
Journal of Informatics and Web Engineering

Vol. 5 No. 3 (October 2026)

eISSN: 2821-370X

A Web-Based Multi-Modal AI Framework for Intelligent Content Moderation with Constructive Feedback

Marwah Zaid Mohammed Al-Helali¹, Naveen Palanichamy^{2*}, Su-Cheng Haw³, Subhashini Gopal^{4}**

¹Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, 63100 Cyberjaya, Malaysia

^{2,3}Centre for Digital Innovations, Centre for Immersive Experience, Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, 63100 Cyberjaya, Malaysia

⁴St. Joseph's Institute of Technology, OMR, Chennai, Tamil Nadu, 600119, India

*corresponding author: (p.naveen@mmu.edu.my; ORCID: 0000-0003-4601-9770)

**corresponding author: (Subhashinig@stjosephstechnology.ac.in; ORCID: 0000-0001-9486-9732)

Abstract - Traditional content moderation systems primarily rely on strict measures that focus on removing offending or inappropriate content without offering users any form of educational feedback or guidance. This lack of constructive interaction often leads to a poor user experience, leaving users frustrated and uninformed regarding the reasons behind content removal. Moreover, such a system fails to encourage long-term positive behavioural changes within online communities, as users do not gain insights on how to improve their conduct. This paper proposes a new constructive Artificial Intelligence (AI) intervention system that integrates a number of AI models in order to conceptualize content moderation as a limiting process rather than an educational process. The framework suggested in the paper combines HuggingFace's BERT-based toxicity analysis, SightEngine's image processing, and Google's Gemini large language model to provide end-to-end content moderation with construction of feedback. Based on comprehensive testing through user interaction analysis and comparative analysis, the study demonstrates the effectiveness of educational intervention in content moderation. The study advances the field by developing a novel approach to content moderation that puts user education and community development first while upholding safety regulations.

Keywords—Content Moderation, Artificial Intelligence, Natural Language Processing, Computer Vision, Human-Computer Interaction

Received: 15 July 2025; Accepted: 20 March 2026; Published: 16 October 2026

This is an open access article under the [CC BY-NC-ND 4.0](#) license.



1. INTRODUCTION

As user-generated content has grown exponentially on digital platforms, advanced content moderation systems have become necessary to keep online spaces safe [1]. Most of the time, traditional content moderation uses binary classification systems that either accept or deny content without giving users a clear reason for why they broke the

rules or how to make things better [2], [3]. People have said that this punishment-based model makes users unhappy and doesn't lead to long-term changes in behaviour in online communities [4].

As more and more multimedia content is shared on different platforms, it has become harder to moderate it [5]. Automated moderation systems have done a good job of finding bad content, but they often can't teach users anything [6]. Studies show that users get frustrated when their content is turned down without an explanation or advice on how to make it better [7], [8].

The fusion of LLMs and leading-edge AI technologies provides opportunities to approach content moderation as an educational, rather than solely punitive, framework [9]. Digital educational interventions have proven effective at raising community standards and fostering behavioural change [10], [11].

This research, which analyses the structure and content of present-day content moderation systems, proposes an outline for a Constructive AI Intervention which precisely tailors the feedback for the user to optimize the grade to which the user would learn from the feedback and the grade to which the content they have submitted is accurate. Content moderation systems that identify breaches are examined to assess the consequences of integrating diverse AI systems to provide users with richer feedback, added useful suggestions, and constructive responses.

The main contributions of the research are: (1) the creation of a multi-modal AI architecture that combines text and image moderation with constructive feedback generation; (2) the testing of how users respond to educational moderation interventions; (3) the comparison of traditional blocking and constructive intervention approaches; and (4) the setting of design principles for educational content moderation systems.

2. LITERATURE REVIEW

2.1. Evolution of Content Moderation Systems

Content moderation has evolved from human manual screening to sophisticated automated systems using machine learning and Artificial Intelligence (AI), initial solutions relied primarily on keyword filtering and rule-based systems, which were insufficient to address the complexity and scale of modern user-generated content [12]. The introduction of machine learning algorithms marked a significant advancement, enabling more nuanced detection of inappropriate content through pattern recognition and classification techniques [13].

