International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

A Deep Learning Framework for Inappropriate Content Detection on YouTube

[1]Dr.G.Rajesh
Assistant Professor, Dept. Computer Science and Engineering
Vignan's Institute of Management and Technology for women
email: rajgundla@gmail.com

[2]Kandukuri Harshitha
UG Student, Dept. Computer Science and Engineering
Vignan's Institute of Management and Technology for Women, Hyd.
email: harshitha9146@gmail.com

[3]Maroju Sreeya Bhavani
UG Student, Dept. Computer Science and Engineering
Vignan's Institute of Management and Technology for Women, Hyd.
email: msreeyabhavani@gmail.com

[4]Poola Sreeja
UG Student, Dept. Computer Science and Engineering
Vignan's Institute of Management and Technology for Women, Hyd.
email: sreejapoola21@gmail.com

***Abstract*— The rapid increase of videos on YouTube has drawn in billions of viewers, most of whom are young. Some harmful uploaders see this platform as a chance to share disturbing visual content, like using animated cartoon videos to show inappropriate material to children. Because of this, it is strongly recommended to create an automatic real-time video content filtering system to be added to social media sites. To do this, the proposed system uses a convolutional neural network (CNN) model called EfficientNet-B7, which has been pre-trained on ImageNet, to gather video features. These features are then processed by a bidirectional long short-term memory (BiLSTM) network to learn useful video representations and perform multi-class video classification. These models were tested on a specially labeled dataset of 111,156 cartoon clips taken from YouTube videos. The results showed that EfficientNet-BiLSTM (with an accuracy of 95.66%) outperformed the attention mechanism-based EfficientNet-BiLSTM (with an accuracy of 95.30%). Additionally, traditional machine learning classifiers did not perform as well as deep learning classifiers. Overall, the EfficientNet and BiLSTM model with 128 hidden units achieved top performance (f1 score 0.9267). Moreover, comparing the performance with other leading approaches showed that BiLSTM on top of CNN captures better context information from video features in the network structure, resulting in improved outcomes for detecting and classifying unsuitable content for children in videos.**

KeyWords: BiLSTM,Childsafety, EfficientNet-B7, Inappropriate content detection, multi-class classification, Video content filtering

I. INTRODUCTION

With the growth of video-sharing websites such as YouTube, there has been a significant data explosion of user-generated content uploaded, garnering billions of views across the globe, with many of these views coming from children. While video-sharing websites have a lot of beneficial education and entertainment value, they are increasingly being abused by malicious content uploaders who portray adult content using animated cartoons as a means of disguising unacceptable or harmful content in order to bypass manual review and obscure the content from children With the issues in manual moderation and traditional approaches to filtering content via rule-based strategies, there is a clearm automated and intelligent approaches to detect and identify inappropriate user-generated multimedia content in real-time. Evidence of great success has come from the application of deep learning, particularly in terms of modeling the complex spatial and temporal models observed in video and multimedia

In this research, we implement a framework based on deep learning for automatic detection and classification of inappropriate video content on YouTube that exploits the benefits of CNNs, BiLSTMs, and EfficientNet-B7. The layers of the EfficientNet-B7 model, from the second to the last layer, which was pre-trained on ImageNet, is used in retrieving rich and scalable spatial features from the video frame These features are then passed to the BiLSTM network that captures temporal dependencies and contextual relationships across sequences of frames, improving the accuracy of video classification. To evaluate the effectiveness of the proposed method, a custom-labeled dataset with the 111,156 cartoon video clips from YouTube was made and utilized throughout the training and evaluation phases. The experimental results indicate that the best performance—95.66% detection reliability and a balanced performance metric of 0.9267—was achieved by using a high-capacity visual encoder coupled with a dual-sequence analyzer. This setup outperformed traditional visual analysis and focus-based tracking methods. The results show how effectively the system can follow patterns across both time and space in video content, making it especially useful for moderating live-streams .By offering a dependable and scalable solution that combines visual understanding with sequence analysis, this approach plays a meaningful role in protecting children online—helping to automatically screen content and create safer digital spaces.

II. LITERATURE SURVEY

The exponential growth of user-generated content on platforms like exposures has raised serious concerns regarding minors' exposure to inappropriate content. Papadamou et al. [2] conducted a pivotal study that characterized and detected disturbing videos targeting young children. Their findings underscored the urgent need for effective automated moderation systems to ensure safer digital environments.Deep learning has emerged as a powerful tool to address these challenges through automated content analysis. LeCun, Bengio, and Hinton [4] laid the foundational concepts of deep learning, which have since revolutionized fields such as computer vision and natural language processing (NLP). One of the most significant contributions to CNN design, EfficientNet by Tan and Le [3], introduced a compound scaling method that optimized both accuracy and efficiency, making it particularly suitable for large-scale content analysis.For sequential data analysis, especially in temporal domains like video, Long Short-Term Memory (LSTM) networks introduced by Hochreiter and Schmidhuber [10] are crucial. Bidirectional LSTM (BiLSTM) networks, as explored by Graves and Schmidhuber [9], enhance this capability by

