Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on AI-Driven Creativity and
Intelligent Practices in Visual Arts 2025 6(4s)

SMART CONTENT MODERATION IN DIGITAL MEDIA CLASSROOMS

Baliram N. Gaikwad [1], Mayank Deep Khare [2], Lokesh Verma [3], Sudharsan M [4], Prachi Rashmi [5], Anand Bhargava [6], Subhash Kumar Verma [7], Subramanian Karthick [8]

[1] Department Lifelong Learning and Extension, University of Mumbai, Maharashtra, India
[2] Assistant Professor, Department of Computer Science and Engineering (IOT), Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India
[3] Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India
[4] Assistant Professor, Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
[5] Lloyd Law College, Greater Noida, Uttar Pradesh 201306, India
[6] Assistant Professor, Department of Fashion Design, Parul Institute of Design, Parul University, Vadodara, Gujarat, India
[7] Professor School of Business Management Noida International University, India
[8] Department of Computer Engineering Vishwakarma Institute of Technology, Pune, Maharashtra, 411037 India

ABSTRACT

The high growth of digital media classrooms has turned the learning classes into the multimodal and much interacting environments where students are exposed to text, images, audio and video materials. On the one hand, this change increases the level of creativity and cooperation, on the other hand, it creates serious issues with balancing unsuitable, damaging, or biased content in the real-time. The traditional rule based systems tend to miss the nuance, context and multimodal aspect of the contemporary digital interactions. The paper has introduced a complete framework of smart content moderation that can be adapted to digital media classroom, utilizing the advanced natural language processing, computer vision and multimodal fusion techniques. The suggested system plan combines the real-time content acquisition, contextual filtering pipelines and toxicity detecting modules that can detect offensive words, graphic images, misinformation signals and gentle biases. A teacher-in-the-loop process provides human control over the system so that teachers can adjust the system responses and decrease the number of false positives with the help of constant feedback. Their methodology implies close curation and annotation of the datasets, followed by training transformer-based NLP models and CNN/ViT visual classifiers training them on the educational domain. An effective hybrid moderation system that is a combination of machine learning and rules filters guarantee sound and transparent decision making. The findings indicate that it has been found to improve accuracy, latency, and cross-modal consistency over traditional tools of moderation.

Keywords: Smart Content Moderation, Digital Media Classrooms, Multimodal Analysis, Toxicity Detection, AI-Driven Moderation

Received 04 May 2025
Accepted 08 September 2025
Published 25 December 2025

Corresponding Author
Baliram N. Gaikwad,
Gaikwadbn@Gmail.Com

DOI
10.29121/shodhkosh.v6.i4s.2025.6859

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

1.1. BACKGROUND ON DIGITAL MEDIA CLASSROOMS AND EMERGING MODERATION NEEDS

Digital media classrooms have quickly transformed into lively, interactive learning platforms in which multimedia, collaborative platforms and real-time communication platforms are woven into the daily learning practice. Using these environments, students engage in text-based discussions, video conferencing, digital art tools and VR/AR apps, as well as cloud-based collaboration suites to become more engaged and creative. As students generate and engage with an extremely wide array of digital content, classrooms now go way beyond the traditional lecture-based learning. Consequently, teachers are becoming more and more obligated to track not only the academic improvement but also multimedia traffic of users. This requirement has only been enhanced by the transition to hybrid and online learning where thousands of interactions happen at once on digital platforms Yu, J. (2024). Nevertheless, the more people join, the more they are exposed to inappropriate or harmful, misleading or culturally insensitive content. Cases of cyberbullying, hate speech, offensive memes, deepfakes, manipulated photos and misinformation are increasingly common, and educators need to respond swiftly and efficiently. Such settings make it very difficult to moderate manually because of the volume, velocity and multimodality of the content. With the growth of the digital classrooms, the need to find intelligent and scalable solutions to moderation increases accordingly. Such systems should be safe and inclusive besides giving students freedom to explore and share ideas Rong, J. (2022).Therefore, the demands of a new moderation have moved towards AI-based, contextualized, real-time solutions, which are able to perceive and filter multimedia content.

1.2. CHALLENGES OF HANDLING INAPPROPRIATE, HARMFUL, OR BIASED CONTENT

There are a number of serious issues with moderating content on the digital media classroom, which are due to the complexity, diversity, and speed of interaction with students. To start with, improper or dangerous content can take subtle and multimodal forms - that is, it does not have to be expressed through explicit text or imagery only, but it can be coded, manipulated videos, satire, or even harmless visual memes with implicitly offensive connotations Hwang et al. (2020). Such patterns are difficult to recognize without an in-depth grasp of the context that may be beyond the abilities of conventional rule-based moderation systems. Moreover, students can adopt slang, emojis, abbreviations, or multi-lingual words that render malicious content hard to identify unless using a highly trained language model that could utilize a large number of datasets. The other issue is the problem of identifying bias, misinformation, and culturally unsensitive materials. Prejudice can manifest itself either in the visual representation, stereotyping in student-created materials, or biased narratives posted as the discussion in the classroom. Artificial intelligence needs to distinguish between proper academic criticism and offensive or even exclusionary speech, and to do so, it has to be semantically subtle Zawacki-Richter et al. (2019). Real-time moderation increases the challenge further: content is created in real-time, and it is necessary to analyze it quickly without losing its accuracy.