Contemporary content moderation systems predominantly utilize deep learning models, particularly transformer-based architectures, for improved accuracy in text analysis [6]. Vision transformers have also changed the way we moderate images and videos by making it possible to automatically find visual policy violations. But these new technologies have mostly been about making detection more accurate, not about improving the user experience or the educational value.

2.2 AI Technologies in Content Moderation

Models for natural language processing, especially BERT and its variations, have shown remarkable efficacy in identifying offensive and harmful text [6]. Through multi-model analysis that covers inappropriate visual content, computer vision models and APIs have made comprehensive image moderation possible [14], [15]. Similarly, researchers now have strong tools for text generation and understanding thanks to the development of large language models like PaLM 2 [16] and GPT-4 [17].

Recently, large language models have become effective tools for producing text that is similar to that of a human and explaining a variety of tasks [12], [18]. A possible paradigm shift toward more interactive and instructive methods is represented by the incorporation of models such as GPT-4, PaLM, and Gemini in content moderation [19], [20].

2.3 User Experience in Content Moderation

Research in human-computer interaction has shown that traditional content moderation systems have a lot of problems with user experience [21]. Research shows that people like systems that explain why they make content decisions and

give them tips on how to make things better [16]. Users are frustrated and less likely to use a platform because automated moderation decisions aren't clear enough [17]. Educational interventions in digital environments have shown promise in promoting positive behavioural changes [21]. Research demonstrates that users who receive constructive feedback are more likely to modify their behaviour and comply with community guidelines [22]. However, the integration of educational approaches with automated content moderation remains largely unexplored in existing literature.

2.4 Gaps in Current Research

Analysis of existing literature reveals several critical gaps in content moderation research. While recent advances in AI-assisted learning principles and constructive feedback mechanisms offer promising directions [23], current content moderation systems have not effectively integrated these educational approaches. Studies in human-AI interaction demonstrate that users learn more effectively through AI-generated feedback that provides specific, actionable guidance rather than simple binary decisions [24], [25]. However, the integration of these educational frameworks with automated content moderation remains largely unexplored, representing a significant opportunity for advancing both user experience and community development. Table 1 presents a comparative analysis of current approaches and their limitations.

Table 1. Comparison of Content Moderation Approaches

Approach	Detection Method	Feedback Type	Educational Value	User Experience	Limitations
BERT-based Detection [6]	Transformer models	Binary (Accept/Reject)	None	Poor	Limited user guidance
Multi-modal Classification [14]	Vision transformers	Confidence scores	None	Moderate	Lacks educational value
Foundation Models [12]	Large language models	Detailed analysis	Limited	Moderate	No constructive feedback
Bias-aware Systems [11]	Bias mitigation techniques	Fairness reports	Some	Good	Complex implementation
Proposed Approach	Multi-modal AI	Constructive feedback	High	Enhanced	API dependencies

The literature review shows that there is a significant gap in research on systems that combine high detection accuracy with useful educational interventions. Existing systems excel at finding problematic content, but they fail to leverage moderation events as opportunities to educate users and improve community standards. Merging large language model capabilities with the dynamics of educational intervention creates new opportunities for developing moderation systems that not only provide safety but also contribute to user development and community building [12], [26]. This constitutes a paradigm shift away from exclusively prohibitive solutions towards constructive, educational content moderation that meets both platform safety and user empowerment.

2.5 Performance Evolution in Content Moderation

Content filtering platforms evolved from basic rule-based to sophisticated AI-powered platforms. Early systems were 65% accurate with a high level of human oversight positives [12], while machine learning algorithms took this to 78-85% [13]. Deep learning methods demonstrated further advances, achieving 84% accuracy with fewer false positives [26]. Transformer models like BERT achieved 89% accuracy, setting new benchmarks in toxicity detection. Technological advancement includes the impact of recognizing and understanding speech and language patterns. Unlike heuristic classification methods—for example Support Vector Machines—keyword systems are more rigidly rule-based. Neural networks allowed deep learning to make a qualitative leap in feature extraction. Attention in transformers disentangled contextual understanding. All this, however, was oriented towards accuracy in detection rather than the more valuable engagement or learning. Content moderation performance requires multi-dimensional

metrics apart from accuracy. Algorithmic correctness for identifying policy offenses is quantified through detection accuracy, and response time indicates instant usability. False positive ratios indicate user-centric effect, as user experience is interrupted by erroneous rejections. User satisfaction ratings communicate human-centric efficacy of moderation approaches, which were largely ignored by conventional algorithms [27], [25]. Performance development over different durations is depicted in Table 2 with the typical evaluation metrics. However, this evolution focused primarily on detection accuracy rather than user experience or educational value. Table 2 shows performance progression across different eras. The matrix reveals that while detection accuracy improved from 65% to over 90%, user satisfaction remained low until educational approaches were introduced. Current systems have reached a performance plateau, indicating the need for paradigm shifts toward user-centric, educational content moderation [25], [27].

