International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

AI-Powered Content Moderation and Harmful Content Detection in Streaming Media

Rashi Mahani[1], Sujal Chauhan[2] , Dr. Deepika Bansal[3], Dr. Tejna Khosla[4]

1,2,3,4 Information Technology ,Maharaja Agrasen Institute of Technology , Rohini Sec-22, Delhi, India

---

Abstract - Online video platforms like YouTube are flooded with new uploads every minute, which makes it hard to control harmful or sensitive content such as violence, nudity, or abusive speech. To tackle this issue, this study introduces SafeStream.ai, a simple web tool that helps viewers watch safer versions of videos. The system automatically spots and skips inappropriate parts using a mix of computer vision, text analysis, and emotion detection. It's built with three main parts: a Next.js interface for users, a Flask service that runs the AI model, and a Neon Postgres database to save results. When a user adds a YouTube link, the system breaks the video into frames, checks each one for unsafe visuals via computer vision, and records time ranges where such content appears. Tests so far show about 67% accuracy in identifying harmful material. SafeStream.ai aims to give users control over what they see online while keeping the experience smooth and accessible. Future improvements will focus on increasing accuracy, speeding up processing, and training with more diverse data. The project also plans to go open-source to invite community input and make the internet a safer place for everyone.

Key Words: Video Platforms, YouTube, Objectionable Content, AI, Computer Vision, Flask, Next.js , Open Source

1. Introduction

Streaming sites have now become the go-to place for watching and sharing videos. Every day, millions of clips are uploaded and viewed, especially on YouTube, where almost anyone can post content. This flood of material has made moderating what people see an enormous challenge. YouTube does use automated tools and human reviewers, but the sheer amount of new uploads means that many harmful videos still slip through the cracks.To help reduce this problem, we built SafeStream.ai, a simple website where users can paste a YouTube link and get back a cleaner version of the video. The tool automatically scans and skips the parts that contain violence, nudity, or offensive language. This paper explains how SafeStream.ai works, how it was built, and how well it performs in making online viewing safer.Because of the massive number of videos on YouTube, it's unrealistic to expect humans to review them all.Even with modern moderation tools, many subtle or context-based problems still go unnoticed, allowing harmful content to circulate. These gaps have real consequences young audiences and other sensitive users may end up viewing disturbing visuals or false information, which can harm their mental well-being and online experience. The concept of SafeStream.ai emerged from the need to find a better way to tackle these challenges. From hate speech and graphic violence to explicit imagery, the spread of such content worries users, researchers, and regulators alike. Surveys also show that people of different ages disagree on how strictly online content should be controlled, as seen in Figure 1 .

