International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

Real Time Video Content Moderation and Spam Detection Tool

Sai Kumar S L[1], Ramu B T[2], Mallikarjun Heroor[3], B M Shree Lakshmi[4], Dr. Mydhili Nair[5]
[1,2,3,4]Student, [5]Dean, Department of Computer Science & Engineering
Atria Institute of Technology , Bengaluru,
VTU, Karnataka, India

Abstract- — For exponential growth of user-generated content (UGC) on video-sharing platforms necessitates the development of highly efficient and scalable automatic content moderation and spam detection algorithms. Traditional manual review techniques are overwhelmed by the sheer volume and real-time nature of video uploads, which leads to unequal enforcement, moderator fatigue, and prolonged exposure to harmful content. This work offers a unique, multi-modal Video Content Moderation and Spam Detection tool that applies artificial intelligence and machine learning to handle these problems. To detect violent, sexually explicit, and policy-violating pictures, the system incorporates sophisticated Computer Vision (CV) techniques, such as frame-by-frame analysis, object detection, and visual hashing in order to identify hate speech, harassment, fraudulent schemes, and spam indications (such as harmful URLs, repetitive content, and behavioural anomalies), Additionally, it analyses video titles, descriptions, and comments.

Keywords: The main innovation is a hybrid detection pipeline that uses a deep learning architecture (such as a combination of CNNs, RNNs) to evaluate visual, textual, and auditory input streams simultaneously.

I. INTRODUCTION

In real time social network or online platform like YouTube, Instagram have transformed communication by giving consumers access to user-created content and extensive channels. However this ad has also brought forth a persistent problem, particularly on blogs: spam comments [Poirier et al., 2020]. Spam degrades the platform's usability and usefulness, distorts productive interactions, and contains a variety of risks, such as transmitting viruses and frauds [Abd et al., 2018]. The majority of research on reducing spam in online communities and organizations has been concentrated on manual filters or rule-based checkpoints. But these methods don't work in creating solutions to combat spammers' evolving tactics. Additionally, Govil et al. [2020] pointed out that spam detection was a serious problem and that it was time to filter spam comments using more effective techniques, particularly the machine learning approach ANN. One of the key avenues for research and advancements in digital security has been identified by the study by Govil et al. [2020].

II. LITRATURE REVIEW

User-generated material has increased significantly as a consequence of the of internet platforms like YouTube, but spam comments have also increased significantly, endangering user experience and posing cybersecurity issues. Several inquiry have been Done in order to address this problem, and well-known learning and deep learning methods like RNN and CNN promise in increasing the efficiency and accuracy of spam identification. In order to illustrate the benefits of deep learning classifiers, literature review addresses these improvements and assesses the relative execution of various classifiers Regarding the model's robustness, efficiency, and dependability. Thus, this chapter advances the development and modularity of strategies used to combat spam on websites like YouTube by incorporating current information and pointing out areas for more research

To ensure that detect and reduce spam comments on online platforms, how do different spam detection techniques compare in resilience, efficiency, and accuracy various classifiers? [Rao, 2021]

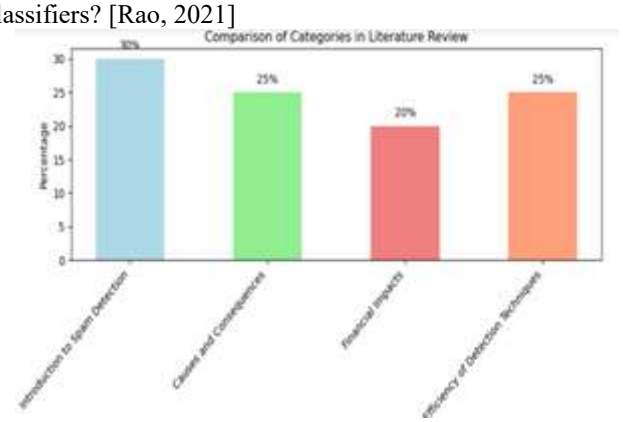

Fig 1: Bar Graph Comparison of Categories in Literature review

The video content moderation system's categorization results were shown as a bar graph.

