International Journal of Computer Applications Technology and Research
Volume 15–Issue 03, 114 – 117, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1503.1017

A Review of AI and Deep Learning Approaches for Content Moderation

Sanika Jadhav [1st]
Department of Computer Science and Engineering
D.Y. Patil College of Engineering and Technology
Kasaba Bawda, India

Ranjita S. Jadhav [2nd]
Department of Computer Science and Engineering
D.Y. Patil College of Engineering and Technology
Kasaba Bawda, India

Mrunal Pardeshi [3rd]
Department of Computer Science and Engineering
D.Y. Patil College of Engineering and Technology
Kasaba Bawda, India

Yash Mane [4th]
Department of Computer Science and Engineering
D.Y. Patil College of Engineering and Technology
Kasaba Bawda, India

Rajvardhan Patil [5th]
Department of Computer Science and Engineering
D.Y. Patil College of Engineering and Technology
Kasaba Bawda, India

Sakshi Patil [6th]
Department of Computer Science and Engineering
D.Y. Patil College of Engineering and Technology
Kasaba Bawda, India

Abstract: The rapid growth of user-generated content on social media platforms has made traditional moderation approaches, such as manual review and keyword-based filtering, increasingly ineffective and error-prone. To overcome these limitations, recent research has focused on automated content moderation techniques based on natural language processing and computer vision. Transformer-based models, including BERT and RoBERTa, enable deeper contextual and multilingual understanding of harmful text, significantly improving moderation accuracy while reducing false positives. For visual content, deep learning models such as convolutional neural networks and vision transformers support real-time detection of violent and inappropriate material. Furthermore, multimodal approaches that combine textual and visual information provide improved detection of complex toxic behavior. Recent advances in large language models further support adaptive and scalable moderation systems aligned with evolving online communities. Collectively, these approaches highlight the need for efficient, explainable, and scalable moderation frameworks to ensure safer digital environments.

Keywords: Content Moderation; Transformer Models; BERT; RoBERTa; Deep Learning; Multimodal Analysis; Toxicity Detection.

1. INTRODUCTION

Online communication, information sharing, and social interaction have all changed as a result of the explosive growth of user-generated content on digital platforms. Issues like hate speech, cyberbullying, false information, and other damaging content have also arisen as a result of this rise. Conventional moderation techniques, such as keyword-based filtering and manual review, are frequently laborious, slow, and unable to adequately capture multilingual expressions, context, or sarcasm.

Researchers and practitioners have been investigating AI-driven content moderation strategies more and more in an effort to overcome these constraints. Automatic identification of harmful text and visual content is made possible by methods that make use of computer vision and natural language processing (NLP). Advanced models enable context-aware, scalable, and multilingual moderation across various online platforms. Examples of these models include transformers for text and vision-based deep learning models for images.

Examining these techniques reveals the most recent developments, difficulties, and weaknesses in automated moderation systems. Managing informal and multilingual content, bias in datasets, making moral decisions, and system transparency are important concerns. Designing more dependable, socially conscious, and efficient moderation solutions for forums, e-commerce, social media, and other digital platforms starts with an understanding of these factors.

2. BACKGROUND AND MOTIVATION

The enormous amount of online content produced every second necessitates automated moderation. For moderators, manual approaches are time-consuming, prone to mistakes, and emotionally taxing. Early computational techniques, such as bag-of-words or keyword matching, were unable to recognize sarcasm or covert hate because they lacked contextual understanding.

Automated systems with multilingual processing, cross-platform adaptability, and semantic understanding are now possible thanks to developments in deep learning and natural language processing. These AI models are able to recognize subtle patterns in text, photos, and videos in addition to identifying overtly offensive content.

3. CONCEPTUAL OVERVIEW OF AI-BASED CONTENT MODERATION

This section presents a literature-derived conceptual overview of transformer-based content moderation techniques. The figure summarizes commonly adopted NLP preprocessing steps, transformer-based contextual representation learning, toxicity scoring, and decision-making workflows reported across existing studies.

