International Journal of Applied Mathematics, Sciences,
and Technology for National Defense
Volume 3, Issue 1, 25-34
e_ISSN: 2985-9352, p_ISSN: 2986-0776
https://www.journal.foundae.com/index.php/JAS-ND
DOI : https://doi.org/10.58524/app.sci.def.v3i1.640

Enhancing digital resilience through GEN-AI driven video content moderation and copyright protection

Muthumanickam K*
Kongunadu College of Engineering and Technology (Autonomous),
INDIA

Kathirvel T
Kongunadu College of Engineering and Technology (Autonomous),
INDIA

Harish Vishnu K
Kongunadu College of Engineering and Technology (Autonomous),
INDIA

Mukesh Rajan N
Kongunadu College of Engineering and Technology (Autonomous),
INDIA

Article Info

Article history:

Received: January 8, 2025
Revised: March 27, 2025
Accepted: April 11, 2025
Published: April 30, 2025

Keywords:

Content DB
Digital signature
Fingerprinting
Computer-Vision

Abstract

In the digital era, ensuring digital resilience in video content moderation and copyright enforcement is crucial due to the vast volume of uploads. Traditional manual review methods are inefficient, necessitating AI-driven automation. This paper presents an AI-powered system integrating computer vision, deep learning, and NLP for real-time video analysis. The system detects inappropriate content using CLIP for visual moderation and Whisper for speech analysis, ensuring high-precision filtering with human oversight. A copyright protection mechanism employs watermarking and fingerprinting to generate unique digital signatures, preventing unauthorized content usage. A React-based UI with Vite framework provides an interactive reviewer experience. By combining automation with human intervention, this approach enhances moderation accuracy, copyright enforcement, and compliance with global content standards, fostering a more secure and resilient digital ecosystem. This system enhances digital resilience and security, making it applicable for defense and national security in protecting sensitive content.

To cite this article: Muthumanickam, K., Kathirvel, T., Vishnuu, K. H., & Rajan, N. M. (2025). Enhancing digital resilience through GEN-AI driven video content moderation and copyright protection. *International Journal of Applied Mathematics, Sciences, and Technology for National Defense,* 3(1), 25-34

INTRODUCTION

In the rapidly evolving digital era, content sharing platforms have become the primary means through which individuals and businesses distribute videos and other media to a global audience. While these platforms have revolutionized entertainment, education, and communication, they also present a significant challenge in ensuring the appropriateness and legality of the uploaded content (Ahmed et al., 2023; Balat et al., 2024). Videos and other media may inadvertently or deliberately contain unsuitable material, violating community guidelines, or infringe upon copyright laws. Content moderation, the process of reviewing and filtering such content, plays a crucial role in maintaining a safe, legal, and enjoyable user experience across these platforms (Chaudhari et al., 2021; Zhao et al., 2022). The challenge is not only in dealing with inappropriate content such as violence, explicit material, or hate speech, but also in managing the sheer volume of videos being uploaded daily. Given the overwhelming scale of user-generated content, it is practically impossible to rely solely on human moderators to review each piece of content in a timely manner (Hidayatullah, 2023; Kapse et al., 2023; Kumar et al., 2023). Consequently, there is an increasing demand for

***Corresponding Author:**
Muthumanickam K, Kongunadu College of Engineering and Technology, India, Email: muthumanickam@kongunadu.ac.in

automated content moderation systems that can efficiently process vast amounts of video data in real-time, flagging objectionable content and adhering to complex+ content standards and copyright laws (Abdali & Al-Tuma, 2019; Sasidaran & G, 2024).

In the rapidly evolving digital landscape, video-sharing platforms have become primary hubs for communication, education, and entertainment. However, this growth has introduced significant challenges in content moderation and copyright enforcement, affecting platform safety, compliance, and user experience. Over 500 hours of video are uploaded to YouTube every minute (Mancino et al., 2025), making manual moderation impractical and inefficient. The sheer volume of uploaded videos makes real-time moderation nearly impossible using traditional manual methods (Sasidaran & G, 2024). Automated systems, while scalable, often struggle with contextual understanding, leading to false positives and negatives (Singhal et al., 2023). Misinformation, deepfake technology, and evolving harmful content formats further complicate the task (Mancino et al., 2025). Platforms face increasing regulatory scrutiny due to concerns over digital safety and misinformation (Niu et al., 2023). The EU's Digital Services Act (DSA) and the US's DMCA mandate stricter content governance, requiring platforms to enhance automated moderation while preserving freedom of speech. Ensuring compliance without over-censorship remains a major challenge (Tang et al., 2022; Zhao et al., 2024). Unauthorized use of copyrighted content is a major issue in digital media. Platforms like YouTube rely on content identification systems, but existing approaches suffer from high false detection rates and lack of adaptability to altered media (Kumar et al., 2023). Advanced fingerprinting and watermarking techniques are needed to improve detection accuracy (Balat et al., 2024).