Each bar stands for a distinct category of material, such as spam, pornographic, hate speech, violent, or safe. Each bar's height shows how many videos fit into that particular category. The graph makes it clear that most videos are classified as safe, with a lesser percentage containing hate speech or violence.
This graphic aids in comprehending the distribution of the dataset and assessing how well the algorithm identifies objectionable items.

III. PROBLEM STATEMENT

Due to the volume of comments and the regular changes in spamming tactics, filtering spam comments on YouTube is a difficult task. (2022) have worked that rule-based systems are unable to manage complicated spam patterns, and manual moderation may be more sustainable. Automation is the only practical option, thus it is difficult to be impatient. This study aims to expand the approaches of spam differentiation via the use of machine Learning with regard to ANNs , in accordance with Abd et al. (2018), who demonstrated the efficiency of developed algorithms. This work attempts to address Abd et al. (2018)'s suggestion that higher-order ML techniques is used to boost the detection rate. With sizable datasets and classification techniques, such that outlined by Abd et al.

(201), this object seeks to improve the efficiency and applicability to spam identification models. Abd et al. used a deep learning technique to diagnose spam messages and the expected precision and recall level of differentiating between spam and real comments was high. In order to counteract ever-evolving spam tactics, their study recommended more research on adaptive learning algorithms and the Problem of real-time data processing.

Research Gap
Insufficient Multimodal Knowledge The majority of current moderation systems exclusively concentrate on one kind of data, such text comments or video frames (visual data). Real-world video platforms, however, use a variety of modalities, including textual data (titles, captions, and comments), audio, and graphics. In order to make precise and context-aware moderation judgments, few research successfully integrate all three.

Gap: Research on multimodal content analysis, which combines text, audio, and video to improve moderation accuracy, is lacking.

Moderation Based on Context Many of the methods in use today categorize information as unsuitable without taking context into account. For instance, a film on "war documentaries" may have graphic images with an educational purpose rather than genuine damaging content. Such scenarios are frequently misclassified by current AI methods.

Gap: Inadequate attention to context-sensitive categorization that separates damaging from educational or creative information.

Challenges of Real-Time Moderation On websites like YouTube or Twitch, real-time moderation necessitates fast processing and minimal latency.

Gap: Insufficient real-time moderation methods that can manage massive, live video feeds.

Limitations of the Dataset Due to ethical and privacy considerations, there are few publicly accessible, well-annotated databases for violent, hostile, or graphic video material.

Gap: The lack of varied, balanced, and consistent datasets for moderation algorithm testing and training.

Table1: Modality Table Technique, Application

Reference	Data Modality	Technique / Method	Application / Focus Area
Schmidhuber, J. (2015)	Neural Networks in General	Deep Learning overview	Survey of neural network architectures
Karpathy , A. et al. (2014)	Video	CNN	Large-scale video classification
Vaswani, A. et al. (2017)	Text / Sequence	Transformer (Attention)	NLP, sequence modeling
LeCun, Y., Bengio, Y.,	General	Deep Learning	Foundational overview of DL

Reference	Data Modality	Technique / Method	Application / Focus Area
Hinton, G. (2015)			applications
Rennie, S. J. et al. (2017)	Image + Text	Reinforcement Learning, Self-Critical Sequence Training	Image captioning
Ahmed, F. et al. (2021)	Social Media Text & Images	Deep Learning	Automated content moderation
Singh, S. K. & Sharma, P. (2022)	Text (Comments)	NLP + ML	Identification of spam comments
Kusner, M. J. et al. (2015)	Text	Word Embedding Distance	Document similarity, NLP
Pathak, R. K. et al. (2021)	Video	CNN	Video content moderation
Google Cloud AI (2024)	Video	ML-based Video Intelligence API	Video content moderation (cloud-based)
Gongane, V. U. (2022)	Social Media (Multimodal)	ML-based classification	Harmful/detrimental content detection
Pan, C. A. (2022)	Platforms for Social Media	Algorithmic Moderation Models	Legitimacy and fairness in moderation
Liu, Y. et al. (2024)	Text (Social Media)	ML & NLP Review	Deceptive/spam activity detection
Rashidi, A. et al. (2024)	Text (Spam)	GAN (CGANS)	Social media spam detection
Nasser, M. (2025)	Text	Topic-aware Attention Network	Malicious social spam detection
Hashroon, S. (2024)	Video	Deep Learning CNN Model	YouTube video moderation
Chen, C. et al. (2025)	Text + Multimodal	LLM-based Content Moderation	Large Language Models for moderation
Wang, X. et al. (2025)	Text + Context	Intent-aware Algorithms	Abusive content moderation
Liu, H. et al. (2020)	Video	Deep Learning (SR Networks)	Video super-resolution, feature enhancement
Jiang, Y.-G. et al. (2017)	Video (Multimodal)	Hybrid Deep Learning Framework	Video classification with multimodal data