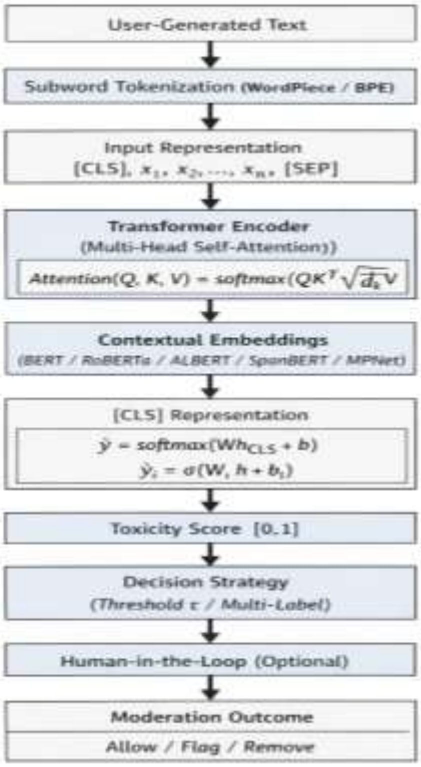

User-Generated Text
Subword Tokenization (WordPiece / BPE)
Input Representation
[CLS], x_1, x_2, ..., x_n, [SEP]
Transformer Encoder
(Multi-Head Self-Attention))
$Attention(Q, K, V) = softmax(QK^T\sqrt{d_k}V$
Contextual Embeddings
(BERT / RoBERTa / ALBERT / SpanBERT / MPNet)
[CLS] Representation
$\hat{y} = softmax(Wh_{CLS} + b)$
$\hat{y}_i = \sigma(W_i h + b_i)$
Toxicity Score [0.1]
Decision Strategy
(Threshold τ / Multi-Label)
Human-in-the-Loop (Optional)
Moderation Outcome
Allow / Flag / Remove

Figure 1: Literature-derived overview of transformer-based deep learning techniques for content moderation.

This illustration represents a synthesized view of surveyed approaches and does not denote a proposed system architecture.

4. TAXONOMY OF CONTENT MODERATION TECHNIQUES

4.1 Manual Moderation

Do not include headers, footers or page numbers in your submission. These Either community flagging or human moderators review the content. Although extremely precise and context-aware, it is not scalable for big platforms. [12] [14].

4.2 Rule-Based / Keyword Filtering

Recognizes preset patterns or words. Easy and quick, but unable to comprehend sarcasm, context, or changing language. Examples of methods include Regex-based detection and blacklist filters. [9] [10].

4.3 Machine Learning-Based

Automatically classifies content using conventional algorithms that have been trained on labeled datasets. Dependent on feature quality, but superior to rule-based methods. Techniques like Naïve Bayes, SVM, and Decision Trees are examples [3][10].

4.4 Deep Learning-Based

Uses contextual embeddings and neural networks to increase accuracy and improve semantic context comprehension. CNN, LSTM, BERT, and RoBERTa are a few examples of techniques [1][11].

4.5 Multimodal Moderation

Combines text, image, and audio signals to make more comprehensive moderation decisions. Example techniques: CLIP, Vision Transformers[13].

4.6 Hybrid Approaches

utilizes the advantages of each approach by combining rule-based, ML/DL models, and occasionally human moderation. Examples of methods include Human-in-the-loop AI and Rule + ML systems. [9] [12] [14].

5. LITERATURE REVIEW

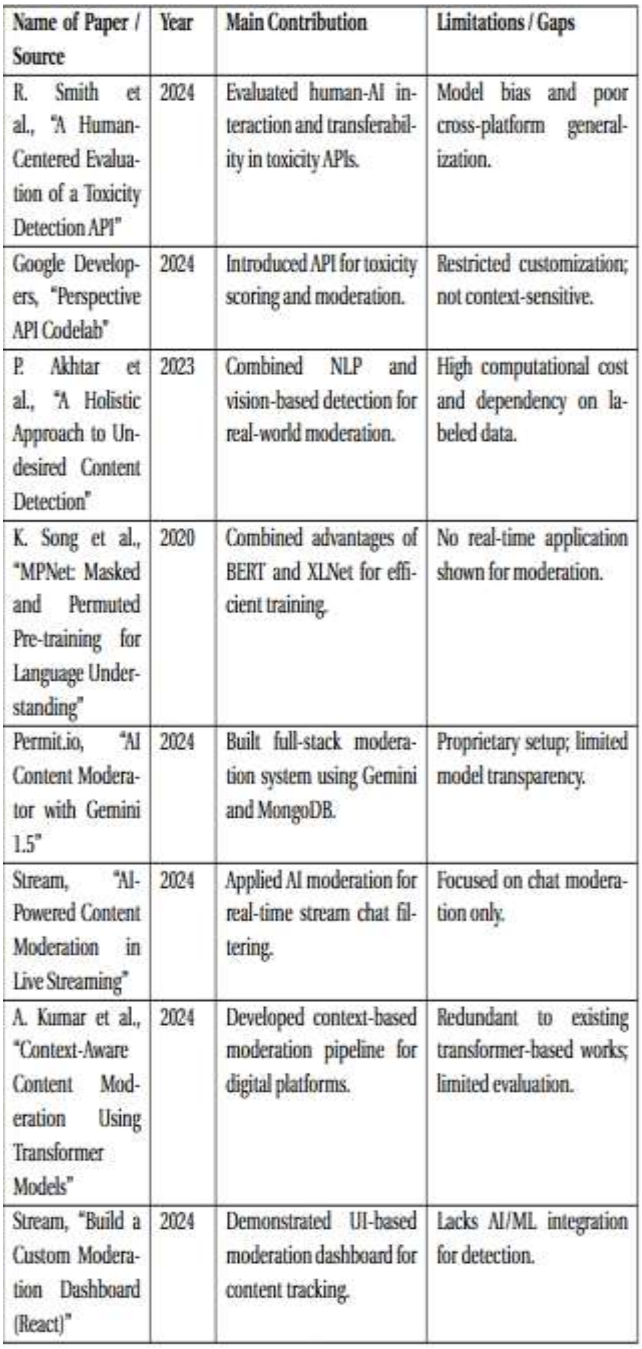
The rapid growth of digital platforms has made effective content moderation essential for maintaining safe online environments. Traditional moderation methods, largely dependent on human reviewers, are time-consuming, inconsistent, and difficult to scale. To address these limitations, AI-based content moderation systems using natural language processing (NLP) and machine learning have become widely adopted [1][2][3].