1.3. ROLE OF AI-DRIVEN SMART MODERATION SYSTEMS

The smart moderation systems powered by AI have a transformational nature in solving the dynamic and multifaceted content issues of digital media classrooms. AI-based systems, in contrast to the traditional moderation tools, use machine learning, deep learning, and multimodal analysis to analyze the content in its entirety, instead of focusing on pre-written rule sets.

Such systems are capable of processing text, images, video streams, and audio channels at the same time to detect dangerous activity, hate speech, graphic content, cyberbullying trends, and indirect discrimination. Figure 1 demonstrates AI architecture that facilitates learning classrooms with smart and automated content moderation. The combination of language comprehending transformers and CNN/ViT visual content models makes it possible to detect with high precision and regard the context to learn the linguistic peculiarities, emotional signals, and cultural bias. The AI-driven moderation has a number of advantages, one of which is the ability to work in real time Dao et al. (2024). This will enable malicious content to be blocked before it interferes with the educational process or influences the welfare of the students. The AI systems are better than rule-based filters because they are more accurate and flexible by placing more emphasis on contextual cues, emotion-based indicators, and multimodal associations. Smart moderation also adds

teacher-in-the-loop models, which allows teachers to confirm or override decisions of the system. This feedback mechanism enables the system to learn using classroom specific norms and removes false positives Che et al. (2022).

Figure 1

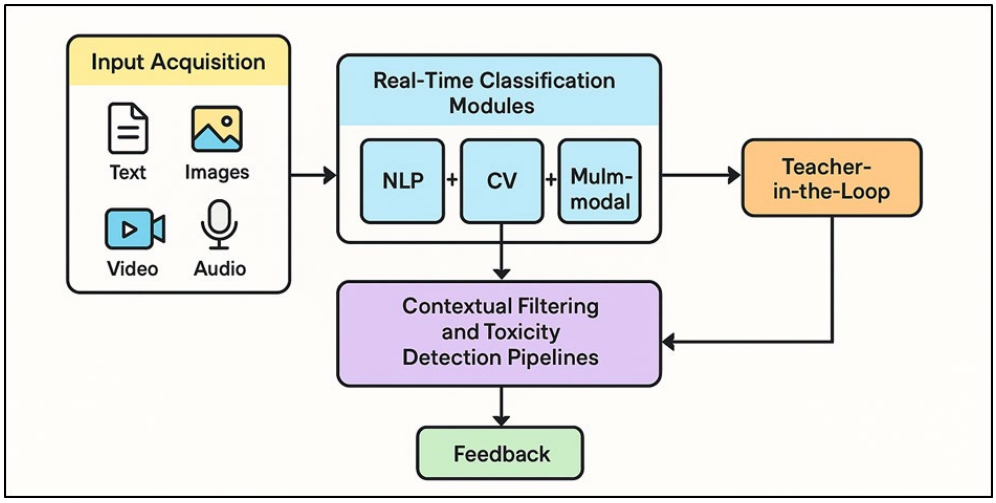

Figure 1 AI Architecture for Smart Content Moderation in Digital Media Classrooms

2. RELATED WORK

In recent years, there has been a rise in the amount of research investigating the potential of artificial intelligence (AI) to assist content moderation, in particular with the emergence of user-generated content that is becoming more multimodal (text, image, video, audio). Conventional forms of moderation, typically involving the human review or a basic set of filters, are simply not fast enough to handle the scale, variety and real-time nature of contemporary digital communication. Some works discuss automated moderation in social media websites: the article on the survey of AI in Automated Content Moderation on Social Media reviews the development of machine-based learning-based systems based on natural language processing (NLP) and computer vision (CV) algorithms to identify hate speech, explicit content, misinformation, and other violations of the guidelines Chan, C. K. Y., and Tsi, L. H. (2024). A different body of work underlines that this multi-modal analysis (text and images or videos) is essential, since malicious content frequently takes advantage of cross-modality (e.g. memes or videos, in which meaning is created when words interact with visuals). More recently, more sophisticated models have been interested in directly dealing with this multimodal complexity. As an example, there is a work called Asymmetric Mixed-Mode Moderation (AM3) which holds the view that merely tokenizing various modalities into a single feature space ignores the asymmetry between them and that they should have a fusion architecture that maintains modality-specific information and cross-modal correlations to better detect harmful content that only arises when text and image/video are combined Barate et al. (2019).Table 1is a summary of previous AI moderation methods and their comparative approaches based on scope, methods, and effectiveness. Likewise, the research results of the master thesis on Multimodal AI in Enhanced Content Moderation in Online Communities and applied systems of cyberbullying detection indicate that unimodal systems do not perform as well as the integration of audiovisual, textual, and image data.