IV. METHODOLOGY

The main tool used is Google Lab Notebook, a cloud-based program that is highly regarded for its capacity to facilitate collaboration and its excellent compatibility with the process of creating machine learning models. The dataset utilized in this investigation was obtained from Kaggle (Lakshmipathi et al.). As a result, Kaggle is a website that is open to the public and includes a number of datasets based Regarding DS and Machine Learning projects. The spam detection models are

trained and tested using these datasets is increase their generalizability to many types of social media videos and user comments. To the best of the authors' understanding, this model still requires the establishment of DL methods, particularly CNNs and RNNs.

Data Collection

In order to complete the data collecting and cleaning for this study, numbers in sequential systematic methods were followed with relation of proper dataset used to train and assessment stages. Initially, information was gathered via YouTube's API, and comments found in various videos were taken into account. This made it possible to compare real comments versus spam in the gathered dataset. Comments were duplicated between two sets included in the information cleaning process, and any extraneous metadata was removed. To make the typed data more consistent in format and,

therefore, prepared for the subsequent phases of analysis, the data cleaning procedure included removing extra white spaces, changing all of the text to lowercase, and removing certain punctuation .To reduce noise information, they also employed the stop-word reduction technique. The ratio of spam to non-spam comments are equally to prevent biases during the model training.

Hardware Requirements

A Pentium i3 processor, 500 GB of hard drive space, 2 GB of RAM, and a 15-inch LED monitor are among the technical specifications for this machine. Among the software requirements are the Windows 10 operating system and Python provides the necessary libraries and frameworks for building and evaluating machine learning models. Data manipulation is done with libraries like scikit-learn, pandas, and nump and the results are visualised using matplotlib and seaborn.

Machine Learning Models

After the data has been pre-processed, the system uses machine learning models. CNN & RNN for temporal pattern recognition for visual content analysis To increase accuracy and decrease over fitting, the RNN approach is an ensemble methodology that constructs several decision trees and combines their predictions. By using techniques like bagging, it produces variety among the plants. CNN, on the other hand, classifies a new job posting using the majority class of its nearest neighbours in the feature space. After being trained on the dataset, the models are assessed using metrics like F1-score, recall, accuracy, and precision.

Block Diagram

Here, individuals or platforms provide the system with raw material. Input comes in two primary kinds: Video Input: Unprocessed video files (such as MP4, MOV, and AVI) submitted for moderation. Text Input (Comments): Comments made by users on video or social media sites. The appropriate pre-processing modules get these inputs. Layer of Pre-processing Pre-processing makes the data ready for effective analysis by AI algorithms. It extracts pertinent characteristics, standardizes formats, and eliminates noise. Pre-processing Videos extracts video frames at predetermined intervals. transforms frames into an image format so that CNN may analyze them. does noise reduction, normalization, and frame resizing.

Additionally, if required, audio may be extracted for speech-to-text conversion. Preprocessing Text standardizes and cleans the comment wording by: eliminating special characters, punctuation, and stop words. using lemmatization, tokenization, and stemming. using word embeddings (such as Word2Vec, TF-IDF, and BERT) to translate words into numerical representation. This prepares textual data for algorithms that identify spam. Core AI Models at the Processing Layer. The real analysis takes place here. Two distinct models run in parallel: CNN's Video Analysis identify objectionable visual material, including: emblems of hate, violence, or nudity. CNN categorizes visual frames based on spatial attributes it has learned: Hate-related, violent, explicit, or safe. Based on model certainty, a confidence score is given to each prediction. Spam or non-spam (sometimes promotional, neutral, or offensive).