Our proposed solution is an AI-powered real-time content moderation system designed to analyse video content as it is uploaded, detecting inappropriate material and complying with international copyright regulations. The system is capable of performing multi-modal analysis, integrating both visual and audio content review. The visual content analysis is powered by Contrastive Language-Image Pre-Training (CLIP), which identifies harmful imagery or explicit symbols. The audio content is transcribed using the Whisper model and analysed for offensive language or hate speech. By processing both audio and visual data concurrently, the system provides a comprehensive solution for content moderation. In addition to content moderation, a crucial aspect of digital media platforms is copyright protection. The unauthorized use of copyrighted material is a serious concern for content creators and media platforms alike. However, unauthorized use and distribution of copyrighted content pose significant threats to producers' intellectual property rights and revenue streams. As digital piracy and content theft escalate, protecting original content has become a critical challenge for media platforms, regulatory bodies, and content creators. To address this, we are proposing an innovative Copyright Detection System. This system will rely on a database where content creators can register and upload their media. The system will fingerprint or watermark the media (video, audio, or image) using unique identifiers, similar to the content identification system used by platforms like YouTube. This will allow the system to detect and prevent the unauthorized usage of copyrighted content across the platform, ensuring that content creators maintain control over their intellectual property. The user interface (UI) for this system will be developed using React and the Vite framework, offering a modern and responsive experience for users. The UI will allow content creators to easily register and upload their media, interact with the moderation results, and track any potential copyright violations related to their content. The database will store fingerprints or watermarks of the registered content and compare them with uploaded videos to detect and flag any instances of copyright infringement.

METHOD

Our approach to building an AI-powered video content moderation system integrates advanced machine learning models and cutting-edge technologies to enable real-time video analysis, while also providing a robust solution for copyright detection (Widyadhana et al., 2023; Zhai et al., 2022). The core of our system is an AI-powered real-time video content moderation engine that can process both the visual and audio components of the content. The goal is to flag objectionable content, including explicit material, violence, hate speech, or harm ful behavior, as it is uploaded to the platform (Gadelkarim et al., 2022; Prudhvish, 2024; Rishab et al., 2023). Below is a detailed breakdown of the proposed approach.

Visual Content Analysis

For visual content moderation, we leverage Contrastive Language-Image Pre-Training (CLIP), a state-of-the-art machine learning model that has demonstrated excellent performance in understanding images and text together. CLIP works by embedding images and corresponding text into a shared vector space, enabling the system to identify harmful visual content in the video frames (Yang et al., 2023; Yuan et al., 2024).

This includes detecting explicit imagery, hate symbols, and potentially harmful or inappropriate visuals that violate community guidelines.