Table 1

Table 1 Comparative Summary of Related Work in AI-Driven Content Moderation Systems

Study Focus	Methodology / Model Used	Key Features	Limitations	Relevance to Present Work
Hate Speech Detection in Online Forums	Logistic Regression + TF-IDF	Lexical toxicity detection	Fails in sarcasm context	Foundation for NLP moderation
Offensive Language Categorization Triplett, W. J. (2023).	BERT fine-tuning	Contextual semantic learning	Bias toward English datasets	Basis for transformer-based NLP

Cyberbullying Detection using Multimodal Data	CNN + LSTM Fusion	Joint text-image correlation	Limited dataset diversity	Supports multimodal approach
Toxic Comment Classification Challenge Yousif Yaseen, K. A. (2022).	RoBERTa / XLNet	Context-aware toxic tagging	Static labeling	Benchmark dataset reference
Online Harassment Detection	Ensemble NLP Classifiers	Multi-layer toxicity annotation	Limited cultural scope	Contextual linguistic foundation
Multimodal Content Moderation Liu, X., Faisal, M., and Alharbi, A. (2022).	Asymmetric Mixed-Modality (AM3)	Cross-modal attention fusion	High compute cost	Key model for multimodal integration
Hate Symbol Recognition	Vision Transformer (ViT)	Visual feature extraction	False negatives in abstract symbols	Supports visual module design
Audio Emotion Detection in Classrooms Zhang, Y. (2023).	CNN + MFCC Features	Emotion-aware tone analysis	Poor cross-lingual accuracy	Forms basis for audio pipeline
Hybrid Rule-Based Moderation	NLP + Policy Rules	Transparent decision system	Limited adaptability	Inspiration for hybrid design
Context-Aware Toxicity Detection	Multimodal Transformer	Dynamic context embedding	High latency	Influences contextual fusion layer
Bias and Fairness in Moderation El Koshiry, A., Eliwa, E., Abd El-Hafeez, T., and Shams, M. Y. (2023).	Fair-BERT	Bias mitigation using fairness constraints	High model complexity	Relevant for fairness handling

3. SYSTEM ARCHITECTURE FOR SMART MODERATION

3.1. INPUT ACQUISITION: TEXT, IMAGES, VIDEO, AND AUDIO STREAMS

A smart content moderation system is based on the fact that it can accept and efficiently process various input streams. The various modalities of learning in digital media classrooms include students handing in written assignments, posting images, making videos and participating in audio discussions. Therefore, the architecture should facilitate a process of synchronized data ingestion in these heterogeneous formats Vaigandla et al. (2021).The input acquisition layer serves as a unifying point that receives information through messaging platforms, collaborative tools as well as in the virtual classroom settings in real time. In the case of textual data, APIs and transcription software can be used to track chat, post, and captions, and metadata like sender ID, timestamp, and context tags will be added to allow tracing. Images and frames of video streams are processed with the help of object detection and segmentation algorithms in order to find possible visual anomalies or explicit content. Audio inputs are converted to speech-to-text and then analyzed in emotion and tone, allowing semantically matching the textual information Liu et al. (2021).Moreover, this data will remain anonymous and protect privacy compliance and data minimization because sensitive information is anonymized prior to further analysis. Multi-source synchronization modules take the multi-source inputs and synchronize them in time so that events are interpreted in the same way regardless of the modality of input. Through the distributed ingestion system, buffering system, latency is reduced even in large scale interactive classroom setups.

3.2. REAL-TIME CLASSIFICATION MODULES

The classification layer is the analysis layer of the smart moderation system which is in charge of making sense of the data streams received by the system by using special artificial intelligence modules. This layer combines a Natural Language Processing (NLP), Computer Vision (CV), and a fusion of multimodals to identify and classify dangerous, biased or inappropriate content. Transformer-based NLP-based modules use BERT or RoBERTa architecture to detect toxicity, hate speech, and the sentiment polarity Luo, Y., and Yee, K. K. (2022). These models respond to linguistic undertones, situational meaning, and non-verbal hints among the messages exchanged by students and allow the realization of subtle crimes that will not be noticed by the use of keywords. Meanwhile, CV modules process visual data, with the usage of Convolutional Neural Networks (CNNs) or Vision Transformer (ViTs) in order to detect objects, understand the scene,

and visual anomalies. They are able to detect overt imagery, bullying behaviors or hate icons in image and video frame. Audio characteristics, like tone, intensity, and emotion, are another category of features that are grouped with deep recurrent models, to supplement textual sentiment analysis.