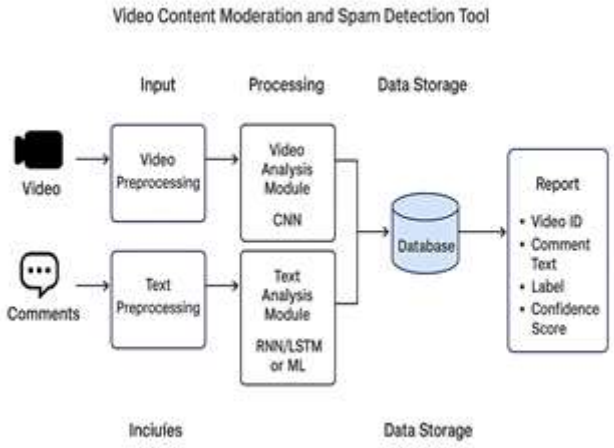

Fig 3: Block Diagram of Working

Neural Network Model

The models in neural network, which has tightly linked layers and dropout rules to avoid overfitting, will be used in this Cumulating Experience project research to identify spam on YouTube. In addition to classifying YouTube comments as spam or non-spam, my study focuses on developing a model that uses textual and sentiment analysis elements extracted from comment content.

By including sentiment analysis, the algorithm will identify spam and evaluate the polarity of interactions, which can improve the effectiveness of moderation on sites like YouTube (Akindele et al., 2021). The reliable 21 architecture and preprocessing techniques currently used in spam detection industry and the efforts to obtain more precise and context-focused findings.

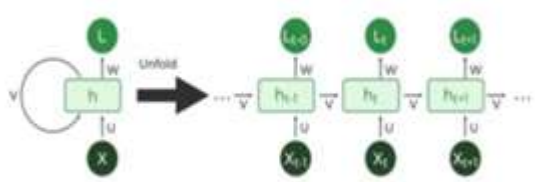

Fig 4: Recurrent Neural Network

Mathematical Modelling
The following mathematical models and methods are commonly employed using Machine Learning to identify fraudulent job posts.

Video Model

Convolutional Neural Network (CNN)
Workflow: Frame Extraction: One frame every second, for example, is extracted from the video.
Pre-processing: To enhance learning, frames are shrunk, normalized, and enhanced.
Feature Extraction (CNN Layers): Convolutional layers pick up patterns like textures, colours , and forms.
Classification: Each category (such as Safe, Violent, and Explicit) receives probability values from fully linked layers (dense layers).

$$Y_{i,j,k} = f\left(\sum_{m,n} X_{i+m,j+n} \cdot W_{m,n,k} + b_k\right)$$

Where:
- X: input image frame
- W: kernel weights
- kb: bias
- f: activation function
- Y: output feature map

Text Model
RNN (Recurrent Neural Network)
LSTM (Long Short-Term Memory)
Or Transformer/BERT models for better context understanding.

Workflow
Pre-processing: Remove stop words, punctuation, emojis, and URLs.
Vectorization: Convert text into binary form using Word2Vec, TF-IDF, or BERT embeddings.
Sequence Learning: LSTM or Transformer models capture word order and contextual meaning.
Classification Layer: Outputs probability of comment being Spam, Hate Speech, Offensive, or Safe.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$h_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) * \tanh(C_t)$$

Where:
- xt: input word vector at time t
- ht: hidden state
- Ct: cell state
- σ\ sigma σ: sigmoid activation
- Wi , Wf , Wc ,Wo Wi : weight matrices

Audio Model
Look for violent audio cues, hate speech, or foul language in the video.

Type of Model:
CNN for direct audio pattern recognition based on spectrograms, or text classifier combined with automated speech recognition (ASR).