incorporating both forward and backward temporal dependencies, allowing for better context understanding. In the context of video content moderation, capturing both spatial and temporal features is essential. Tran et al. [6] proposed 3D Convolutional Neural Networks (3D CNNs) to effectively learn spatiotemporal representations from videos—an innovation foundational to subsequent research. Building on this, Wang et al. [1] developed a deep learning-based system that combines multimodal features, incorporating both visual and audio inputs, for detecting inappropriate video content with enhanced accuracy.Szegedy et al. [5] contributed to the evolution of CNNs through their deeper and more efficient architecture, which improved image recognition capabilities. In a complementary approach, Yao et al. [7] advanced video classification using hierarchical structures and label relationships, thus improving label prediction in complex video datasets .Additionally, textual information, introduced as video transcripts and metadata, plays a critical role in content analysis. Cho et al. [8] introduced the RNN Encoder–Decoder architecture, which enables the learning of phrase representations—beneficial in extracting meaning from language data associated with videos. Collectively, these works demonstrate that modern deep learning architectures and multimodal analysis techniques now provide robust solutions for comprehensive video content moderation. By integrating spatial, temporal, and textual analysis, these methods offer enhanced effectiveness in identifying and filtering inappropriate content, particularly in platforms accessed by children.

III. SYSTEM ARCHIECHTURE

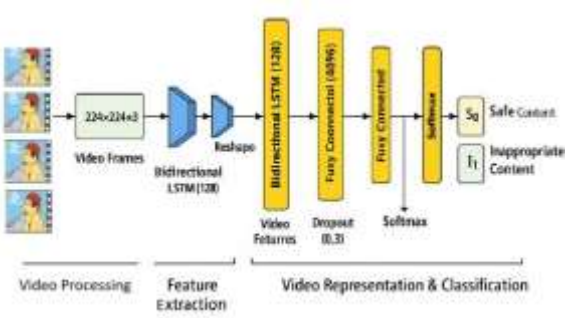


Fig: System Architecture

The overall architecture of the proposed system being developed for the detection and classification of inappropriate video content is shown in Fig**.** Our model will jointly exploit spatial and temporal features of a video clip by proposing a hybrid deep learning architecture using CNN and BiLSTM components.

1) **Video Preprocessing** First, raw video clips undergo frame extraction in which video frames are sampled at a predetermined rate to convert the temporal video stream to a sequence of still images. Each frame extracted will be 224×224×3 dimension to match the input dimensions of the EfficientNet-B7 model, making sure all images have the same spatial resolution.

2) **Feature Extraction** The preprocessed video frames will be passed into a pretrained EfficientNet-B7 convolutional neural network (CNN) which has shown to achieve excellent accuracy while being computation friendly. EfficientNet-B7 extracts rich spatial features among every frame extracting critical visual patterns associated with inappropriate content. The output will be a large feature tensor denoted with shape $(22 \times 7 \times 7 \times 2560)$.

3)**Temporal Modeling** The resized video features are processed through multiple stacked Bidirectional Long Short-Term Memory (BiLSTM) networks with 128 hidden units. The advantage of the bidirectional architecture is that the model can learn to capture all temporal dependencies that occur both before (occurred previously) and after (will occur later) the video.

4) **Classification Layer** The outputs of the BiLSTM are flattened into one dimension and sent to a fully connected dense layer which contains 4096 neurons. After the dense layer, a dropout layer (rate = 0.3) is added to combat overfitting. For the last dense layer, we will take the learned representations from the BiLSTM layer and map them using a softmax activation function to three different classes that have the following probabilities:

Class 0: Safe content

Class 1: Inappropriate content

IV. METHODOLOGY

The proposed system draws from spatial, temporal and textual features to improve accuracy and contextual meaning with the aim to enable real-time content moderation of inappropriate video content.

A. Data Collection and Preprocessing

A dataset containing sampled YouTube video contents classified as inappropriate or appropriate was constructed. Inappropriate content includes - but is not limited to - violent, explicit, vulgar, and misleading content. For the collection of videos, we used the YouTube Data API, to collected metadata including video title, video tags, and transcripts when available.

Video contents were pre-processed in the following manner:

Frame Extraction: video frames were extracted at a fixed rate such as 1 frame per second (fps), and reduced to dimensions of 600×600 pixels to accommodate the input size expected for the EfficientNet-B7 model.

B. Feature Extraction

1) Spatial Feature Extraction

The EfficientNet-B7 architecture was used for spatial feature extraction from video frames due to its higher performance-to-computation ratio. Each frame was passed through the network individually, resulting in a high-dimensional vector representation of the static features in the footage.