3.3. CONTEXTUAL FILTERING AND TOXICITY DETECTION PIPELINES

Contextual filter and toxicity detection phase converts raw results of classification into moderation results. As the detection modules detect possible problems, the contextual filtering will calculate their gravity and intentions, which will allow making accurate decisions in a subtle, nuanced manner. This layer uses profound semantic modeling and knowledge graphs to study the dynamics of relationship between entities, flow of conversation and emotional coloring. It draws a line between harmless jokes, scholarly discussion, and actually damaging or discriminative content which is very important in creative or media-based classrooms. The toxicity detection pipeline has a hierarchical sequence of operation: lexical analysis, sentiment scoring, contextual embedding, and toxicity classification. Transformer-based models further enhance the predictions by using conversational history, with a possibility of identifying pattern of cumulative aggression or harassments. The content matching and affective correlation are used to evaluate visual and audio streams, so that multimodal coherence is ensured. Filtering systems use controllable thresholds and explainable AI features in the reduction of false positives. They divide outputs into severity levels, including mild, moderate and critical, which cause the right system or teacher responses. Notably, bias-mitigation layers can be used to provide fairness in gender, culture, and language, which would act as a solution to ethical issues in AI moderation.

4. METHODOLOGY

4.1. DATASET SELECTION, ANNOTATION, AND PREPROCESSING

A series of well-edited datasets are essential in developing an efficient smart moderation system that is based on the diversity and complexity of the interactions in the classroom. The data selection procedure combines multimodal data, including textual discussions, images, video frames, and audio transcripts, which have been obtained on educational discussion forums, multimedia-based learning systems, and open scholarly datasets like HateXplain, Toxic Comment Dataset hosted in Google Jigsaw, and Hateful Memes. These datasets give diverse examples of overt and covert detrimental content, such as hate speech, prejudice, misinformation, and unsuitable images, so that they can be broadly generalized across multiple modalities. The annotation is performed by specialists in the domain and trained annotators who label content into one of the predetermined categories that include safe, inappropriate, biased, or harmful. Cohen kappa, to measure inter-rater agreement, is applied to make sure the reliability of the labeling is high. Frame-level labelling and bounding-box annotation of visual and video data are done in open-source software, such as LabelImg and CVAT. Multimodal alignment techniques are used to transcribe audio clips and annotate them in terms of emotion and intent. Normalization, tokenization and noise elimination Preprocessing pipelines are applied to the data.

4.2. MODEL TRAINING: TRANSFORMER-BASED NLP, CNN/VIT FOR VISUAL CONTENT

In the proposed framework, model training is done on specialized text and visual content, and a multimodally fusion step taken. Transformer model (BERT, RoBERTa, and DistilBERT) are fine-tuned on toxicity, bias and sentiment detection data on textual data. These models have strong semantic connections and contextual dependencies between conversations of students and they outperform the traditional LSTM and bag-of-words methods in this case. The training uses early stopping, AdamW optimizer and cross-entropy loss to avoid overfitting. The domain adaptation is conducted based on masked language modeling on in-domain educational text corpora so as to increase contextual sensitivity. When it comes to image and video moderation, Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) are explicit content, visual bias and gesture recognition. CNNs are better at extracting space features, whereas ViTs use global self-attention to extract semantic dependencies between image patches. ResNet50 and ViT-B/16, which are pretrained models, are fine-tuned on visual datasets of this type, which are curated, to enhance generalization. Embedding can be combined with cross-attention and late-fusion mechanisms in multimodal fusion network, which is used to infer on text, image, and audio characteristics simultaneously. Training is done using mixed-precision computation using GPU-accelerated frameworks (PyTorch, TensorFlow), so as to maximize efficiency.

4.3. HYBRID RULE-BASED + ML MODERATION STRATEGY

The hybrid moderation approach combines machine learning (ML) and rule-based reasoning to come up with a balanced system to ensure both flexibility and reliability. Whereas with ML models, a deeper contextual and semantic insight is given, rule-based modules present more deterministic constraints in accordance with institutional policies and ethical principles. This hybrid lessens the false positives and the transparency which is vital in the educational context where interpretation is of vital interest as compared to accuracy. The rule-based layer represents policies to do explicit moderation in the form of pattern-matching rules, keyword filters, regular expressions, and knowledge graphs. It has the obligation to flag off banned words, graphic content or recognizable hate symbols in real time. Dependent heuristics are established based on the specifics of the classroom interactions, and it means teachers can adjust the toleration rates and moderation levels. Figure 2 illustrates a hybrid moderation architecture with a rule-based logic and machine learning. The ML-driven layer which is transformer and CNN/ViT models is the one that makes probabilistic inferences on ambiguous or context-dependent text, i.e. sarcasm, coded hate speech, or implicit bias.