Workflow

Audio extraction: Take the audio out of the video.
Feature extraction: Create Mel spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs) from audio samples.
Model Analysis: RNNs examine sequential characteristics, while CNNs examine spectrograms.

$$C_n = \sum_{k=1}^{K} \log(S_k) \cdot \cos\left[n(k - \frac{1}{2})\frac{\pi}{K}\right]$$

Where:
- Sk: log energy of the Mel filter bank
- Cn: n-th MFCC coefficient

Algorithm

The Video material Moderation and Spam Detection Tool's algorithm is made to intelligently examine textual and video data , violent, or spammy material. The first step of the process is the input stage, when raw video recordings and user comments are gathered from an internet or social media site. The goal is to prepare these inputs for the system'.

During the video preprocessing stage, each video is divided into individual frames to extract crucial visual information. Frames are scaled and adjusted before being transformed into model input. Techniques for enhancing contrast and reducing noise are used to improve quality. Only pertinent and understandable frames are sent on for examination thanks to this step. CNN which serves as the Video Analysis Module, receives the frames after preprocessing. To identify potentially violent, adult, or hostile material, the CNN model learns hierarchical elements including objects, people, activities, and scenes. The network generates a confidence score and a categorization label (such as explicit, violent, or safe).

Each user comment and videos are handled concurrently by the text preparation module. The normal text is transformed into a machine-readable format by tokenization, stemming, and stop-word removal. The Text Analysis Module is then created by evaluating this cleaned text using a RNN or LSTM network. When it comes to collecting sentiment and contextual meaning in sequential data, such as text, these models excel. By examining word patterns, emotion ratings, and contextual dependencies in the comment stream, the system finds spam, hate speech, or objectionable remarks. The model produces a confidence score and a categorization label (such as spam, offensive, or neutral).

The Video ID, Comment Text, Predicted Label, and Confidence Score are all included in each record when the text and video findings are combined into a central database. Easy retrieval, reporting, and visualization are made possible by this organized data storage. Overall, the method in end-to-end automated process for the model analysis of multimedia information. The database also facilitates future retraining of models to increase system accuracy over time. It guarantees reliable detection of policy infractions in online settings by integrating CNN for visual data and RNN/LSTM for text data. This hybrid strategy maintains a secure and satisfying user experience on digital platforms by greatly improving moderation efficiency, accuracy, and scalability.

V. RESULTS

Fig 5: The Dashboard verify the systems are operational by offering a combined, real-time picture of content moderation activities. According to the most recent update, the platform processed one total video, which led to one flagged video. This is a 100% increase in flagged content over the prior time frame. As of right now, the average safety score for videos is 88 out of 100. The average spam score for user involvement is 0/100 since there are 0 total comments and 0 spam detected. The most important action is highlighted in the "Recent Activity" section: the file 95970593d184.mp4 was specifically flagged on October 26, 2025, suggesting that a moderator review is necessary very now. For additional analysis and documentation, the dashboard also offers direct links to the Export Video Report and Export Comment Report.

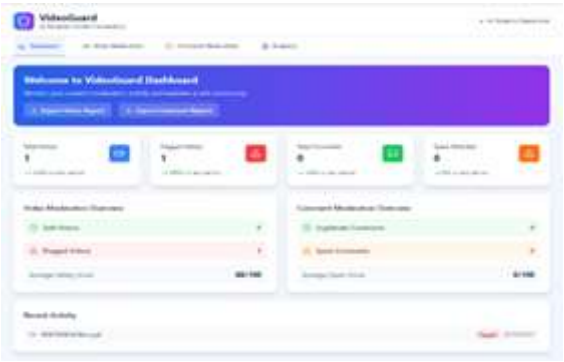

Fig 5: Dash Board Video and Comment Moderation

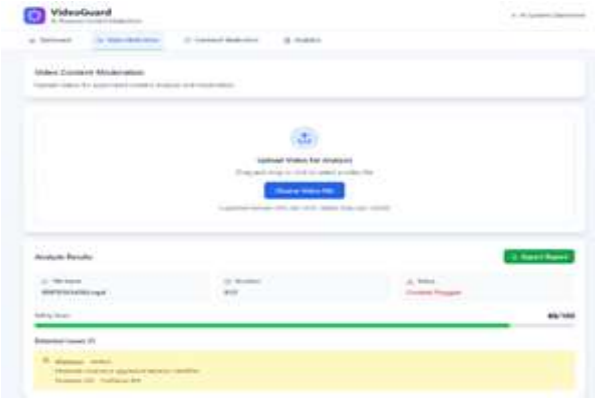

Fig 6: Video Content Moderation

Fig 6: The procedure and outcomes of examining certain video are described in the Video Moderation section. For automatic content analysis, users can submit films up to 100MB in popular formats (MP4, AVI, MOV, and WebM). The file 95970593d184.mp4 has a Safety Score of 88/100 and