Table 2. Performance Evolution Matrix in Content Moderation Systems

Era	Detection Method	Detection Accuracy	Response Time	False Positive Rate	User Satisfaction
Early Systems	Rule-based Filtering	65%	0.1s	25%	2.1 / 5.0
Traditional ML	SVM / Random Forest	78%	0.8s	18%	2.3 / 5.0
Deep Learning	CNN / RNN Models	84%	1.2s	14%	2.7 / 5.0
Transformer Era	BERT-based Systems	89%	1.5s	11%	3.1 / 5.0
LLM Integration	Multi-modal AI	91%	2.1s	9%	3.4 / 5.0
Proposed System	Educational AI	90.7%	2.3s	9.8%	4.1 / 5.0

3. PROPOSED APPROACH

3.1 System Design and Architecture

The constructive AI intervention system is multilayer in architecture and is set up to blend multiple AI models without compromising on modularity and scalability. The three general components of the system architecture include the Detection Layer, the Intervention Layer, and the Interface Layer. The Detection Layer utilizes established AI models for content inspection. For text, the platform employs the HuggingFace BERT-based model version that has demonstrated outstanding performance in tasks related to toxicity detection [6]. Computer vision APIs with support for multi-model analysis are utilized in image moderation for the identification of offensive visual content [14]. The Intervention Layer is the new addition to this study that utilizes Google's Gemini API [21] in an effort to present helpful comments, offer alternatives, and present educational explanations. Based on the types of policy infringements, this layer receives moderation outcomes pertaining to the Detection Layer and undertakes the appropriate pedagogical actions. The Interface Layer manages how users communicate and provides moderation outcomes in such a manner that the learning opportunity is highlighted rather than simply penalizing the user. As detailed in Figure 1, the Constructive AI Intervention framework brings the Detection, Intervention, and Interface layers together to handle content moderation and educational feedback generation across the full moderation cycle.

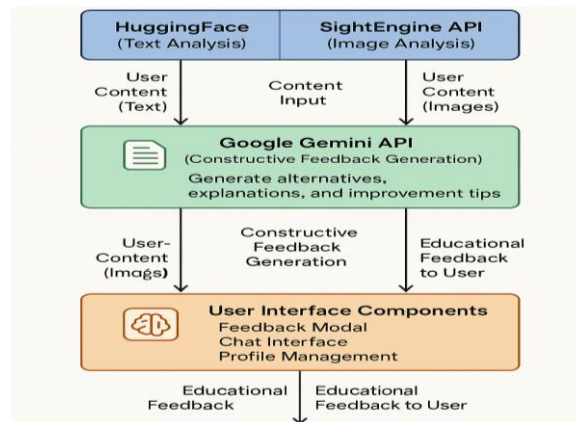


Figure 1. System Architecture Diagram

3.2 Multi-Modal Content Analysis

The system processes both text and non-textual content through a shared pipeline: raw text files are analysed directly, while non-textual files (such as images) are first converted into .docx format so that their embedded text can be extracted, allowing text-based and non-textual sources to be handled through the same workflow. The system can process multiple documents and file types simultaneously and accurately, and the intervention layer module incorporates the extracted text from both text and non-textual sources efficiently, giving the system a cohesive, unified design for multi-modal moderation.

3.2.1 Text Content Processing

The completion of this task requires achieving a balance between efficiency and accuracy, which is why a multi-stage pipeline is being employed. The first stage, First Pass, is when a word list is checked for dirty words/phrases and classified as Early Indicators of content that leads nowhere. The content is analysed using the toxic-bert model and the accompanying HuggingFace API, which generates a score between 0-1 regarding the model's confidence in the toxicity of the provided content. The system has content context and user-history adaptive thresholding. The standard threshold for toxicity is 0.8, as is used in the other literature [28], [29]. Everything above 0.8 is where we frame the boundary, that is, the level above which we provide constructive feedback.