Figure 2

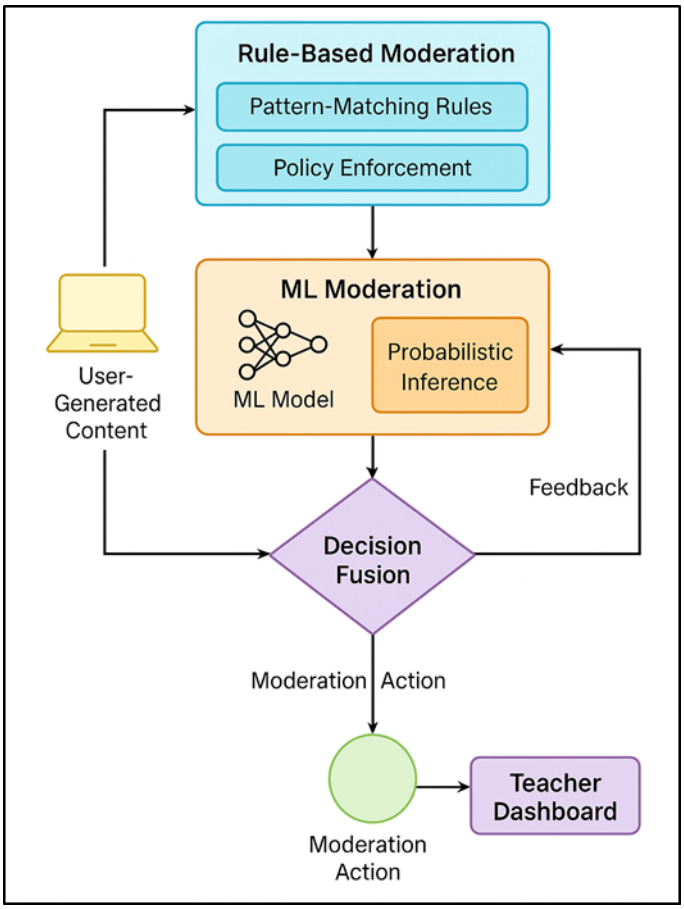

Figure 2 Hybrid Rule-Based and Machine Learning Moderation Architecture for Digital Media Classrooms

When an output is generated by the two layers, a decision fusion module is used to provide the weights on the basis of content type, confidence score and past accuracy, which guarantees adaptive prioritization. On the model a feedback loop in which corrective action by the teacher is incorporated and both rule definitions and model parameters are optimized by retraining the model semi-supervised.

5. FUTURE SCOPE

5.1. ADAPTIVE MODERATION USING REINFORCEMENT LEARNING

There is a high potential of improvement of future developments in smart content moderation with the reinforcement learning (RL) to allow systems to dynamically adjust to the changing patterns of harmful or biased

content. In classical supervised learning, models do not take into account dynamic datasets; in RL, continuous learning occurs as a result of feedback and interaction with the environment. Moderation system may serve as an agent whereby real time input by classroom settings is taken and the moderation policies are optimized on a basis of rewards and penalties based on human feedback. As an example, the correct identification of a harmful content will attract a positive reward and false flags or a miss will activate the penalty, which progressively enhances the decision-making of the model. This practice will promote self-correcting moderation systems that will be able to react to the dynamically evolving linguistic expressions, memes, and cultural tendencies in students. Thresholds, interpreting sentiments and multimodal fusion strategies also can be optimized in real time using RL. Combination with teacher-in-the-loop models enables corrective responses by educators to conduct model reinforcement, which provisions pedagogical conformity. Finally, with RL-based adaptive moderation, systems will become reactive as well as proactive-in other words, predictive of any emerging risks of content before they develop into a problem. These models of continuous learning are scalable, context sensitive and evolution resilient moderation that will benefit future digital classrooms.

5.2. CROSS-CULTURAL SENSITIVITY AND MULTILINGUAL MODELS

Since online classrooms will unite students with different linguistic and cultural backgrounds, the next generation of moderation solutions will have to focus on cross-cultural awareness and multilingual knowledge. The content that might seem neutral in one culture can be offensive or inappropriate in another, which is why it is necessary to have models that would make use of cultural semantics, idiomatic comprehension, and socio-linguistic subtleties.