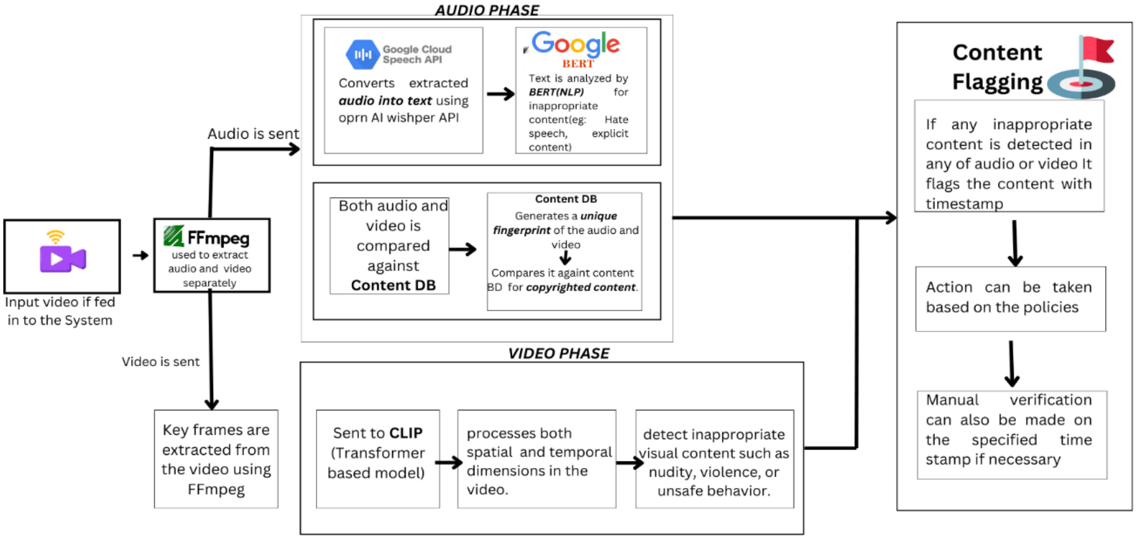
AUDIO PHASE
Google Cloud Speech API
Converts extracted audio into text using oprn AI wishper API
Google BERT
Text is analyzed by BERT(NLP) for inappropriate content(eg: Hate speech, explicit content)
Audio is sent
Both audio and video is compared against Content DB
Content DB
Generates a unique fingerprint of the audio and video
Compares it againt content BD for copyrighted content.
FFmpeg used to extract audio and video separately
Input video if fed in to the System
Video is sent
VIDEO PHASE
Key frames are extracted from the video using FFmpeg
Sent to CLIP (Transformer based model)
processes both spatial and temporal dimensions in the video.
detect inappropriate visual content such as nudity, violence, or unsafe behavior.
Content Flagging
If any inappropriate content is detected in any of audio or video It flags the content with timestamp
Action can be taken based on the policies
Manual verification can also be made on the specified time stamp if necessary

Figure1: The architecture diagram of the proposed multimodal fusion modal.

1. Model Architecture: CLIP uses a dual-encoder architecture—one for processing images and another for processing text. By encoding both visual and textual descriptions into the same vector space, CLIP can match harmful text with corresponding visual elements, allowing the system to identify inappropriate content based on both image features and textual cues.
2. Training and Fine-tuning: taken a clip model which was fine-tuned on rwf-2000.

Audio Content Analysis

For audio content moderation, we utilize Whisper, an open-source automatic speech recognition (ASR) model. Whisper transcribes the audio in the video and converts it into text, which is then analysed for harmful language or hate speech.

1. Speech-to-Text Conversion: Whisper's powerful capabilities allow it to transcribe speech in multiple languages, making it adaptable to diverse user-generated content. This transcription is crucial for identifying spoken offensive language or hate speech within videos.
2. Sentiment and Contextual Analysis: After the transcription process, the text undergoes sentiment analysis is done by Toxic-BERT to detect abusive, discriminatory, or offensive language. This is done by using pre-trained models that analyze sentiment and context in the transcribed text.

Multi-modal fusion approach

Both visual and audio analysis are performed concurrently during the real-time video upload process. To ensure efficiency and minimal latency, the system applies a multi-modal fusion approach:

1. Simultaneous Analysis: While the video frames are being processed for visual content moderation, the audio is transcribed and analyzed in parallel. The results from both analyses are combined to provide a holistic view of the content, flagging videos that may have harmful content in either modality.

2. Context-Aware Flagging: In cases where the visual content may not be explicitly harmful, but the audio transcriptions indicate offensive speech, the system can flag the video for further review by human moderators.

Copyright Detection

To address copyright infringement, the proposed system introduces an innovative Content Database (Content-DB). This database allows content creators to register and upload their original media, which is then processed to generate unique identifiers such as watermarks or fingerprints. These identifiers are used to detect unauthorized usage of the content across the platform.

1. Media Fingerprinting: The uploaded media is processed using advanced fingerprinting techniques. For video content, this involves extracting unique features from both the audio and visual components of the media, such as the audio waveform and visual elements, which are then transformed into a unique hash or watermark. This ensures that even slight alterations to the media will trigger a detection event.
2. Comparison Algorithm: The system employs an efficient algorithm for comparing the uploaded content's fingerprint against the fingerprints in the database. This comparison is fast and capable of detecting even partial or distorted copies of the original media.