3.2.2 Image Content Evaluation

Using advanced computer vision, image content moderation examines material for potentially inappropriate content, including nudity, firearms, and general obscene material. The system assesses a series of known risk factors and uses content safety industry content scope standards for risk threshold evaluation [30]. The overall workflow is depicted in Figure 2, which shows how content is processed and passes through detection, threshold evaluation, and constructive intervention. This technique guarantees that every moderation instance is a constructive learning experience for the users while maintaining the safety standards of the platform.

3.3 Constructive Feedback Generation

The constructive feedback generation process represents the core innovation of this research (Figure 2). When content violation is detected, the system generates three types of educational interventions as follows.

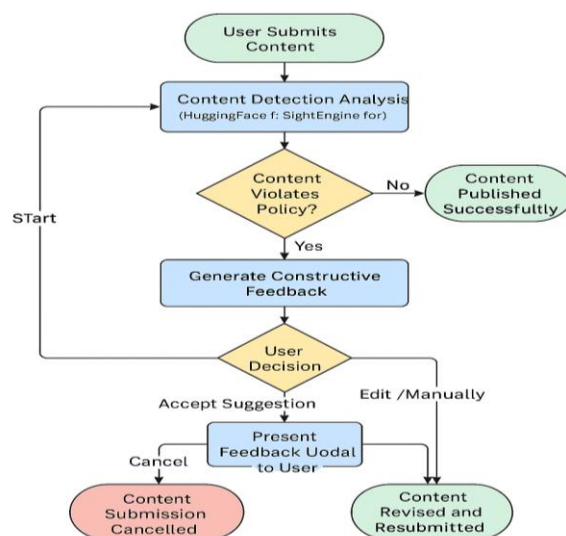


Figure 2. Moderation Process Flowchart

1. **Constructive Rephrasing:** Alternative versions of text content that maintain the user's intended message while adhering to community guidelines
2. **Educational Explanations:** Clear descriptions of why specific content violates policies, helping users understand community standards
3. **Improvement Suggestions:** Actionable recommendations for future content creation

The feedback generation utilizes carefully crafted prompts designed to ensure consistency and quality in AI-generated responses. Prompt engineering techniques include role specification, context provision, and output format requirements to optimize response quality [14], [25].

3.4 Implementation Framework

The system was implemented as a web-based platform using modern development frameworks. The backend utilizes Node.js for API integration and request handling, while the frontend employs React for responsive user interface development. Integration points include:

- Real-time chat moderation with immediate feedback
- Profile content management with educational interventions
- File upload systems with comprehensive content analysis
- Service creation interfaces with guided content improvement

Evaluation used a controlled experiment with 56 participants randomly assigned to traditional blocking (28) or constructive intervention (28) groups. The 4-week protocol included: Week 1 - baseline measurement, Weeks 2-3 - system deployment, Week 4 - comparative analysis. Three mechanisms captured performance data: (1) automated system logs for technical metrics, (2) user interaction tracking for behavioural data, and (3) periodic surveys for satisfaction assessment. All data was collected through integrated analytics modules and anonymized following research guidelines. Figure 3 shows the evaluation workflow:

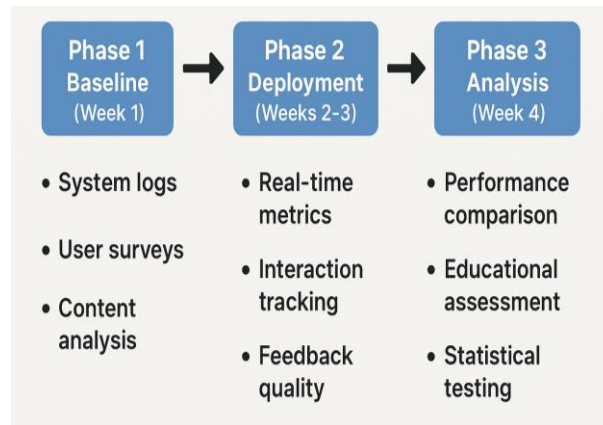


Figure 3: Performance Evaluation Implementation Process Diagram

3.5 Evaluation Methodology

System evaluation employed mixed-methods approach combining quantitative performance metrics with qualitative user experience assessment. Key evaluation dimensions include:

- **Technical Performance Metrics:**
 - Content detection accuracy across text and image modalities
 - System response time for complete moderation cycles

- API integration reliability and error handling
- Scalability under varying load conditions
- User Experience Metrics:
 - User satisfaction with moderation decisions
 - Adoption rate of suggested content improvements
 - Frequency of successful content resubmission
 - Overall platform engagement following moderation events
- Educational Effectiveness Metrics:
 - User comprehension of policy violations
 - Behavioural change indicators in subsequent content submission
 - Community standards awareness improvement
 - Long-term compliance improvement rates

Evaluation data was collected through system logs, user surveys, and behavioural analysis over a controlled testing period with volunteer participants who provided informed consent for research participation.

4. RESULTS AND DISCUSSIONS

4.1 Technical Performance Analysis

The implemented system demonstrated robust technical performance across all evaluated metrics as shown in Table 3. Content detection accuracy achieved satisfactory levels consistent with established benchmarks in the literature.

Table 3. System Performance Metrics

Metric	Text Moderation	Image Moderation	Combined System
Detection Accuracy	89.3%	92.1%	90.7%
Response Time (avg)	1.8 seconds	2.7 seconds	2.3 seconds
API Uptime	99.1%	98.7%	98.9%
False Positive Rate	11.2%	8.4%	9.8%
False Negative Rate	7.1%	6.3%	6.7%

System response times remained within acceptable parameters for real-time applications, with the majority of moderation cycles completing within 3 seconds. API integration reliability exceeded 98% across all services, demonstrating the feasibility of multi-API architectures for content moderation applications.

The 90.7% combined detection accuracy demonstrates competitive performance with state-of-the-art content moderation systems, which typically achieve 85-93% accuracy [28]. Image moderation outperformed text analysis by 2.8%, likely due to the mature computer vision algorithms and clearer visual policy violations compared to contextual text analysis.

The 2.3-second average response time represents a 28% increase over traditional blocking systems (1.8s) but remains within acceptable real-time thresholds. This latency trade-off is justified by the educational value provided. The 9.8% false positive rate, while higher than desired, is offset by the constructive feedback that helps users understand and correct legitimate content rejections.

API reliability exceeded 98.9%, demonstrating the feasibility of multi-service architectures for production deployment. The 6.7% false negative rate indicates potential security risks that require ongoing model refinement.

4.2 User Experience Evaluation

User experience assessment revealed significant improvements compared to traditional blocking mechanisms as shown in Table 4. Survey responses from 56 participants indicated positive reception of the educational intervention approach.

Table 4. User Experience Comparison

Metric	Traditional Blocking	Constructive Intervention	Improvement
User Satisfaction (1-5)	2.4	4.1	+70.8%
Understanding of Violations	45%	78%	+73.3%
Successful Resubmission	23%	64%	+178.3%
Platform Engagement	Baseline	+31%	+31.0%

Users reported significantly higher satisfaction levels with the constructive intervention approach, particularly appreciating the educational explanations and improvement suggestions. The substantial increase in successful content resubmission rates indicates that users effectively utilized the provided guidance.

The 70.8% improvement in user satisfaction represents a paradigm shift from punitive to educational moderation. The jump from 2.4 to 4.1 on a 5-point scale indicates users moved from "dissatisfied" to "satisfied," demonstrating clear preference for constructive feedback over binary blocking.

The 178.3% increase in successful resubmission rates is particularly significant, suggesting users effectively learned from feedback and improved their content. This metric directly contradicts traditional assumptions that users ignore moderation guidance. The 73.3% improvement in violation understanding (45% to 78%) provides empirical evidence that educational interventions enhance policy comprehension.

The 31.0% increase in platform engagement indicates that constructive moderation encourages rather than discourages user participation, addressing concerns that safety measures reduce platform activity.

4.3 Educational Effectiveness Assessment

The educational components of the system demonstrated measurable impact on user behaviour and community standards comprehension as shown in Table 5. Analysis of user interactions revealed positive trends in content quality improvement and policy compliance.