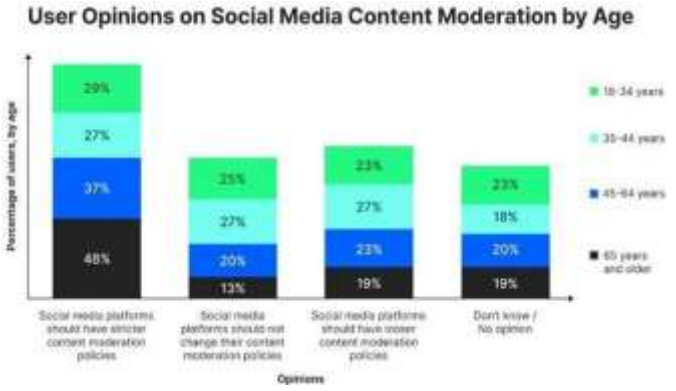
In 2022, a Pew Research survey revealed that 64% of YouTube users had encountered inappropriate content on the platform, despite extensive moderation efforts [3]. This statistic highlights the growing gap between user expectations for a safe online environment and the limitations of current content moderation systems. Social media platforms, particularly those hosting vulnerable groups like women, children, and minorities, continue to struggle with harmful content. In 2023, a third of internet users expressed that platforms should ban harmful content, focusing on extremist material, hate speech, and violence. Among the most vulnerable to this gap are children and teenagers, who form a large segment of YouTube's audience. In fact, a 2021 survey by Common Sense Media reported that 78% of parents are deeply concerned about their children's exposure to inappropriate material while using the platform [4]. Their concerns are not unfounded; younger users are more impressionable and, thus, more susceptible to the lasting negative effects of encountering objectionable content online. Complicating matters further, it's not just a matter of moderating more — but moderating better. Research by Gorwa and Guilbeault (2020) demonstrated that automated moderation tools often struggle to strike the right balance [5]. Many content models are inaccurate. They can't always see the difference between harmful material and normal talk about a serious topic. Sometimes even useful videos , like ones explaining history or medical issues get deleted, while some hate speech or violent clips stay online.The job also affects the people who do it. In 2019, Roberts wrote that moderators often struggle with stress and trauma because they have to view disturbing content every day [6]. It shows that depending only on people for this work isn't practical or fair in the long run.Because of problems like these, researchers have started to look for other ways. Chandrasekharan et al. (2018) found that when users can choose their own filters, online spaces become safer and more welcoming [7]. Everyone's comfort level is different, so letting people control what they see makes sense.Overall, today's systems don't really protect users or moderators. We need smarter ideas such as SafeStream.ai that can handle content more effectively and make watching videos online safer for everyone.

2. Body of Paper

The body of this paper presents the design, methodology, and evaluation of **SafeStream.ai**, an AI-powered system developed to identify and skip harmful content within YouTube videos. As described in Sec. 2, existing moderation systems struggle with contextual understanding, scalability, and real-time detection, which often leads to gaps in user safety. To address these limitations, SafeStream.ai integrates computer vision, natural language processing, and timestamp-level content filtering, allowing users to receive a sanitized viewing experience. The system architecture, detailed in Sec. 3.1, combines a Next.js-based frontend, a Flask-driven AI microservice, and a Neon Postgres database to ensure efficient

processing and storage of moderation results. The methodology outlined in Sec. 3.3 enables frame extraction, object detection, sentiment analysis, and automated skipping of unsafe segments during playback. Implementation results discussed in Sec. 5 demonstrate that the MobileNetV3-Small model achieves a baseline accuracy of 67% in detecting violent, explicit, or unsafe content, validating the system's capability while leaving scope for improvement. Overall, the findings confirm that SafeStream.ai provides a practical and user-controlled approach to content moderation, offering enhanced safety while maintaining viewing continuity.

Figure -1: User Opinion On Social Media Content Moderation



This research introduces SafeStream.ai, a flexible AI-based platform built to help users find and skip harmful parts within YouTube videos. It works by combining frame-level object detection, sentiment analysis, and timestamp mapping to locate inappropriate segments. The paper explains the system's overall design, how it was developed, and the results observed from testing this real-time content moderation tool.

3. Literature Survey

Over the last ten years, AI has improved a lot in the area where content moderation is concerned. Back in 2012, Krizhevsky and his team created AlexNet, a deep learning model that did extremely well on the ImageNet challenge. It became one of the main reasons computer vision improved so much later on.A few years later, in 2018, Chandrasekharan and his colleagues looked at Reddit's community bans from 2015. They found that when users themselves helped with moderation, hate speech went down and the platform became more inclusive. This showed how community led moderation can actually work if there are clear rules.

In 2019, Roberts talked about how hard this job can be for human moderators in her book Behind the Screen. She mentioned that many of them suffer from stress and trauma after seeing disturbing content every day. Around that time, Cambridge Consultants, along with Ofcom, tried to resolve this by using AI filters that could detect images automatically, like detecting if an image had a weapon. This helped lower the amount of harmful material that AI systems had to review.