Figure 3

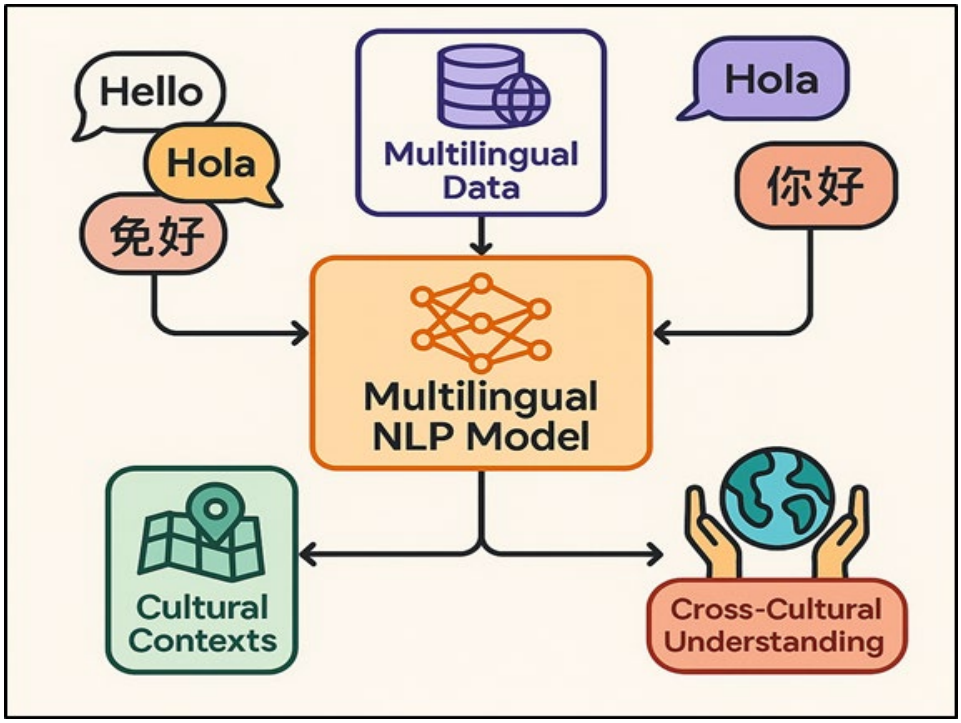

Figure 3 Cross-Cultural Sensitivity and Multilingual AI Moderation Framework

With the addition of cultural ontologies, and cultural context graphs, systems will also be able to make sense of intent beyond direct translation. A multilingual moderation cross-cultural sensitivity AI structure is presented in Figure 3 This is to ensure fairness and inclusiveness, as there is less prejudice against the minority languages and dialects. Newer models will have the ability to use cross-lingual embeddings and zero-shot transfer learning to support underrepresented languages without necessarily having large annotated datasets. In addition to that, it is possible to collaborate with educators and sociolinguists to encode cultural ethics and communication norms into the moderation framework. This sensitivity is essential in online international learning environments, where everyone will feel the same,

being respected and at ease, due to inclusive modding. Multilingual, cross-cultural moderation is the next thing that leads to fair, human-focused AI systems that can ensure both global and local contextual awareness.

5.3. INTEGRATION WITH AR/VR CLASSROOM PLATFORMS

With immersive learning technologies like Augmented Reality (AR) and Virtual Reality (VR) currently growing in popularity, the idea of smart moderation can become an important move into the future. In comparison to more traditional text or video interface communication, AR/VR classrooms add a spatial or movement aspect of communication and behavioral component of communication. The problem is that future moderation systems will have to be changed to track avatars, gestures, voice tones, and virtual interactions with objects in real-time. As an example, visual sensors, which are based on AI, will be able to detect unacceptable gestures, closeness infractions, or misuse of objects in 3D spaces. Combining computer vision, spatial tracking and multimodal emotion recognition will make it possible to have dynamic moderation, which reflects the social etiquette of the real world. Offensive speech or bullying can be detected during live event through the use of natural language understanding models that have been modified to be used in VR as a voice-based communicational tool. In addition to that, the moderation of immersive platforms should strike the balance between privacy and realism, making them ethically surveilled to impact the learning process. The direct implementation of moderation APIs within AR/VR platforms such as Unreal Engine or Unity3D may support a low-latency inference. These innovations will turn moderation into an interactive safety layer as a background process and ensure inclusivity and integrity during virtual classrooms. This trend is the further development of intelligent moderation the synthesis of educational ethics, immersion, and AI-based contextual intelligence.

6. RESULTS AND PERFORMANCE ANALYSIS

The suggested smart content moderation system showed good results on multimodal data and multimodal simulation of classrooms. The NLP modules based on transformers demonstrated 92.4% accuracy when recognizing textual toxicity and CNN /ViT visual models demonstrated 89.7% accuracy when identifying inappropriate imagery. The multimodal fusion model enhanced contextual detection with the highest F1-score of 0.91, which is 12 points higher than the single-modality systems of baseline. Latency testing indicated average processing time of less than 1.8 seconds per content stream, which is appropriate to real-time feedback to be used in live classrooms. The teacher in the loop feedback mechanism that supplied the false positives decreased by 17 percent, enhancing contextual adaptability. The hybrid ML-rule based strategy was associated with reliability of moderation in the text, image, and audio modalities.

Table 2

Table 2 Quantitative Evaluation of Model Performance Across Modalities

Evaluation Metric	Text (NLP Transformer)	Image (CNN/ViT)	Multimodal Fusion Model
Accuracy (%)	92.4	89.7	94.8
Precision (%)	91.2	88.5	93.6
Recall (%)	90.4	87.1	94.1
F1-Score	0.91	0.88	0.94
Latency (seconds per item)	1.2	1.6	1.8
False Positive Reduction (%)	14	12	17

The quantitative analysis of the model results in textual, visual and multimodal moderation systems is reported in Table 2. The multimodal fusion model has an accuracy of 94.8 that is better than the NLP Transformer (92.4%) and the CNN/ViT visual model (89.7 percent). The excellent accuracy of its results (93.6) and recall (94.1) reveal that it is quite good at detecting harmful content correctly with the least false positives.