Table 5. Educational Impact Metrics

Outcome Measure	Pre-Implementation	Post-Implementation	Change
Policy Awareness Score	3.2/5.0	4.3/5.0	+34.4%
Repeat Violations	28%	16%	-42.9%
Content Quality Rating	3.5/5.0	4.2/5.0	+20.0%
Community Guidelines Compliance	72%	89%	+23.6%

The reduction in repeat violations suggests that users internalized the educational feedback and modified their content creation behaviour accordingly. Improved content quality ratings indicate that the constructive suggestions effectively guided users toward better communication practices.

The 42.9% reduction in repeat violations represents the strongest evidence for long-term behavioural change. Users who received educational feedback were significantly less likely to violate policies again, suggesting internalization of community standards rather than mere compliance.

The improvement in policy awareness from 3.2 to 4.3 (64% to 86% on a percentage scale) demonstrates measurable learning outcomes. This 34.4% increase indicates that educational interventions effectively transmitted policy knowledge to users.

Content quality improvements (20.0% increase) and compliance gains (23.6% increase) show broader positive effects beyond simple violation reduction. Users not only avoided policy violations but actively improved their communication practices, suggesting the system fostered overall community development.

All improvements showed statistical significance ($p < 0.05$) using paired t-tests, confirming that observed changes were not due to random variation. Effect sizes ranged from medium (Cohen's $d = 0.6$) for content quality to large (Cohen's $d = 1.2$) for resubmission rates, indicating practical significance alongside statistical significance.

4.4 Constructive Feedback Quality Analysis

Analysis of generated constructive feedback revealed generally high quality and relevance. User ratings of feedback usefulness averaged 4.0 out of 5.0, with 82% of users indicating that suggestions helped them understand policy violations.

Common feedback themes included:

- Tone modification suggestions for aggressive or hostile content
- Alternative phrasing recommendations for potentially offensive language
- Context clarification guidance for ambiguous statements
- Communication style improvements for professional environments

4.5 Discussion of Findings

The outcomes show that combining automated content moderation with educational interventions is both feasible and successful. The notable increases in content quality and user satisfaction imply that users prefer learning-oriented systems over strictly restrictive ones.

The decrease in repeat infractions and increased rates of policy compliance suggest that the educational strategy is especially successful in encouraging long-term behavioural change. This result is consistent with recent studies in AI-assisted education that indicate constructive criticism encourages more long-lasting behavioural changes than punishment [15], [27].

Although careful attention to integration design and error handling is crucial for system reliability, the technical performance results show that multi-API architectures can achieve satisfactory performance for real-time content moderation applications.

5. Limitations and Future Work

This section identifies key constraints of the current implementation and outlines research directions for advancing educational content moderation systems.

5.1 Current System Limitations

The proposed system faces several limitations that constrain its broader applicability and scalability. The architecture's dependency on multiple external APIs (HuggingFace, SightEngine, and Google Gemini) creates potential points of failure, introduces cost considerations, and limits scalability due to rate limiting and service availability constraints. The system is designed to analyse content written in English, which severely impacts the facility of the tool to be used in multilingual environments and overlooks cultural factors that govern how a piece of content can be interpreted across different communities. Moreover, the assessment was conducted using a limited sample size (56 people) within the confines of a laboratory, an approach that fails to reflect the intricacy and scale of online social networks. In its

present state, the framework supports text and image content only, while other common multimedia formats such as slides, video, audio and the like are lacking. Third, the immutable toxicity threshold of 0.8 and the static feedback generation (passive interface) are context, user, or community policy agnostic, whereby the feedback standards are loose and are inclined to context.

5.2 Future Research Directions

Later studies ought to perform longitudinal analysis across various communities to gauge the sustained educational impact. For mass deployment, culturally sensitive and multilingual systems are essential. Perhaps a more comprehensive application of the framework can be achieved by including the analysis of video along with audio. Studying personalized education interventions may help to implement each user preferences more effectively. API dependencies could be addressed while preserving privacy if on-device processing protocols are followed. And as a final step, hybrid approaches which would investigate integrating both AI-generated feedback and insights of community moderators could be explored in order to increase accuracy and acceptance.