Recent research highlights the effectiveness of transformer-based models such as BERT, RoBERTa, ALBERT, and MPNet in detecting harmful and context-dependent content. By leveraging self-attention mechanisms, these models capture semantic relationships within text more effectively than traditional classifiers, leading to improved performance in identifying hate speech, misinformation, and offensive language across large datasets[7] [11].

Beyond technical performance, human-centered and ethical considerations remain central to AI-driven moderation. Research shows that models may exhibit biases and limited generalization, underscoring the need for transparency, continuous monitoring, and human oversight. Additionally, regional and cultural differences significantly influence how moderation systems are designed and enforced[12].

To balance automation and accountability, several studies propose hybrid moderation frameworks that integrate deep learning models with human-in-the-loop review mechanisms. Such systems, often supported by real-time dashboards, combine the scalability of AI with human judgment to ensure more reliable and responsible content moderation [1][13].

6. KEY LITERATURE FINDINGS



Name of Paper / Source	Year	Main Contribution	Limitations / Gaps
R. Smith et al., "A Human-Centered Evaluation of a Toxicity Detection API"	2024	Evaluated human-AI interaction and transferability in toxicity APIs.	Model bias and poor cross-platform generalization.
Google Developers, "Perspective API Codelab"	2024	Introduced API for toxicity scoring and moderation.	Restricted customization; not context-sensitive.
P. Akhtar et al., "A Holistic Approach to Undesired Content Detection"	2023	Combined NLP and vision-based detection for real-world moderation.	High computational cost and dependency on labeled data.
K. Song et al., "MPNet: Masked and Permuted Pre-training for Language Understanding"	2020	Combined advantages of BERT and XLNet for efficient training.	No real-time application shown for moderation.
Permit.io, "AI Content Moderator with Gemini 1.5"	2024	Built full-stack moderation system using Gemini and MongoDB.	Proprietary setup; limited model transparency.
Stream, "AI-Powered Content Moderation in Live Streaming"	2024	Applied AI moderation for real-time stream chat filtering.	Focused on chat moderation only.
A. Kumar et al., "Context-Aware Content Moderation Using Transformer Models"	2024	Developed context-based moderation pipeline for digital platforms.	Redundant to existing transformer-based works; limited evaluation.
Stream, "Build a Custom Moderation Dashboard (React)"	2024	Demonstrated UI-based moderation dashboard for content tracking.	Lacks AI/ML integration for detection.

Figure 2: Summary of Recent Studies on AI-based Content Moderation (2023–2025).

7. EVALUATION METRICS

The To assess model performance, the following metrics are commonly used:

- Accuracy: The proportion of correctly predicted samples out of all samples.
- Precision: Measures how many of the items labeled as "hate/offensive" were actually correct.
- Recall: Measures how many of the actual "hate/offensive" items were correctly identified.
- F1-Score: The harmonic mean of precision and recall. Balances the two.
- Confusion Matrix: A table showing TP, TN, FP, FN counts.

8. CHALLENGES AND LIMITATIONS

The Despite significant advancements, several challenges persist:

- Contextual Ambiguity: Difficulty detecting sarcasm or implicit hate.
- Multilingual Complexity: Limited resources for low-resource languages.
- Dataset Bias: Models may inherit cultural or labelling bias.
- Real-time Moderation: High latency in live stream moderation.
- Ethical Concerns: Balancing free speech with censorship risks.
- Explainability: Most deep learning models lack transparency in decision-making.