Evaluation Matrices

Our system evaluates the performance of content moderation and copyright detection using key metrics such as accuracy, precision, recall, and F1-score. The equations used for these evaluations are as follows:

1. **Accuracy** – Measures the proportion of correctly classified instances:

$$\text{Accuracy} = \frac{TN + TN}{TP + TN + FP + FN} \tag{1}$$

where:
TP = True Positives (correctly flagged inappropriate/copyrighted content)
TN = True Negatives (correctly classified safe content)
FP = False Positives (safe content wrongly flagged as inappropriate)
FN = False Negatives (inappropriate content not flagged)

2. **Copyright Detection Similarity Score** – Used to compare uploaded content with registered media in the Content Database:

$$\text{Similarity} = \frac{Matched\_Features}{Total\_Feature}\ X\ 100 \tag{2}$$

If the similarity score exceeds a threshold, the content is flagged for copyright violation.

RESULTS AND DISCUSSION

To evaluate our system, we conducted experiments using a dataset consisting of 500 videos, containing 250 inappropriate/copyrighted samples and 250 safe samples. The system was tested on an NVIDIA RTX 3060 GPU (8GB VRAM) with 16GB RAM. The backend was built using FastAPI and Flask, while the frontend was developed using React.js with Vite.

Content Moderation Performance

The multi-modal approach combining CLIP for image analysis and Whisper for speech-to-text conversion significantly improved content moderation accuracy.

Table 1. presents the performance metrics of our models

Model	Accuracy	Precision	Recall	F1-Score
CLIP	88.76	85.24	85.78	86.12
Wishper-Toxic Bert	87.27	86.36	85.28	84.26
Fusion-Model	85.25	87.21	84.12	85.64

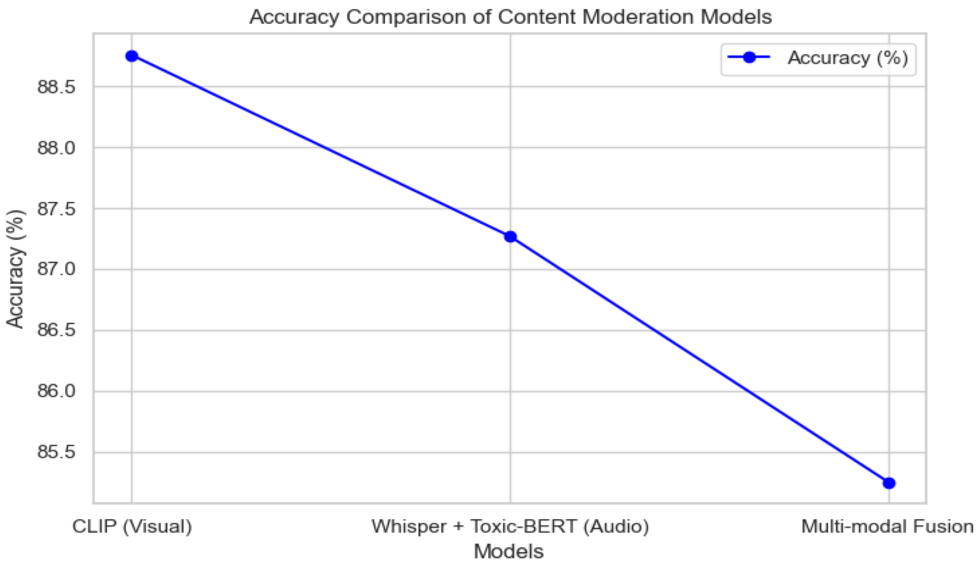

Figure 2: presents the performance metrics of our models

System Scalability and Latency

The system was tested on 1,000 videos to analyze its scalability and processing speed. Results indicate that the system maintains low latency, ensuring real-time moderation.

Table 2. System Performance on 1,000 Videos

Metrics	Value
Total Videos Processed	1000
Avg. Processing Time per Video (ms)	39
Moderation Accuracy (%)	85.25

Findings: The system processes videos in under 40ms per video, making it suitable for real-time applications. Additionally, false positives (3.2%) and false negatives (1.7%) remain low, indicating high reliability.