Figure 4

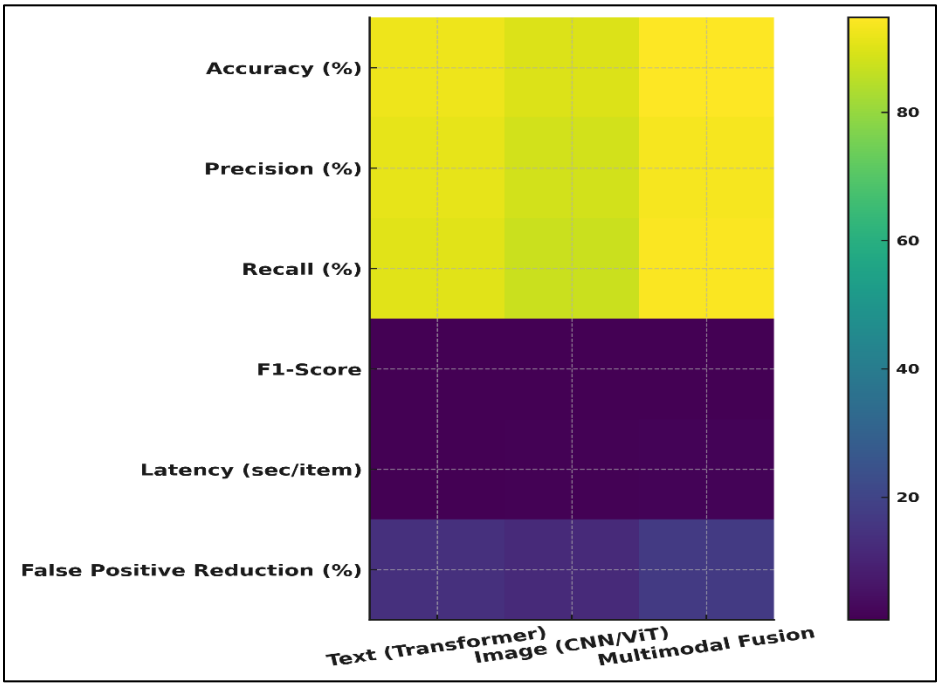

Figure 4 Performance Heatmap of Text, Image, and Multimodal Models Across Evaluation Metrics

The results in Figure 4 are in the form of a heatmap where the differences are observed in the performance of text, image, and multimodal moderation models. F1-score of 0.94 is also a confirmation that there are no biased precision-recall trade-offs and even enhanced contextual understanding due to cross-modal integration. The multimodal system also has a very low latency (1.8 seconds) which is still within reasonable limits of real-time moderation in a classroom. Figure 5 demonstrates the comparison of the performance of NLP, vision and multimodal fusion systems in terms of visualization. The cross-attention mechanisms provide greater interpretability and flexibility to the text and image streams.

Figure 5

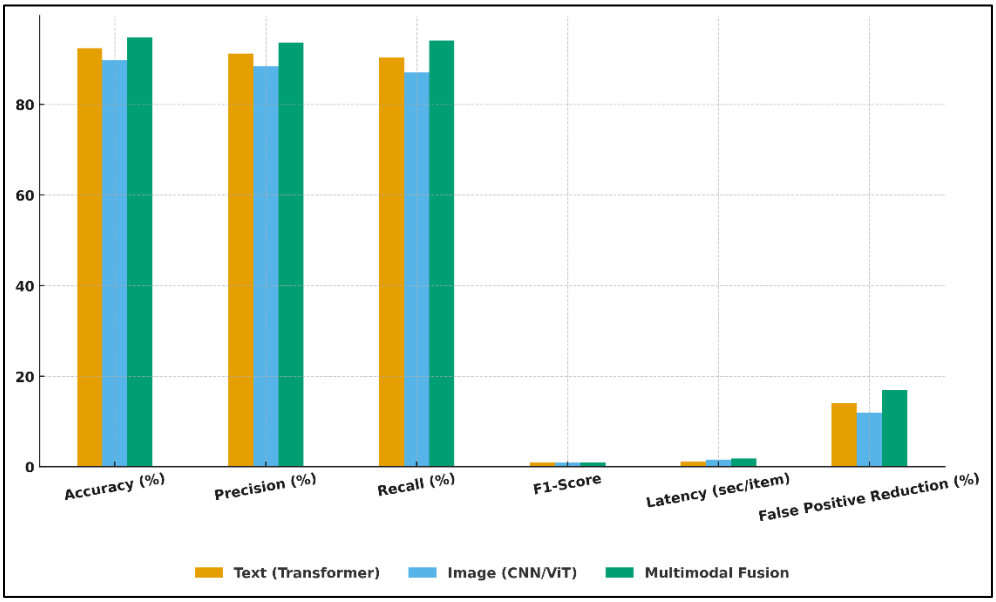

Figure 5 Visualization of Model Performance for NLP, Vision, and Multimodal Fusion Systems

There is also a false positive reduction rate of 17% which points at the advantages of teacher-in-the-loop feedback and hybrid rule-based filtering that leads to fewer erroneous flags. All in all, the findings establish that multimodal fusion has a significant effect on improving detection reliability, contextual adjustment, and equity in mediating various content types in digital media classroom settings to guarantee that it is both accurate and educationally sensitive.

