** and academic frameworks using **OpenCV** have been employed for nudity detection and face recognition. While effective at filtering basic explicit content, these approaches often struggle with sophisticated manipulations like deepfakes. In 2018, a study highlighted the increasing risk of deepfake media in political campaigns, noting that even advanced detection systems had limited success in real time. This underscores the importance of implementing at least basic visual moderation in prototypes to cover the most obvious harmful imagery.

5. Reputation and Behavior Tracking Systems
Beyond content, researchers have also examined user behavior. Reputation-based scoring has been widely studied in online communities such as Stack Overflow, Reddit, and gaming platforms. A 2017 survey on online trust systems emphasized that dynamic scoring encourages accountability and deters repeated violations. However, most implementations remain specific to their platforms and are not integrated with AI-based content moderation. This gap motivates the inclusion of a **unified reputation mechanism** in the proposed system.

Gap Analysis
From the above review, several clear gaps emerge:

- **Transparency Deficit:** Most systems fail to explain why content was flagged, leading to user dissatisfaction.
- **Multilingual Weakness:** Advanced models exist but are resource-heavy and impractical for academic or lightweight deployment.
- **Limited Age-Based Safety:** Few systems focus on protecting minors from posting or consuming unsafe material.
- **Inadequate Behavior Tracking:** Reputation or scoring systems are rarely integrated with moderation engines.
- **Prototype Feasibility:** Commercial tools are powerful but not replicable in an academic setting, leaving a gap for lightweight, demonstrable systems.

Objectives of the Project
Objective 1: Core Content Moderation System
The central objective is to build an intelligent moderation system capable of identifying and managing harmful digital content. This involves multiple layers of analysis:
Objective 2: User Management & Reputation System
Moderation is incomplete without user accountability. This objective introduces mechanisms for user registration, authentication, and behavior-based reputation scoring.
Objective 3: Administrative Control Panel
While AI-driven moderation is powerful, human oversight is essential for fairness and adaptability. This objective focuses on equipping administrators with tools for control and monitoring.
Objective 4: Demonstration Social Media Platform
To showcase the system in a realistic setting, a prototype social media platform will be implemented.

Proposed System
The proposed system, *AI-Based Content Moderation System (Project Clean)*, is designed as an integrated platform that combines text analysis, image filtering, reputation scoring, and administrative oversight into a single functional prototype. The system operates as a lightweight, offline-capable web application, making it practical for both academic demonstration and future scalability.
Short Overview of How the System Works
At the core of the system lies the **content moderation engine**, which continuously analyzes user-generated content in real time. When a user attempts to post text or upload an image, the system first processes the input through the moderation module:

- **Text content** is passed through natural language processing (NLP) models and rule-based filters to detect abusive language, hate speech, propaganda, or other forms of harmful expression.
- **Image content** is checked using a lightweight computer vision model capable of identifying explicit or inappropriate visuals.

Once the analysis is complete, the system either approves the content, blocks it, or flags it for further review by administrators.
Every user in the system has a **reputation score** that is dynamically adjusted based on their behavior. Positive interactions such as respectful posting increase the score, while repeated violations reduce it. Low-scoring users face restrictions such as limited content reach or account suspension.
The **administrative dashboard** provides moderators with full control, allowing them to review flagged content, manage users, and configure moderation policies. To validate these features, the system includes a **demo social media platform** where posting, commenting, and user interaction can be tested under real-time moderation conditions.

Outcomes
The proposed *AI-Based Content Moderation System (Project Clean)* is expected to produce both **technical outcomes** and **academic/social outcomes**. These outcomes validate the feasibility of the project, its alignment with defined objectives, and its usefulness as a demonstrable prototype.

Technical Outcomes

1. **Functional Moderation Engine**
   - Working text and image moderation modules that detect and block harmful content in real time.

   - Clear explanations provided to users for every blocked or flagged post.
2. **Reputation-Based User Management**
   - Dynamic scoring system to promote positive behavior and penalize repeat violators.
   - Tier-based enforcement (trusted, restricted, suspended) demonstrated effectively.
3. **Administrative Dashboard**
   - Fully operational interface for moderators to review flagged content, manage users, adjust rules, and analyze trends.
   - Audit logs and analytics available for evaluation and reporting.
4. **Demo Social Media Platform**
   - End-to-end prototype simulating a real platform with posts, comments, feeds, and moderation feedback.
   - Offline-ready execution on student laptops using lightweight components (React, Node.js, Flask, SQLite).

Conclusion

The project *AI-Based Content Moderation System (Project Clean)* successfully demonstrates the design and implementation of an intelligent, lightweight, and transparent moderation platform. By combining **text analysis, image filtering, reputation scoring, and administrative oversight**, the system addresses several critical challenges faced by existing moderation tools:

- Lack of transparency in moderation outcomes.
- Weak support for multilingual content (especially Hindi/English mixed content).
- Inability to track and penalize repeat offenders effectively.
- Over-reliance on cloud services, making offline demonstrations difficult in academic settings.