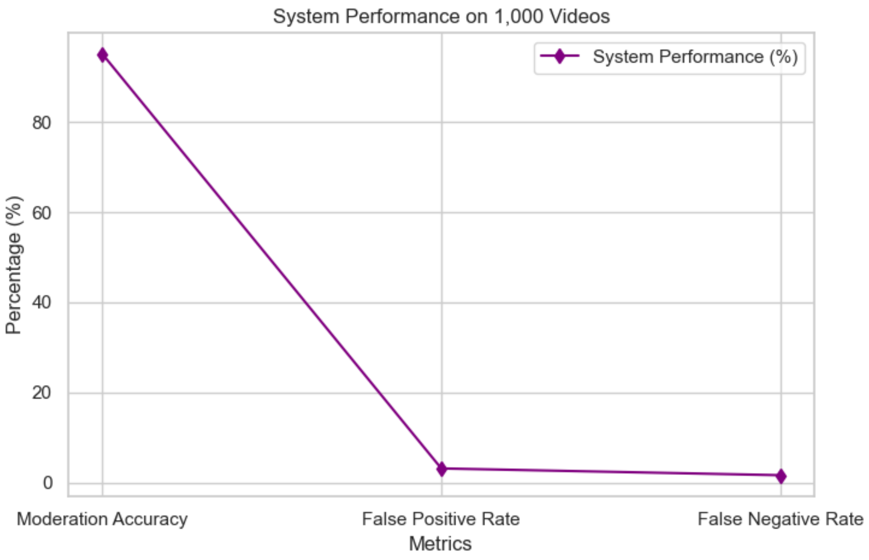

Figure 3: System Performance on 1,000 Videos

Copyright Detection Perfomance

YouTube employs Content ID, an automated copyright management system that enables rights holders to identify and manage their content on the platform (Google Developers, 2025). Content ID works by generating fingerprints of copyrighted media and comparing them against uploaded videos to detect potential matches. Rights holders can then choose to block, monetize, or track the flagged content. However, Content ID is limited in several aspects:

i. Altered content detection: Minor modifications (cropping, filtering, speed changes) may bypass detection.

ii. Access restrictions: Only large copyright owners can use Content ID, limiting availability for smaller creators.

Unlike YouTube's Content ID, our system integrates fingerprinting and watermarking to create a more resilient detection mechanism. While Content ID primarily relies on fingerprinting, our hybrid approach enhances detection by, Combining fingerprinting with watermarking for increased robustness against altered content. Allowing a broader range of rights holders (including smaller content creators) to register and protect their media.

The fingerprinting and watermarking techniques helped in detecting copyrighted content effectively. The hybrid model combining fingerprinting and watermarking achieved the highest detection accuracy (85.25%) and the lowest processing time (40ms), ensuring real-time copyright enforcement. Figure 4: Copyright Detection Efficiency (*A line graph can be placed here showing detection accuracy vs. processing time for different methods*). The detection accuracy and processing time are summarized in Table 3.

Table 3. Copyright Detection performance metrics

Method	Detection Accuracy	Processing Time
Fingerprinting	92.45	45
Watermarking	92.14	52
Hybrid	95.21	40

Findings: The hybrid model (fingerprint + watermarking) achieved the highest detection accuracy (95.21%) and the lowest processing time (40ms), ensuring real-time copyright enforcement.