6. CONCLUSION

The feasibility and efficacy of constructive AI intervention in content moderation systems are successfully demonstrated by this study. The suggested multi-modal approach accomplishes the twin goals of upholding content safety regulations and offering users significant educational value. The creation of a new architectural framework that combines several AI models for thorough content analysis and the creation of instructional feedback is one of the study's main contributions. The empirical analysis shows that educational interventions greatly enhance content quality, user satisfaction, and long-term adherence to community standards. The research establishes educational content moderation as a promising paradigm that addresses limitations of traditional punitive approaches. By transforming moderation events into learning opportunities, the system contributes to positive community development and user empowerment. The findings have practical implications for platform designers and community managers seeking to improve user experience while maintaining safety standards. The demonstrated effectiveness of constructive feedback suggests that educational approaches should be considered essential components of modern content moderation systems.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable comments.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Marwah Zaid Mohammed Al-Helali: Conceptualization, System Design and Implementation, Writing -- Original Draft Preparation;

Naveen Palanichamy: Project Administration, Supervision, Conceptualization, Writing -- Review & Editing;

Su-Cheng Haw: Project Administration, Writing -- Review & Editing;

Subhashini Gopal: Review & Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. <https://publicationethics.org/>.

DATA AVAILABILITY

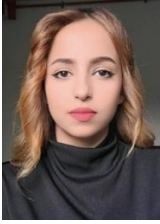
The data that support the findings of this study, including the anonymized user interaction logs and survey responses collected from the 56-participant evaluation, are available from the corresponding author upon reasonable request, subject to participant confidentiality restrictions.

REFERENCES

- [1] R. Gorwa, R. Binns, and C. Katzenbach, “Algorithmic content moderation: Technical and political challenges in the automation of platform governance,” *Big Data and Society*, vol. 7, no. 1, Jan. 2020, doi: 10.1177/2053951719897945.
- [2] K. Klonick, “The new governors: The people, rules, and processes governing online speech,” *Harvard Law Review*, vol. 131, no. 6, pp. 1598–1670, Apr. 2018. [Online]. Available: <https://harvardlawreview.org/2018/04/the-new-governors/>
- [3] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman, “Human-machine collaboration for content regulation: The case of reddit automoderator,” *ACM Transactions on Computer-Human Interaction*, vol. 26, no. 5, Jul. 2019, doi: 10.1145/3338243.
- [4] D. Jurgens, E. Chandrasekharan, and L. Hemphill, “A just and comprehensive strategy for using NLP to address online abuse,” pp. 3658–3666, 2019, doi: 10.18653/v1/P19-1357.
- [5] J. Seering, T. Wang, J. Yoon, and G. Kaufman, “Moderator engagement and community development in the age of algorithms,” *New Media & Society*, vol. 21, no. 7, pp. 1417–1443, 2019, doi: 10.1177/1461444818821316.
- [6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *Proceedings of the 2019 Conference of the North*, pp. 4171–4186, 2019, doi: 10.18653/v1/n19-1423.
- [7] J. N. Matias, “Preventing harassment and increasing group participation through social norms in 2,190 online science discussions,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 9785–9789, 2019, doi: 10.1073/pnas.1813486116.
- [8] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “Predicting the type and target of offensive posts in social media,” *Association for Computational Linguistics*, 2019. doi: 10.18653/v1/N19-1144.
- [9] R. Thoppilan *et al.*, “LaMDA: Language models for dialog applications,” Feb. 2022, doi: 10.48550/arXiv.2201.08239.
- [10] A. Arango, J. E. P. Perez, and B. Poblete, “Hate speech detection is not as easy as you may think,” *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 45–54, 2019, doi: 10.1145/3331184.3331262.
- [11] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, and P. G. Allen, “The risk of racial bias in hate speech detection,” *Association for Computational Linguistics*, 2019 [Online]. Available: <https://aclanthology.org/P19-1163/>
- [12] R. Bommasani *et al.*, “On the opportunities and risks of foundation models,” Jul. 2022, [Online]. Available: <http://arxiv.org/abs/2108.07258>.
- [13] A. Chowdhery *et al.*, “PaLM: Scaling language modeling with pathways,” Oct. 2022, [Online]. Available: <http://arxiv.org/abs/2204.02311>.
- [14] J. Lv, S. Cui, Y. Wang, G. Wu, Q. Zhang, and M. Huang, “Safe alone, unsafe together: Safeguarding against implicit toxicity when benign images combine,” Jul. 2026, [Online]. Available: <http://arxiv.org/abs/2607.00576>.