9. OBSERVATIONS AND IMPLEMENTATION GUIDANCE

Although there are still noticeable gaps between research models and practical application, it is clear from examining previous studies and implementations that AI-powered content moderation has advanced considerably. When implemented in real-world settings, many systems encounter difficulties despite performing well in research settings. According to the study in [10], live-streaming platforms can efficiently identify harmful or toxic messages through real-time text analysis using APIs like OpenAI and Perspective API. This study demonstrates how AI works best when paired with rule-based decision logic and user behavior tracking, which enables dynamic moderation system adaptation.

From a broader perspective, the report discussed in [12] emphasizes that technological efficiency alone is insufficient for responsible moderation. It underlines the importance of ethical and contextual decision-making, especially in linguistically diverse regions like India. The report also identifies key challenges such as the lack of region-specific datasets, limited transparency in AI-based decisions, and inadequate policy alignment for moderation fairness.

Overall, this survey indicates that while tools like those presented in [10] demonstrate the technical feasibility of automated moderation, successful large-scale adoption demands integrating context-aware models, inclusive multilingual datasets, and transparent decision framework

10. FUTURE RESEARCH DIRECTIONS

To enhance the effectiveness and fairness of moderation systems, future work may focus on:

- Multilingual : Developing universal language models that understand cultural nuances.
- Explainable AI (XAI): Making moderation decisions transparent and interpretable.
- Federated and Privacy-preserving Learning: Training models without sharing raw data.

- Multimodal Fusion: Integrating text, image, audio, and video analysis.
- Ethical Frameworks: Establishing standardized, fair, and transparent AI governance.
- Continuous Learning: Adapting models to emerging slangs, memes, and contextual shifts.

11. CONCLUSION

This survey provides a comprehensive overview of content moderation systems, tracing their evolution from manual and rule-based techniques to advanced AI-driven approaches leveraging natural language processing and deep learning. Recent transformer-based architectures, such as RoBERTa and MPNet, have demonstrated strong performance in multilingual and context-aware moderation tasks [1], [3], [11]. Despite these advancements, several challenges remain, including dataset bias, ethical decision-making, model explainability, and scalability in real-world deployments [12], [14].

Existing studies emphasize the importance of scalable infrastructures and transparent decision-making frameworks to support real-time applications, such as live-stream moderation dashboards and toxicity detection systems [9], [10], [13]. Furthermore, social and regional inclusivity continues to be a critical concern, particularly in linguistically diverse regions such as India, where fairness and cultural sensitivity play a significant role in content moderation systems [12].

Based on the literature published between 2023 and 2025, this survey highlights that future content moderation systems should prioritize contextual awareness, fairness, and transparency to ensure responsible deployment [14]. Achieving sustainable and trustworthy moderation ecosystems will require the integration of hybrid AI–human moderation approaches, region-specific datasets, and well-defined ethical governance frameworks [9], [13].

12. REFERENCES

[1] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. arXiv preprint arXiv:2004.09297.

[2] Google Developers. 2024. Perspective API Codelab. Available: https://developers.google.com/codelabs/setup-perspective-api

[3] P. Akhtar et al. 2023. A Holistic Approach to Undesired Content Detection in the Real World. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 12.

[4] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. Transactions of the Association for Computational Linguistics, vol. 8, pp. 64–77.

[5] ExoAPI. 2024. Leveraging Large Language Models for Content Moderation in React. Available: https://exoapi.dev/blog/ai-powered-content-moderation-react-large-language-models

[6] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. In Proceedings of the International Conference on Learning Representations (ICLR).

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, pp. 4171–4186.

[8] Y. Liu et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[9] Permit.io. 2024. AI Content Moderator: Why and How to Build One with Gemini 1.5. Available: https://www.permit.io/blog/ai-powered-content-moderation-gemini-and-nextjs

[10] arXiv. 2024. Context-Aware Content Moderation Using Transformer Models for Detecting Harmful Digital Content. Available: https://arxiv.org/abs/2412.16114

[11] A. Kumar et al. 2024. Context-Aware Content Moderation Using Transformer Models for Detecting Harmful Digital Content. Available: https://www.researchgate.net/publication/391017674

[12] Social Media Matters. 2024. Is Content Moderation Working in India? Social Media Matters Research Report. Available: https://www.socialmediamatters.in/our-work/research/is-content-moderation-working-in-india

[13] Stream. 2024. Build a Custom Moderation Dashboard (React). Available: https://github.com/GetStream/moderation-dashboard-example

[14] M. Sap et al. 2023. A Human-Centered Evaluation of a Toxicity Detection API: Testing Transferability and Unpacking Latent Attributes. In Proceedings of the ACM Conference.