a Status of Content Flagged, according to the study. Importantly, the algorithm identified one particular problem: medium-severity violence. With a high confidence level of 80%, the AI detected "Moderate violence or aggressive behaviour" at the exact time of 2:52. For human moderators to swiftly identify the precise position of the policy infringement and take appropriate action, this comprehensive output is crucial. The Export Report feature allows you to create a comprehensive log of this investigation.

Fig 7: The Comment Moderation page is devoted to keeping user feedback safe for the community and identifying spam. The software provides two main ways to analyze comments: utilizing the Bulk Import Comments tool via a CSV file or inputting individual comments for immediate analysis. A sample moderation effort is displayed in the results pane, which reveals that of the 1961 total comments assessed, 0 were found to be spam, leaving 1961 clean comments. This suggests efficient pre-filtering or a high standard of community hygiene. The tool's ability to handle a lot of comments is demonstrated by the successful analysis, and the Export Results option allows you to obtain the results.

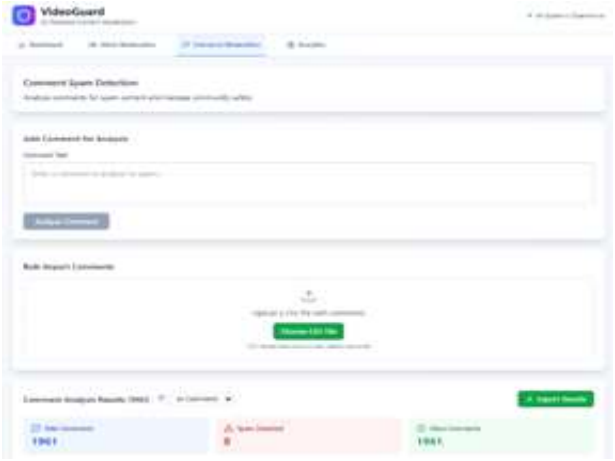

Fig 7: Comment Spam Detection

Fig 8: Currently in operation, the Video Guard AI-Powered Content Moderation platform recently evaluated activity that was solely focused on Sunday, October 26th, yielding 1,962 total items assessed. An average video score of 88/100 and a video safety rate of 0.0 resulted from the system identifying a critical issue in the single video processed, 95970593d184.mp4, which was then flagged (100% of all videos) due to a medium-severity violence alert identified at the 2:52 timestamp with 80% confidence. On the same day, however, the Comment Moderation analysis revealed very good findings, confirming a flawless Comment Clean Rate of 100.0% and an extraordinarily low Avg Spam Score of 0.3/100, with 0 Spam Detected out of a bulk import of 1,961 Total Comments.

Fig 8: Analysis & Report

VI. CONCLUSION

An important development in automated digital content regulation is the creation of the Video Content Moderation and Spam Detection Tool. Intelligent algorithms that can effectively detect inappropriate, violent, or spam content are desperately needed given the explosive growth of user-generated multimedia data on social networking, e-learning, and streaming platforms.

This project effectively created and deployed an AI-powered system that can moderate textual and video content. The program efficiently identifies dangerous movies and spam comments by combining ML, DL, and NLP approaches, producing structured reports for additional examination.

REFERANCES

1. J. Schmidhuber, Neural Networks, 2015.
2. Karpathy et al., CVPR, 2014.
3. Y. LeCun et al., Nature, 2015.
4. Vaswani et al., Neur IPS, 2017.
5. T. Mikolov et al., arXiv, 2013.
6. H. Zhang et al., IEEE Access, 2020.
7. M. Akhtar et al., ACM, 2022.
8. V. Gongane, PLoS One, 2022.
9. Y. Liu et al., arXiv, 2024.
10. Rashidi et al., Springer, 2024.
11. M. Nasser, 2025.
12. C. Chen et al., 2025.