In 2020, Gorwa and Guilbeault discussed several major issues with AI moderation. They mentioned that many of these systems are still unreliable and sometimes unfair, as they depend heavily on basic keyword filtering. Gillespie also pointed out that these tools often remove too much content, especially posts from minority groups, simply because they fail to understand context. Even after all these improvements, users were still unhappy with how moderation worked.

In 2021, Common Sense Media reported that nearly 78% of parents were worried about their children seeing inappropriate videos on YouTube. The following year, a study by Pew Research showed that 64% of users continued to come across harmful videos despite YouTube's moderation efforts.

In 2022, Shah Fadia Anwar and her team introduced the idea of AI as a Service (AIaaS) for content moderation. They used semantic and sentiment analysis to detect immoral or harmful content on cloud platforms. This model made it easier to manage large amounts of user data.

In 2023, Arora and Nakov worked on detecting cyberbullying and hate speech using natural language processing. They showed that looking at both text and context helps AI make better decisions. That same year, other studies said that current systems still find it difficult to handle unpredictable content like livestreams or deepfakes.

By 2024, an AI company called Stream reported an 85% accuracy rate for finding harmful content in live streams using multimodal models. Their work also reduced the need for human moderators by about 40%. Another report from that year studied how AI detects fake or AI-generated media, but it still found problems with accuracy and robustness.

Most recently, in 2025, Cloudinary AI developed a deepfake detection system that used metadata and lip-sync checking to spot manipulated videos. It reduced how fast fake videos spread—from about 14 hours to just 22 minutes. Another study that year showed that most people still prefer some level of community moderation because AI sometimes makes mistakes.

