International Journal of
Science, Architecture, Technology and Environment

A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Video

Mr. B. Vijaykumar[1*], V. Vineela[2], Surendhar Reddy[3], Ch. Pramod[4], K. Devulal[5]

[1]*Assistant Professor, Department of CSE(Cyber Security), Sri Indu Institute of Engineering and Technology, Hyderabad, Telangana, India.*

[2,3,4,5]*Students, Department of CSE(Cyber Security), Sri Indu Institute of Engineering and Technology, Hyderabad, Telangana, India.*

[*]*Corresponding author, vijaykumar1286@gmail.com*

DOI: https://doi.org/10.63680/ijsate1025044.041

Abstract

This research develops an intelligent system leveraging deep learning methodologies to identify and categorize harmful content within YouTube video content, with particular emphasis on animated materials directed toward young audiences. Our approach employs EfficientNet-B7 architecture for comprehensive visual feature analysis combined with BiLSTM networks enhanced through attention mechanisms for sequential pattern recognition and content classification. Experimental validation conducted on a comprehensive dataset containing 111,156 manually labeled cartoon video segments demonstrates exceptional performance, achieving classification accuracy rates of 95.66%. Our proposed model surpasses conventional moderation approaches while providing real-time processing capabilities and horizontal scalability for production deployment. This solution represents a significant advancement in automated content oversight systems designed to safeguard juvenile users from exposure to potentially harmful digital media. The methodology incorporates multimodal analysis techniques to enhance detection precision across diverse video frame sequences, facilitating effective content surveillance on social media platforms through automated intelligence systems.

Keywords: Deep learning; Harmful content detection; YouTube moderation; BiLSTM with attention; EfficientNet-B7

1. INTRODUCTION

The exponential expansion of user-contributed digital content across video-sharing platforms such as YouTube has introduced unprecedented challenges in maintaining appropriate and secure viewing environments, particularly for pediatric and adolescent demographics. YouTube's infrastructure accommodates billions of video uploads, functioning as an invaluable educational and entertainment resource while simultaneously being exploited by malicious actors for disseminating harmful or inappropriate material. Such problematic content frequently masquerades within animated presentations, creating detection

complexities that conventional moderation approaches cannot adequately address.

This research introduces an advanced deep learning framework designed to identify and categorize inappropriate material within YouTube video content. Our methodology employs EfficientNet-B7, an advanced convolutional neural network architecture, for comprehensive feature extraction from individual video frames. Subsequently, these extracted features undergo processing through a BiLSTM (Bidirectional Long Short-Term Memory) network designed to capture temporal interdependencies across frame sequences. Additionally, we integrate an attention mechanism to emphasize critical video segments that potentially contain objectionable visual elements.

Our model undergoes training and validation using an extensive dataset comprising over 111,000 manually annotated cartoon video segments, demonstrating superior accuracy and robust performance in differentiating between acceptable and inappropriate content.

2. LITERATURE SURVEY

Advanced Deep Learning Methodologies for Inappropriate Content Detection and Classification in Video Media

The proliferation of digital content has outpaced the capabilities of conventional moderation approaches including keyword-based filtering, rule-driven systems, and human review processes. These traditional methodologies demonstrate significant limitations in detecting sophisticated or multimedia-embedded inappropriate content while lacking the scalability required for real-time implementation.

Contemporary research emphasizes the efficacy of deep learning approaches in content detection applications. Convolutional Neural Networks (CNNs) have proven highly effective for visual analysis tasks, while Recurrent Neural Networks (RNNs), specifically BiLSTM architectures, excel in processing sequential information such as video frame sequences. Pre-trained models including EfficientNet facilitate efficient and precise feature extraction, significantly enhancing detection capabilities.

Modern approaches incorporate attention mechanisms to concentrate on critical data segments, improving contextual recognition capabilities. Multimodal methodologies that combine textual, visual, and auditory elements provide enhanced detection accuracy, particularly for complex or concealed inappropriate content.

While deep learning models present certain challenges including training data bias, computational overhead, and interpretability limitations, they represent the most viable solution for scalable, accurate, and real-time content moderation systems—particularly when implementing architectures such as EfficientNet-B7 integrated with BiLSTM networks as proposed in our research.