The developed prototype integrates **React.js (frontend), Node.js + Express (backend), Python Flask (AI service), and SQLite (database)** into a cohesive workflow. It operates efficiently on student laptops, requires minimal setup, and runs completely offline after the initial installation of dependencies.

The system's **key contributions** include:

- A moderation engine with explainable decisions for both text and image content.
- A reputation model that dynamically updates user behavior scores and enforces progressive penalties.
- An administrative dashboard offering oversight, analytics, and configuration management.
- A demo social media platform that enables real-time testing and academic demonstration of the system's capabilities.

In conclusion, the project validates the feasibility of creating a **transparent, offline-ready, and academically valuable content moderation platform** that can also inspire further research and real-world applications.

References

[1] P. Dhruv, and S. Naskar, "Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review," in Proc. Mach. Learn. Inf. Process.: Proc. ICMLIP 2019, Singapore, 2020, pp. 367–381.

[2] S. S. Sohail et al., "Decoding ChatGPT: A taxonomy of existing research, current challenges, and possible future directions," J. King Saud University-Comput. Inf. Sci., vol. 35, no. 8, Sep. 2023, Art no. 101675, doi: 10.1016/j.jksuci.2023.101675.

[3] G. P. Patrinos et al., "Using ChatGPT to predict the future of personalized medicine," Pharmacogenomics J., vol. 23, no. 6, pp. 178–184, Sep. 2023, doi: 10.1038/s41397-023-00316-9.

[4] V. U. Gongane, M. V. Munot, and A. D. Anuse, "Detection and moderation of detrimental content on social media platforms: Current status and future directions," Social Netw. Anal. Mining, vol. 12, no. 1, p. 129, Sep. 2022, doi: 10.1007/s13278-022-00951-3.

[5] T. Brown et al., "Language models are few-shot learners," in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 1877–1901.

[6] T. Gillespie, "Content moderation, AI, and the question of scale," Big Data Soc., vol. 7, no. 2, Aug. 2020, Art no. 2053951720943234, doi: 10.1177/2053951720943234.

[7] K. Yousaf, and T. Nawaz, "A deep learning-based approach for inappropriate content detection and classification of youtube videos," IEEE Access, vol. 10, pp. 16283–16298, Jan. 2022, doi: 10.1109/ACCESS.2022.3147519.

[8] M. Moustafa, "Applying deep learning to classify pornographic images and videos," 2015, arXiv Preprint arXiv:1511.08899.

[9] M. Anas, A. Saiyeda, S. S. Sohail, E. Cambria, and A. Hussain, "Can generative AI models extract deeper sentiments as compared to traditional deep learning algorithms?," IEEE Intell. Syst., vol. 39, no. 2, pp. 5–10, Mar./Apr. 2024, doi: 10.1109/MIS.2024.3374582.

[10] M. Franco, O. Gaggi, and C. E. Palazzi, "Analyzing the use of large language models for content moderation with chatgpt examples," in Proc. 3rd Int. Workshop Open Challenges Online Soc. Netw., 2023, pp. 1–8, doi: 10.1145/3599696.3612895.

[11] D. Kumar, Y. AbuHashem, and Z. Durumeric, "Watch your language: Large language models and content moderation," 2023, arXiv Preprint arXiv:2309.14517.

[12] Y. Du et al., "Enhancing job recommendation through LLM-based generative adversarial networks," Proc. AAAI Conf. Artif. Intell., vol. 38, no. 8, pp. 8363–8371, Mar. 2024, doi: 10.1609/aaai.v38i8.28678.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90. 80 IEEE Intelligent Systems

[14] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv Preprint arXiv:1409.1556.

[15] M. Nadeem, S. S. Sohail, L. Javed, F. Anwer, A. Saudagar, and A. Muhammad, "Vision-enabled large language and deep learning models for image-based emotion recognition," Cogn. Computation, vol. 16, no. 5, pp. 1–14, May 2024, doi: 10.1007/s12559-024-10281-5.

[16] OpenAI, "GPT-4 technical report," 2023, arXiv:2303.08774.

[17] M. M. Soliman, M. H. Kamal, M. A. E. M. Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence recognition from videos using deep learning techniques," in Proc. Ninth Int. Conf. Intell. Comput. Inf. Syst. (ICICIS), Cairo, Egypt, 2019, pp. 80–85, doi: 10.1109/ ICICIS46948.2019.9014714.

[18] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT," IEEE Intell. Syst., vol. 38, no. 2, pp. 15–23, Mar./Apr. 2023, doi: 10.1109/MIS.2023.3254179.

[19] Q. M. Areeb et al., "Filter bubbles in recommender systems: Fact or fallacy—A systematic review," Wiley Interdisciplinary Rev.: Data Mining Knowledge Discov., vol. 13, no. 6, Art no. e1512, Aug. 2023, doi: 10.1002/ widm.1512

[20] E. Cambria, X. Zhang, R. Mao, M. Chen, and K. Kwok, "SenticNet 8: fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing," In Proc. Int. Conf. Human Comput. Interact. (HCII), 2024