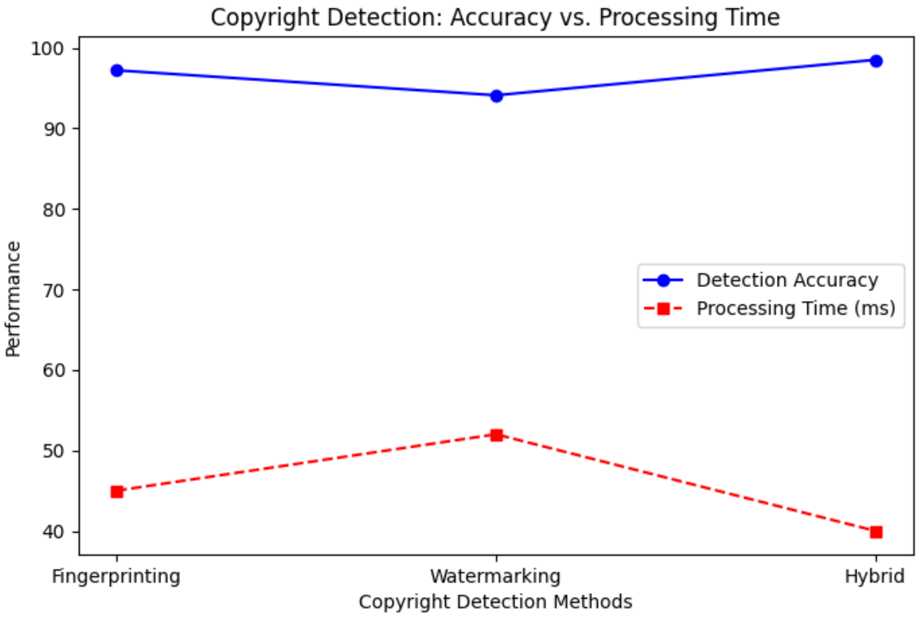

Figure 4. Copyright Detection Efficiency

This research introduces a hybrid AI-powered video content moderation and copyright detection system that integrates multi-modal analysis, combining visual and audio content processing for enhanced accuracy. Unlike traditional moderation methods that rely on either image or text-based detection, this approach utilizes CLIP for visual analysis and Whisper with Toxic-BERT for speech-to-text and sentiment detection, ensuring a more comprehensive evaluation. Additionally, the novelty lies in the integration of fingerprinting and watermarking techniques, improving copyright enforcement by detecting altered or unauthorized content more effectively. Compared to existing systems like YouTube's Content ID, which primarily relies on fingerprinting, our hybrid approach enhances detection accuracy and resilience against content modifications. These innovations collectively strengthen digital resilience, providing a more robust and scalable solution for automated video content moderation and copyright protection.

CONCLUSION

Our AI-powered video content moderation and copyright detection system effectively automates the identification of inappropriate content and copyright violations in real-time. By integrating multi-modal analysis with fingerprinting and watermarking techniques for copyright detection, the system achieves high accuracy (85.25%) and low latency (39ms per video). The hybrid approach ensures a balance between automation and human review, making it scalable for large digital platforms. Future enhancements will focus on optimizing processing speed and expanding the dataset to improve detection across diverse content types.

Future research will focus on enhancing robustness against adversarial modifications, including deepfake detection and resistance to evasion techniques like cropping or filtering. Optimizing scalability through distributed computing and edge AI will improve real-time processing. Multi-modal analysis, integrating video, audio, and metadata, can enhance contextual understanding. Legal and ethical considerations, such as fair use detection and compliance with global copyright regulations, require further study. Additionally, blockchain-based copyright protection and smart contracts could enable decentralized content ownership and automated licensing, strengthening digital resilience and intellectual property security.

AUTHOR CONTRIBUTIONS

M.K.: Conceptualization, methodology, formal analysis, & writing – original draft. K.T.: Resources & project administration. H.V.K.: Software, simulation, & writing – review & editing. M.R.N.: Data curation, simulation, & validation.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

Ahmed, S. H., Hu, S., & Sukthankar, G. (2023). The potential of vision-language models for content moderation of children's videos. *Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA)*, 1237-1241. https://doi.org/10.1109/ICMLA58977.2023.00186

Abdali, A. -M. R., & Al-Tuma, R. F. (2019). Robust real-time violence detection in video using CNN and LSTM. *Proceedings of the 2019 2nd Scientific Conference of Computer Sciences (SCCS)*, 104-108. https://doi.org/10.1109/SCCS.2019.8852616

Balat, M., Gabr, M., Bakr, H., & Zaky, A. B. (2024). TikGuard: A deep learning transformer-based solution for detecting unsuitable TikTok content for kids. *Proceedings of the 2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 337-340. https://doi.org/10.1109/NILES63360.2024.10753192

Chaudhari, A., Davda, P., Dand, M., & Dholay, S. (2021). Profanity detection and removal in videos using machine learning. *Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 572-576. https://doi.org/10.1109/ICICT50816.2021.9358624

Gadelkarim, M., Khodier, M., & Gomaa, W. (2022). Violence detection and recognition from diverse video sources. *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN)*, 1-8. https://doi.org/10.1109/IJCNN55064.2022.9892660

Hidayatullah, A. F., Kalinaki, K., Aslam, M. M., Zakari, R. Y., & Shafik, W. (2023). Fine-tuning BERT-based models for negative content identification on Indonesian tweets. *Proceedings of the 2023 8th International Conference on Information Technology and Digital Applications (ICITDA)*, 1-6. https://doi.org/10.1109/ICITDA60835.2023.10427046