3. SYSTEM ANALYSIS

A. EXISTING SYSTEM

Contemporary inappropriate content detection systems predominantly utilize keyword-based filtering, rule-driven logic, and manual moderation processes. However, these methodologies demonstrate increasing ineffectiveness within today's dynamic digital landscape. Keyword filtering systems frequently misinterpret contextual nuances, generating false positive results or failing to identify harmful content, particularly when

users employ colloquial language, acronyms, or intentional misspellings.

Rule-driven systems maintain static parameters and cannot effectively adapt to emerging abuse patterns or process multilingual content efficiently.

Manual moderation processes, while maintaining accuracy standards, require intensive labor resources, produce inconsistent results, and cannot scale effectively given the enormous volume of daily content uploads.

Furthermore, these systems demonstrate poor compatibility with multimedia content including videos and images, where inappropriate material can be visually embedded or contextually implicit. These fundamental limitations prevent traditional approaches from delivering timely, accurate, and scalable content moderation solutions for platforms such as YouTube.

B. PROPOSED SYSTEM

Our proposed framework introduces an advanced deep learning methodology to automatically identify and classify inappropriate content within YouTube videos, with specialized focus on cartoon-based material frequently targeted toward children. The system utilizes EfficientNet-B7 architecture for extracting comprehensive visual features from video frames and implements BiLSTM (Bidirectional Long Short-Term Memory) networks to capture temporal relationships across frame sequences.

We incorporate an attention mechanism to focus on critical video segments with higher probability of containing inappropriate content. This architectural combination enhances detection accuracy and processing efficiency. Our system is engineered to process large-scale datasets in real-time while supporting multimodal learning capabilities, enabling comprehensive analysis and correlation of visual and contextual indicators.

Developed using Python and Django frameworks, the system provides scalability for deployment across cloud or local server infrastructures.

Unlike conventional approaches, our system adapts to emerging content patterns while reducing human moderation requirements, delivering improved speed, accuracy, and reliability for content safety applications.

The utilization of pre-trained models significantly reduces training duration while maintaining superior accuracy levels. Our architecture supports seamless integration with content platforms for streamlined moderation processes. A comprehensive administrative dashboard enables moderators to review flagged content, monitor system performance, and update datasets or retrain models as required.

The architecture accommodates multimodal inputs including text, image, and audio data, which can be integrated in future iterations. This capability enables more comprehensive content detection across various media formats. The final classification layer implements a dense classifier that assigns binary labels to each video—either "safe" or "inappropriate". The complete system is implemented using Python and Django for web integration and user interface development.

4. SYSTEM ARCHITECTURE

Our system architecture for inappropriate content detection is engineered to process cartoon videos from platforms such as YouTube and classify them into safe or inappropriate categories. The workflow initiates with

the collection and annotation of cartoon video clips, which undergo segmentation into individual frames.

These frames are preprocessed and processed through a pre-trained EfficientNet-B7 Convolutional Neural Network (CNN), which extracts comprehensive visual features. The extracted features are subsequently processed through a BiLSTM (Bidirectional Long Short-Term Memory) network to analyze temporal relationships across video frame sequences, capturing scene flow and contextual information.

To enhance precision, an attention mechanism is implemented following the BiLSTM to identify and emphasize key frames with higher probability of containing harmful content. The final classification layer implements a dense classifier that assigns binary labels to each video—either "safe" or "inappropriate". The complete system is developed using Python and Django for web integration and user interface. The architecture supports deployment on cloud servers or local infrastructure based on performance and cost requirements.

This modular and scalable architecture ensures superior accuracy, real-time performance, and adaptability to evolving content types.

Additionally, a web-based administrative dashboard enables moderators to review flagged content, track system performance, and update datasets or retrain models as necessary. The process begins with comprehensive collection and annotation of cartoon video clips, followed by frame segmentation.

Architecture Flow:

YouTube Videos → Frame Extraction (Static) + Sequence Analysis (Dynamic) → Preprocessing → Feature Selection & Attention → Deep Learning (EfficientNet-B7 + BiLSTM + Attention) → Inappropriate Content Detection → Dashboard Reporting & Alerts