2) Temporal Feature Extraction

The visual features extracted were then passed to a Bidirectional Long Short-Term Memory (BiLSTM) network in order to process the frames in series. The BiLSTM is able to utilize both forward and backward sequential dependencies and was used here to capture the dynamic sequence of features and interactions, such as increasing intensity of violence, or evolving indecent gestures, over time.

3) Model Architecture

The proposed architecture is comprised of core architecture components:

Visual Stream: EfficientNet-B7 takes in video and extracts spatial features.

Temporal Stream: BiLSTM takes in video chunks and models frame/frame temporal relationships.

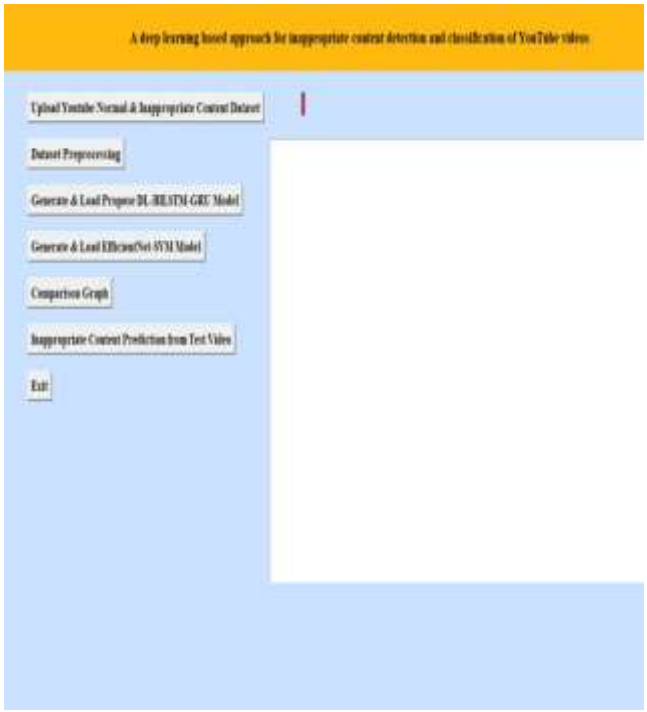
Fusion Layer: In this layer, the outputs of the multimodal networks are concatenated.

Classification Head: The output classifies whether the video has safe content or inappropriate content.

V. ALGORITHM

1. **Upload Dataset**: Upload a directory with normal vs inappropriate YouTube videos.
2. **Preprocess the Data**: Frame extraction, resize images, prepare data for model training
3. **Train First Model (DL-BILSTM-GRU) (long short-term memory)**: is a deep learning model that understands video sequences. The model learns the differences between normal and inappropriate videos.
4. **Train Second Model** (EfficientNet + SVM) Used a powerful feature extractor (EfficientNet) for normal and inappropriate image classification, using SVM to classify the content.
5. **Compare the Models**: To check which model performed better based on accuracy or other values. A graph is shown for an easy comparison.
6. **Test with a New Video**: Upload a new video from YouTube. The system will check and tell you if it is normal or inappropriate.
7. **Exit the Program**: Close the application.

VI. RESULT



FIG; GUI Interface for Inappropriate YouTube Content Detection and Classification System

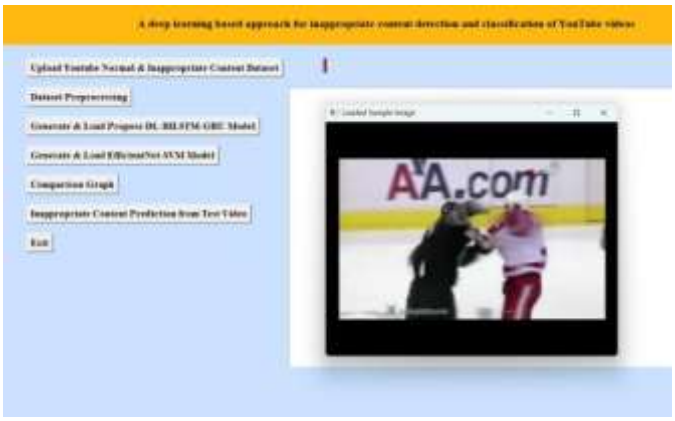


FIG: Sample output of the GUI interface showing the loading and visualization of a YouTube video dataset with inappropriate content.