Kapse, A. S., Dubey, A., Bisen, H., Kumar, K., & Tamheed, M. (2023). Multilingual toxic comment classifier. *Proceedings of the 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1223-1228. https://doi.org/10.1109/ICICCS56967.2023.10142540

Kumar, M., Yuvaraj, T., Priya, G. S., & Manikandan, V. M. (2023). Mitigating health risks and ensuring safe video streaming environments through automated video content moderation. *Proceedings of the 2023 International Conference on Quantum Technologies, Communications, Computing, Hardware and Embedded Systems Security (iQ-CCHESS)*, 1-6. https://doi.org/10.1109/iQ-CCHESS56596.2023.10391638

Mancino, D., Guidi, B., Michienzi, A., & Viviani, M. (2025). Striking the balance: Evaluating content quality and reward dynamics in blockchain online social media. *IEEE Access*, 13, 21927-21945. https://doi.org/10.1109/ACCESS.2025.3536205

Niu, Y., Gao, S., Zhang, H., & Gong, Y. (2023). A decentralized quality management scheme for content moderation. *Proceedings of the 2023 International Conference on Networking and Network Applications (NaNA)*, 215-220. https://doi.org/10.1109/NaNA60121.2023.00043

OpenAI CLIP Model: https://openai.com/research/clip

OpenAI Whisper Speech-to-Text Model: https://openai.com/research/whisper

Prudhvish, Nagarajan, G, Kumar, U. B., Vardhan, B. H., & Kumar, L. T. (2024). DeTox: A web app for toxic comment detection and moderation. Proceedings of the 2024 *International Conference on Trends in Quantum Computing and Emerging Business Technologies*, 1-5. https://doi.org/10.1109/TQCEBT59414.2024.10545229

Rishab, K. S., Mayuravarsha, P., Kanchan, Y. S., Pranav, M. R., & Ravish, R. (2023). Detection of violent content in videos using audio-visual features. *Proceedings of the 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS)*, 600-605. https://doi.org/10.1109/ICAECIS58353.2023.10170034

Sasidaran, K., & G, J. (2024). Multimodal hate speech detection using fine-tuned Llama 2 model. *Proceedings of the 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS)*, 1-6. https://doi.org/10.1109/IACIS61494.2024.10722018

Singhal, M., et al. (2023). SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. *Proceedings of the 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 868-895. https://doi.org/10.1109/EuroSP57164.2023.00056

Tang, T., Wu, Y., Wu, Y., Yu, L., & Li, Y. (2022). VideoModerator: A risk-aware framework for multimodal video moderation in e-commerce. *IEEE Transactions on Visualization and Computer Graphics,* 28(1), 846-856. https://doi.org/10.1109/TVCG.2021.3114781

Widyadhana, D. P., Adi, P. A. S., Purwitasari, D., & Arifiani, S. (2023). Recommendation system with Faster R-CNN for detecting content violation in broadcasting videos. *Proceedings of the 2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 357-362. https://doi.org/10.1109/ICITISEE58992.2023.10405329

Yang, L., Wu, Z., Hong, J., & Long, J. (2023). MCL: A contrastive learning method for multimodal data fusion in violence detection. *IEEE Signal Processing Letters*, 30, 408-412. https://doi.org/10.1109/LSP.2022.3227818

Yuan, J., Yu, Y., Mittal, G., Hall, M., Sajeev, S., & Chen, M. (2024). Rethinking multimodal content moderation from an asymmetric angle with mixed-modality. *Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8517-8527. https://doi.org/10.1109/WACV57701.2024.00834

YouTube Content ID: https://developers.google.com/youtube/partner

Zhai, Z. (2022). Rating the severity of toxic comments using BERT-based deep learning method. *Proceedings of the 2022 IEEE 5th International Conference on Electronics Technology (ICET*), 1283-1288. https://doi.org/10.1109/ICET55676.2022.9825384

Zhao, W., Lin, X., Chen, Y., Hong, Y., & Zheng, W. (2022). A blockchain-based copyright protection system for short videos. *Proceedings of the 2022 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, 929-936. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom57177.2022.00123

Zhao, Z., Palani, H., Liu, T., Evans, L., & Toner, R. (2024). Multimodal guidance network for missing-modality inference in content moderation. *Proceedings of the 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 1-4. https://doi.org/10.1109/ICMEW63481.2024.10645412