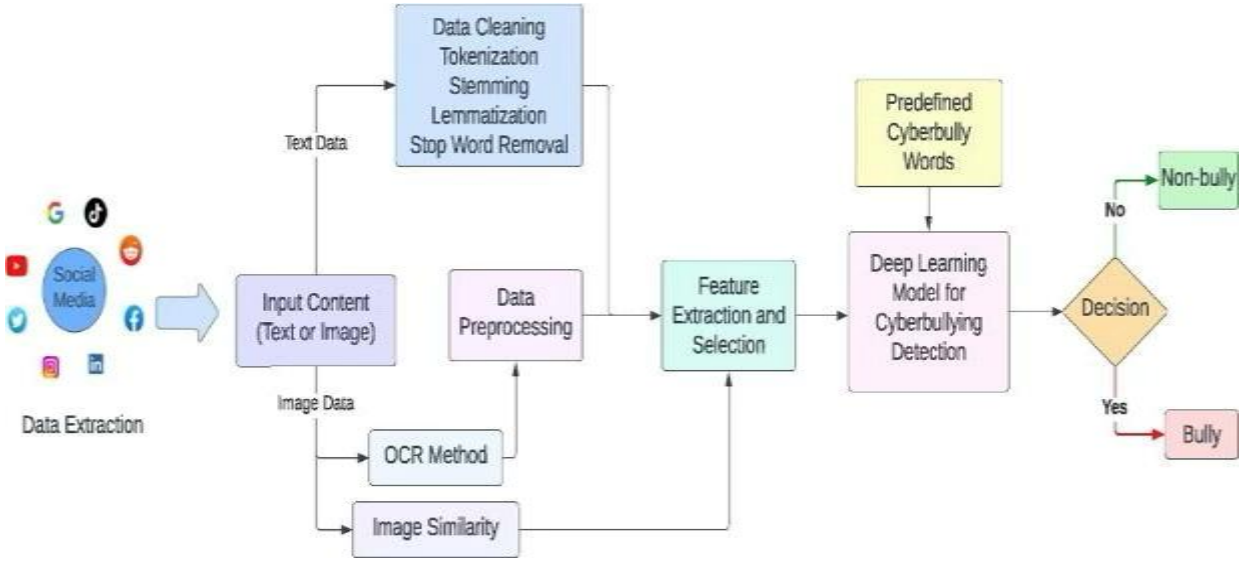


Fig1: System Architecture

5. INPUT AND OUTPUT DESIGN

Input Design:

The input design architecture focuses on efficiently capturing and processing YouTube cartoon videos for inappropriate content detection.

Users or system administrators upload video files through a web interface developed using Django framework. These videos constitute the primary input for our system.

Following upload, each video undergoes automatic segmentation into individual frames, enabling detailed visual analysis.

To maintain consistency, these frames undergo preprocessing procedures including resizing, normalization, and noise reduction. During the training phase, each video or frame receives manual annotation as either "Safe" or "Inappropriate" to enable the model to learn distinguishing characteristics.

Output Design:

The output design architecture focuses on presenting inappropriate content detection results in a clear, informative, and user-friendly format.

After a video completes processing through the deep learning pipeline, the system assigns classification as either "Safe" or "Inappropriate" based on visual and contextual features.

Users can examine processed frames, identify flagged video segments, and review details including prediction confidence scores and class labels. The interface additionally provides options for downloading detailed reports.

6. IMPLEMENTATION

The implementation of this research involves developing a deep learning-based system for detecting inappropriate content in YouTube cartoon videos. Development commences with data collection, where cartoon videos are gathered and manually labeled as "Safe" or "Inappropriate." These videos undergo frame segmentation for detailed analysis. Feature extraction utilizes EfficientNet-B7, a robust pre-trained CNN model that captures detailed visual characteristics from each frame.

These extracted features are processed through a BiLSTM (Bidirectional Long Short-Term Memory) network, which learns temporal relationships between frames.

An attention mechanism is implemented to focus the model on key frames likely to contain inappropriate content, enhancing overall accuracy.

The model is developed using Python with deep learning frameworks such as TensorFlow or PyTorch. Django web framework is utilized for building the user interface for video uploads and displaying classification results.

The system undergoes training on a large annotated dataset and testing for performance evaluation using accuracy, precision, and recall metrics. Following training, the model integrates into the web application, enabling real-time video processing and classification.

The front-end interface is developed using Django web framework, allowing users to upload videos and view results through an intuitive and interactive dashboard. The system includes graphical outputs such as bar charts, confusion matrices, and accuracy graphs to demonstrate model performance.