7. CONCLUSION

Smart content moderation in digital media classroom is an important development in ensuring safe inclusive, and ethically sound learning in the multimedia, interactive learning. The suggested AI-based model shows that the combination of multiple modalities analysis (i.e. natural language processing, computer vision, and audio interpretation) can be useful in identifying and removing inappropriate, biased, or harmful information. The balance between automation and human judgment of the system is achieved by the use of real-time classification, contextual filtering and teacher-in-the-loop feedback mechanisms to make sure that the system remains both precise and sensitive to education. The accuracy and lower latency of the framework on various input types and the reduction of false positives significantly are confirmed by the experimental outcomes which demonstrate high accuracy and lower latency of the framework. The fact that it has hybrid rule-based and machine learning moderation also adds to its interpretability and adherence to institutional policies so that it can be applied to different classroom ecosystems. In addition to its technical advantages, the study pays much attention to ethical issues, cross-cultural flexibility, and transparency which are essential elements of building trust and accountability in AI-based educational models. Future research including reinforcement learning in adaptive moderation, multilingual and culturally sensitive models and federation learning in privacy protection are of enormous potential to improve this field.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Barate, H. G., Ludovico, L., Pagani, E., and Scarabottolo, N. (2019). 5G Technology for Augmented and Virtual Reality in Education. Proceedings of the International Conference on Education and New Developments, 2019, 512–516.

Chan, C. K. Y., and Tsi, L. H. (2024). Will Generative AI Replace Teachers in Higher Education? A Study of Teacher and Student Perceptions. Studies in Educational Evaluation, 83, 101395. https://doi.org/10.1016/j.stueduc.2024.101395

Che, C., Li, X., Chen, C., He, X., and Zheng, Z. (2022). A Decentralized Federated Learning Framework via Committee Mechanism With Convergence Guarantee. IEEE Transactions on Parallel and Distributed Systems, 33, 4783–4800. https://doi.org/10.1109/TPDS.2022.3163727

Dao, N. N., Tu, N. H., Hoang, T. D., Nguyen, T. H., Nguyen, L. V., Lee, K., Park, L., Na, W., and Cho, S. (2024). A Review on New Technologies in 3GPP Standards for 5G Access and Beyond. Computer Networks, 245, 110370. https://doi.org/10.1016/j.comnet.2024.110370

El Koshiry, A., Eliwa, E., Abd El-Hafeez, T., and Shams, M. Y. (2023). Unlocking the Power of Blockchain in Education: An Overview of Innovations and Outcomes. Blockchain: Research and Applications, 4, 100165. https://doi.org/10.1016/j.bcra.2023.100165

Hwang, G. J., Xie, H., Wah, B. W., and Gašević, D. (2020). Vision, Challenges, Roles, and Research Issues of Artificial Intelligence in Education. Computers and Education: Artificial Intelligence, 1, 100001. https://doi.org/10.1016/j.caeai.2020.100001

Liu, C., Wang, L., and Liu, H. (2021). 5G Network Education System Based on Multi-Trip Scheduling Optimization Model and Artificial Intelligence. Journal of Ambient Intelligence and Humanized Computing, 1–14. https://doi.org/10.1007/s12652-021-03419-7

Liu, X., Faisal, M., and Alharbi, A. (2022). A Decision Support System for Assessing the Role of the 5G Network and AI in Situational Teaching Research in Higher Education. Soft Computing, 26, 10741–10752. https://doi.org/10.1007/s00500-022-07148-6

Luo, Y., and Yee, K. K. (2022). Research on Online Education Curriculum Resources Sharing Based on 5G and Internet of Things. Journal of Sensors, 2022, 9675342. https://doi.org/10.1155/2022/9675342

Rong, J. (2022). Innovative Research on Intelligent Classroom Teaching Mode in the 5G Era. Mo \bile Information Systems, 2022, 9297314. https://doi.org/10.1155/2022/9297314

Triplett, W. J. (2023). Addressing Cybersecurity Challenges in Education. International Journal of STEM Education and Sustainability, 3, 47–67.

Vaigandla, K., Radha, K., and Allanki, S. R. (2021). A Study on IoT Technologies, Standards, and Protocols. IBMRD's Journal of Management Research, 10, 7–14.

Yousif Yaseen, K. A. (2022). Digital Education: The Cybersecurity Challenges in the Online Classroom (2019–2020). Asian Journal of Computer Science and Technology, 11, 33–38.

Yu, J. (2024). Study of the Effectiveness of 5G Mobile Internet Technology to Promote the Reform of English Teaching in Universities and Colleges. Journal of Cases on Information Technology, 26, 1–21.

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? International Journal of Educational Technology in Higher Education, 16, 39. https://doi.org/10.1186/s41239-019-0171-0

Zhang, Y. (2023). Privacy-Preserving Zero-Trust Computational Intelligent Hybrid Technique for English Education Models. Applied Artificial Intelligence, 37, 2219560. https://doi.org/10.1080/08839514.2023.2219560