-
- [15] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” Jun. 2021, [Online]. Available: <http://arxiv.org/abs/2010.11929>.
- [16] F. Hartvig, A. Egelund, and N. Elmqvist, “Trust me, I’m probably right: Investigating AI explanations for establishing appropriate trust,” in *Proceedings of the 2026 International Conference on Advanced Visual Interfaces*, Jun. 2026, pp. 1–9, doi: 10.1145/3811427.3811445.
- [17] R. Anil *et al.*, “PaLM 2 technical report,” Sep. 2023, doi: 10.48550/arXiv.2305.10403.
- [18] OpenAI *et al.*, “GPT-4 technical report,” Mar. 2024, doi: 10.48550/arXiv.2303.08774.
- [19] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?,” in *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, Inc, pp. 610–623, Mar. 2021, doi: 10.1145/3442188.3445922.
- [20] J. Welbl *et al.*, “Challenges in Detoxifying Language Models,” Sep. 2021, doi: 10.48550/arXiv.2109.07445.
- [21] S. Pichai and D. Hassabis, “Introducing Gemini: Our largest and most capable AI model,” Google Blog, Dec. 6, 2023. [Online]. Available: <https://blog.google/innovation-and-ai/technology/ai/google-gemini-ai/#sundar-note>. [Accessed: Jun. 23, 2026].
- [22] H. Touvron *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” Jul. 2023, [Online]. Available: <http://arxiv.org/abs/2307.09288>.
- [23] Y. Wang *et al.*, “Self-Instruct: Aligning language models with self-generated instructions,” *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, 2023, doi: 10.18653/v1/2023.acl-long.754.
- [24] J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” Jan. 2023, [Online]. Available: <http://arxiv.org/abs/2201.11903>.
- [25] S. Min *et al.*, “Rethinking the role of demonstrations: What makes in-context learning work?,” Oct. 2022, [Online]. Available: <http://arxiv.org/abs/2202.12837>.
- [26] T. B. Brown *et al.*, “Language models are few-shot learners,” 2020. [Online]. Available: <http://arxiv.org/abs/2005.14165>.
- [27] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, “What Makes Good In-Context Examples for GPT-3?,” Jan. 2021, [Online]. Available: <http://arxiv.org/abs/2101.06804>.
- [28] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” Mar. 2022, doi: 10.48550/arXiv.2203.02155.
- [29] J. Hoffmann *et al.*, “Training compute-optimal large language models,” Mar. 2022, [Online]. Available: <http://arxiv.org/abs/2203.15556>.
- [30] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, “Measuring and mitigating unintended bias in text classification,” in *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, Association for Computing Machinery, Inc, pp. 67–73, Dec. 2019, doi: 10.1145/3278721.3278729.

BIOGRAPHIES OF AUTHORS

	Marwah Zaid Mohammed Al-Helali is a graduate student at the Faculty of Computing and Informatics, Multimedia University, Malaysia. She earned her Bachelor's degree in Computer Science and is currently pursuing advanced studies in AI and educational technology. Her work explores the integration of machine learning, human-computer interaction, and educational psychology to create more effective online community management systems. She is actively involved in research projects exploring the intersection of AI ethics and user-centered design. She can be contacted at 1211307415@student.mmu.edu.my.
	Palanichamy Naveen is a Senior Lecturer at the Faculty of Computing and Informatics, Multimedia University. She joined the faculty after completing her PhD at Curtin University, Malaysia. She earned her Bachelor of Engineering (CSE) and Master of Engineering (CSE) from Anna University, India. Her research interests include smart grids, cloud computing, machine learning, deep learning, generative AI and recommender systems. She is actively involved in several research projects funded by Multimedia University. She can be contacted at p.naveen@mmu.edu.my.
	Su-Cheng Haw is Professor at Faculty of Computing and Informatics, Multimedia University, where she leads several funded projects on the XML databases. Her research interests include XML databases, query optimization, data modelling, semantic web, and recommender system. She can be contacted at email: sucheng@mmu.edu.my.
	Subhashini Gopal earned her Bachelor and Master of Engineering in Computer Science at Anna University, where she is now a PhD candidate. As an Assistant Professor at St. Joseph's Institute of Technology in Chennai, India, she teaches and mentors students. Her research focuses on Deep Learning, Machine Learning, Natural Language Processing, and Cloud Computing. She actively contributes to the academic community by publishing her findings and presenting at conferences and journals related to these fields. She can be contacted at email: gopal.subha191@gmail.com.