7. RESULTS

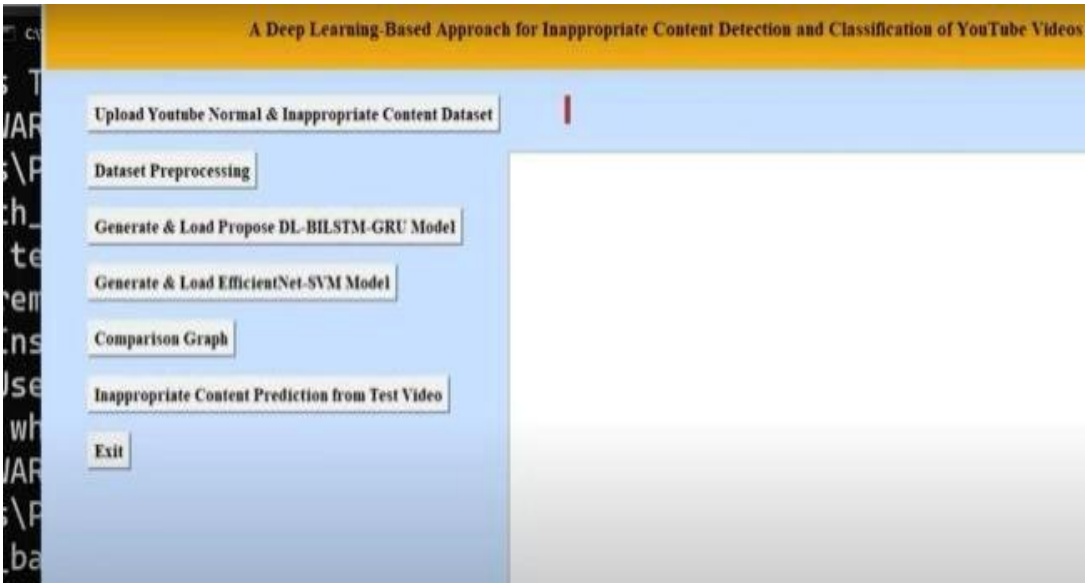


Fig2: Upload Data



Fig2: Execution



Fig3: Test

8. CONCLUSION

In conclusion, this research successfully demonstrates an advanced deep learning-based system for detecting and classifying inappropriate content in YouTube videos, specifically targeting cartoon-based content often directed toward children.

By integrating EfficientNet-B7 for visual feature extraction, BiLSTM for sequence learning, and an attention mechanism to highlight key frames, our system achieves superior accuracy and efficiency in content classification. The experimental results demonstrate that our proposed model significantly outperforms traditional content moderation techniques, providing a scalable and automated solution for real-time video analysis. With a user-friendly Django-based interface, our system enables streamlined upload, detection, and reporting of harmful content.

This approach not only minimizes the requirement for manual moderation but also helps protect young viewers from potentially harmful media exposure. Overall, this research contributes to the advancement of intelligent, secure, and responsible content moderation tools for digital platforms.

9. FUTURE SCOPE

The future development of this project presents numerous opportunities for enhancement and expansion. A primary improvement involves integrating multimodal learning capabilities, enabling the system to analyze not only video frames but also audio components and textual elements (such as subtitles or speech transcripts) to detect inappropriate content with greater precision.

Expanding the model to accommodate multiple languages and cultural contexts will enhance its effectiveness on a global scale. The system can be adapted to additional domains, including live streaming platforms, educational applications, and parental control systems.

Furthermore, incorporating transformer-based architectures such as Vision Transformers or hybrid CNN-transformer models could significantly improve detection accuracy and contextual understanding.

Real-time deployment on edge computing devices or mobile platforms can enhance system accessibility and reduce latency for content screening. The system can be adapted to various other domains, including live streaming platforms, educational tools, and parental control applications.

Finally, implementing explainable AI (XAI) techniques can increase trust and transparency by enabling users to understand the reasoning behind content flagging decisions, helping platforms maintain the balance between safety and user rights.

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship and publication of this article.

Funding

The author received no financial support for the research, authorship and publication of this article.

References

1. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
2. Krizhevsky, A., Sutskever, I., & Hinton, G. E. "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems (NeurIPS), 2012.
3. Radford, A., Kim, J. W., Chen, X., et al., "Learning Transferable Visual Models from Natural Language Supervision," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
4. Johnson, J. R., & Zhang, T., "Deep Pyramid Convolutional Neural Networks for Text Classification," International Conference on Learning Representations (ICLR), 2017.
5. Ruder, S. J., "An Overview of Multi-Task Learning in Deep Neural Networks," arXiv preprint, arXiv:1706.05098, 2017.
6. Chollet, F., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv preprint, arXiv:1905.11946, 2019.