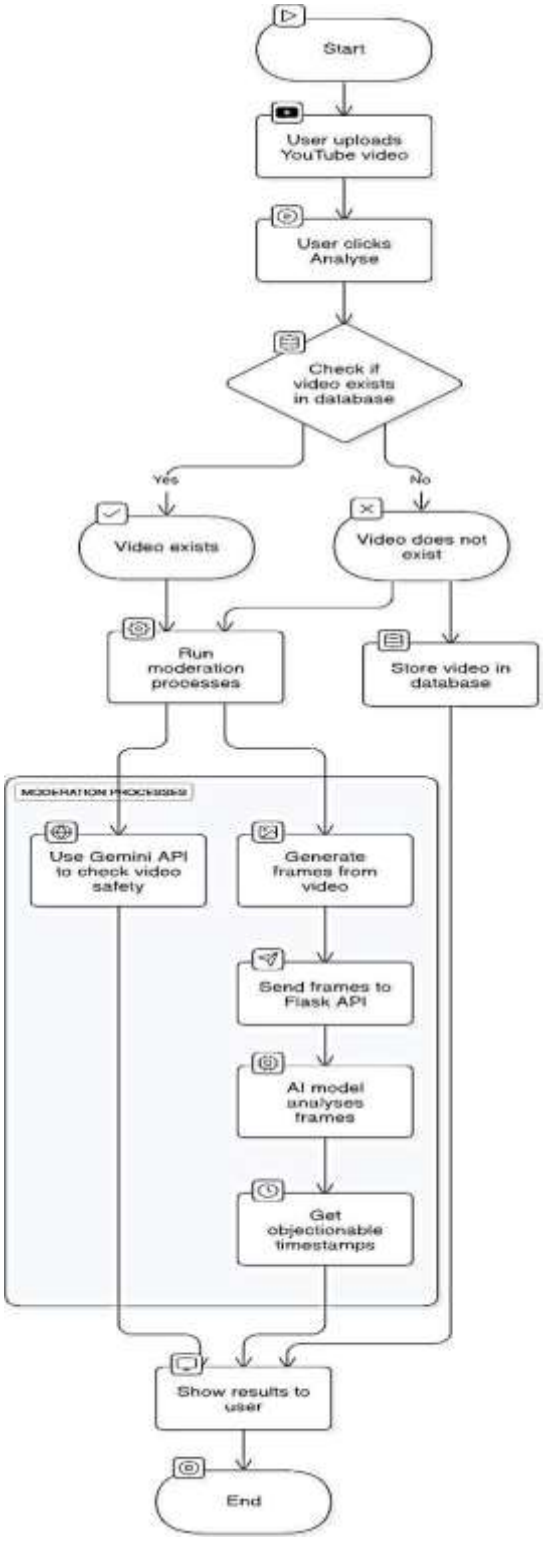
4. Research Gaps

While significant progress has been made in both textual and visual content moderation, several critical gaps remain. Most existing solutions focus on platform-level moderation rather than empowering users directly. Many systems are still reliant on keyword-based filters, which lack contextual nuance, and few combine both NLP and Computer Vision in a real-time, frame-level analysis. Additionally, multimodal moderation models are often not open-sourced, limiting community engagement and transparency. Our work addresses these shortcomings by introducing SafeStream.ai, a modular, user-centric system that performs timestamp-level moderation on YouTube videos using object detection, sentiment analysis, and natural language understanding. Unlike prior models, it provides real-time, user-initiated sanitization, ensuring safety while preserving viewing experience.

5. Methodology

The team began by designing a structured and scalable database schema to support smooth interactions between users, submitted YouTube links, and corresponding video analysis data. Each user account maintains essential credentials, while every submitted video entry stores metadata such as video title, duration, frame timestamps marking inappropriate content, and sanitized video links. The AI-generated results are persistently stored, enabling future access, audits, or reporting. This schema was crafted to ensure secure, efficient, and user-specific content management across the platform. The system architecture was carefully divided into three core components, each optimized for performance, modularity, and clarity of responsibility.

As illustrated in Figure 2, the architecture includes: Frontend – Developed using Next.js, the user interface enables users to seamlessly input YouTube URLs, track analysis progress, and retrieve sanitized video results. Backend – Built using Next.js API routes, the backend handles routing, video metadata processing, secure database interactions, and response generation. It also manages communication between services. AI Microservice – Created with help of Flask, this module is core for content moderation. It runs a PyTorch-based object detection model (MobileNetV3-Small) to evaluate extracted video frames for objectionable content. When someone puts in a video link, the backend first downloads the file and then breaks it into frames with OpenCV. After that, every frame goes to the AI part of the system, which checks for things like nudity, fights, or weapons. If the model thinks a frame is unsafe, the system notes the time so it knows where the bad parts are. Those time marks are later used to skip, blur, or mute that section when the video plays. To keep the process quick, I used asynchronous code with webhooks and multithreading. This lets the AI microservice tell the backend when it's done without stopping everything else from running. The results, mainly the time stamps and what type of content was found, are saved in the Neon Serverless Postgres database so they can be viewed or checked anytime. The whole setup works well and can change easily later on. More models, new filters, or even support for other video platforms can be added without much trouble. SafeStream.ai is meant to stay flexible and useful outside the lab too.



3. CONCLUSIONS

The online version of the volume will be available in LNCS Online. Members of institutes subscribing to the Lecture Notes in Computer Science series have access to all the pdfs of all the

online publications. Non-subscribers can only read as far as the abstracts. If they try to go beyond this point, they are automatically asked, whether they would like to order the pdf, and are given instructions as to how to do so.

ACKNOWLEDGEMENT

The heading should be treated as a 3rd level heading and should not be assigned a number.

REFERENCES

1. Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121
2. Bruce, K.B., Cardelli, L., Pierce, B.C.: Comparing Object Encodings. In: Abadi, M., Ito, T. (eds.): Theoretical Aspects of Computer Software. Lecture Notes in Computer Science, Vol. 1281. Springer-Verlag, Berlin Heidelberg New York (1997) 415–438
3. van Leeuwen, J. (ed.): Computer Science Today. Recent Trends and Developments. Lecture Notes in Computer Science, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995)
4. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)