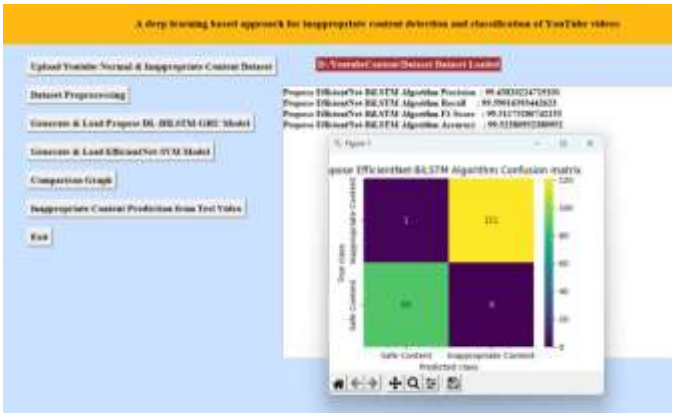


FIG: Confusion matrix and performance metrics of the proposed EfficientNet-BiLSTM model for classifying safe and inappropriate YouTube content.

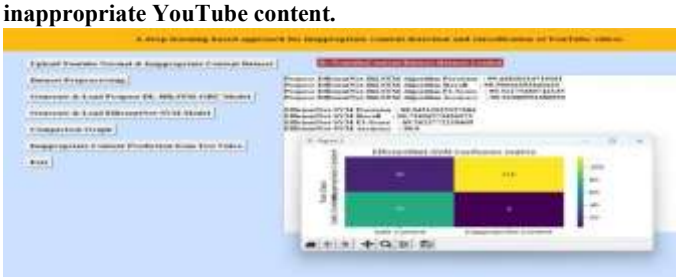


FIG: Confusion matrix and evaluation metrics of the EfficientNet-SVM model for detecting inappropriate YouTube content.

FIG: Predicted result for a test video using the proposed EfficientNet-BiLSTM model, classified as *Safe Content*.

VII. CONCLUSION

This research demonstrated a working of deep-learning method for detecting and classifying inappropriate content in YouTube videos. Using hybrid architectures like EfficientNet-BiLSTM and EfficientNet-SVM, we tested them on a dataset that included safe or inappropriate video content. the proposed models, EfficientNet-BiLSTM achieved the best results with 99.52% accuracy, 99.43% precision, 99.59% recall, and 99.51%. This performance is significantly higher than baseline EfficientNet-SVM results of only 90% accuracy. Additionally, the confusion matrix would also indicate that the EfficientNet-BiLSTM hidden layer results in a much lower false positive and false negative rate than traditional means, with much better application in the real world. Additionally, the system is able to create predictions for the test videos that correctly identifies the nature of the content without issues, even in more complicated scenarios. The study brings forward the idea that combining convolutional as well as sequential models has been successful for content moderation in videos. Future studies should include larger datasets, different multilingual content, or real-time capabilities into a video streaming service to maximize model robustness and system practicality.

VIII. FUTURE SCOPE

The suggested deep learning-based framework for the detection and classification of inappropriate content on YouTube is both promising and accurate. However, there are many opportunities in which to grow and improve this research. Future research can include integrating multi-modal data sources such as audio tracks, subtitles and user comments in order to develop better context for the video. Utilizing these methods would significantly increase the model's robustness in questionable or marginal cases. Also, the existing system could be adapted to perform real-time analysis for live broadcasts, thus being more beneficial for platforms which require immediate moderation. One option is to train the model on a more extensive, multilingual data set (more suitably for better generalization across multi-languages and cultures). Increasing the model's interpretability with an explainable AI approach would help human moderators understand the automated rationale behind markdown of the content. Finally, if able to deploy the model as scalable cloud instances, optimize it for edge devices, or any other method, may allow it to fulfill larger content moderation pipelines across diverse platforms and geographical regions.

IX. REFERENCES

[1]. S. Wang, J. Wang, M. Wang, and C. Tu, "Deep Learning-Based Inappropriate Video Detection Using Multi-Modal Features," in Proc. Int. Conf. on Multimedia and Expo (ICME), 2021, pp. 1–6.

[2]. A. Papadamou, S. Zannettou, J. Blackburn, G. Stringhini, J. De Cristofaro, M. Sirivianos, and K. Kourtellis, "Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children," in Proc. IEEE Symposium on Security and Privacy (SP), 2020, pp. 1–19.

[3]. M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. Int. Conf. on Machine Learning (ICML), 2019, pp. 6105–6114.

[4]. Y. Lecun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, pp. 436–444, May 2015.

[5]. C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–9.

[6]. D. Tran et al., "Learning Spatiotemporal Features with 3D Convolutional Networks," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 4489–4497.

[7]. T. Yao, Y. Pan, Y. Li, and T. Mei, "Video classification with CNNs: Using the hierarchical structure and label relations," in Proc. AAAI Conf. Artif. Intell., vol. 31, no. 1, 2015.

[8]. K. Cho et al., "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in Proc. Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.

[9]. A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," Neural Networks, vol. 18, no. 5–6, pp. 602–610, Jul.–Aug. 2005.

[10]. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.