DEEP LEARNING FOR DETECTING AND CLASSIFYING INAPPROPRIATE CONTENT IN YOUTUBE VIDEOS

[1]JANARDHAN KOMAROLU, Assistant Professor, Dept. of CSE
[2]SRIKANTH, MCA Student, Dept. of MCA
Rajeev Gandhi Memorial College of Engineering and Technology
Nandyal, 518501, Andhra Pradesh, India.

ABSTRACT

The exponential growth of videos on YouTube has attracted billions of viewers among which the majority belongs to a young demographic. Malicious uploaders also find this platform as an opportunity to spread upsetting visual content, such as using animated cartoon videos to share inappropriate content with children. Therefore, an automatic real-time video content filtering mechanism is highly suggested to be integrated into social media platforms. In this study, a novel deep learning-based architecture is proposed for the detection and classification of inappropriate content in videos. For this, the proposed framework employs an

ImageNet pre-trained convolutional neural network (CNN) model known as EfficientNet-B7 to extract video descriptors, which are then fed to bidirectional long short-term memory (BiLSTM) network to learn effective video representations and perform multiclass video classification. An attention mechanism is also integrated after BiLSTM to apply attention probability distribution in the network. These models are evaluated on a manually annotated dataset of 111,156 cartoon clips collected from YouTube videos. Experimental results demonstrated that EfficientNet-BiLSTM (accuracy D 95.66%) performs better than attention mechanism based EfficientNet-BiLSTM (accuracy D 95.30%) framework. Secondly, the traditional machine learning classifiers perform relatively poor than deep learning classifiers. Overall, the architecture of EfficientNet and BiLSTM with 128 hidden units yielded state-of-the-art performance (f1 score D 0.9267). Furthermore, the performance comparison against existing state-of-the-art approaches verified that BiLSTM on top of CNN captures better contextual information of video descriptors in network architecture, and hence achieved better results in child inappropriate video content detection and classification.

1. INTRODUCTION

The creation and consumption of videos on social media platforms have grown drastically over the past few years. Among the social media sites, YouTube predominates as a video sharing platform with plethora of videos from diverse categories. According to YouTube statistics [1], the global user base of YouTube is over 2 billion registered users and more than 500 hours of video content is uploaded every minute. Consequently, billions of hours of videos are available where users of all age groups can explore generic as well as personalized content [2]. Considering such a large-scale The associate editor coordinating the review of this manuscript and approving it for publication was Aasia Khanum . crowdsourced database, it is extremely challenging to monitor and regulate the uploaded content as per platform guidelines. This creates opportunities for malicious users to indulge in spamming activities by misleading the audiences with falsely advertised content (i.e., video, audio or text). The most disruptive behavior by malicious users is to expose the young audiences to disturbing content, particularly when it is fabricated as safe for them. Children today spend most of their time on the Internet and the YouTube platform for them has distinctly established itself as an alternative to traditional screen media (e.g., television) [3], [4]. The YouTube press release [5] also confirmed the high popularity of this social media site among younger audiences compared to other age groups, and the reason for this high level of approval is due to fewer restrictions [6].

Unlike television, children can be presented with any type of content on the Internet due to lack of regulations. Exposing children to disturbing content is considered as one among other internet safety threats (like cyberbullying, cyber predators, hate etc.) [7]. Bushman and Huesmann [8] confirmed that frequent exposure to disturbing video content may have a short-term or long-term impact on children's behavior, emotions and cognition. Many reports [9]–[12] identified the trend of distributing inappropriate content in children's videos. This trend got people's attention when mainstream media reported about the Elsagate controversy [13], [14], where such video material was found on YouTube featuring famous childhood cartoon characters (i.e., Disney characters, superheroes, etc.) portrayed in disturbing scenes; for instance, performing mild violence, stealing, drinking alcohol and involving in nudity or sexual activities. In an attempt to provide a safe online platform, laws like the children's online privacy protection act (COPPA) imposes certain requirements on websites to adopt safety mechanisms for children under the age of 13. YouTube has also included a ''safety mode'' option to filter out unsafe content. Apart from that, YouTube developed the YouTube Kids application to allow parental control over videos that are approved as safe for a certain age group of children [15]. Regardless of YouTube's efforts in controlling the unsafe content phenomena, disturbing videos still appear [16]–[19] even in YouTube Kids [20] due to difficulty in identifying such content. An explanation for this may be that the rate at which videos are uploaded every minute makes YouTube vulnerable to unwanted content. Besides, the decision-making algorithms of YouTube rely heavily on the metadata of video (i.e., video title, video description, view count, rating, tags, comments, and community flags).

Hence, filtering videos based on the metadata and community flagging is not sufficient to assure the safety of children [21]. Many cases exist on YouTube where safe video titles and thumbnails are used for disturbing content to trick children and their parents. The sparse inclusion of child inappropriate content in videos is another common technique followed by malicious uploaders. Fig. 1 displays an example among such cases where video title and video clips are safe for children (as shown in Fig. 1(a)) but included inappropriate scenes in this video (as shown in Fig. 1(b) and Fig. 1(c)). The concerning thing about this example, including many similar cases, is that these videos have millions of views with more likes than dislikes, and have been available for years. Many other cases (as shown in Fig. 1(d)) also identified where videos or the YouTube channel is not popular, yet contains child unsafe content especially in the form of animated cartoons. It is evident from examples that this problem persists irrespective of channel or video popularity. Furthermore, YouTube has disabled the dislike feature of videos which resulted in viewers being incapable of getting the indirect video content feedback from statistics. Since the YouTube metadata can be easily manipulated, it is suggested to better use video features for detection of inappropriate content than metadata features associated with videos [22].

2. EXISTINGSYSTEM

Rea *et al.* [37] proposed a periodicity-based audio feature extraction method which was later combined with visual features for illicit content detection in videos.

The machine learning algorithms are usually employed as classifiers Liu *et al.* [38] classified the periodicity-based audio and visual segmentation features through support vector machine (SVM) algorithm with Gaussian radial basis function (RBF) kernel. Later on, they extended the framework [39] by applying the energy envelope (EE) and bag-of-words (BoW)-based audio representations and visual features.

Ulges *et al.* [23] used MPEG motion vectors and Mel-frequency cepstral coefficient (MFCC) audio features with skin color and visual words. Each feature representation
is processed through an individual SVM classifier and combined in a weighted sum of late fusion. Ochoa *et al.* [40] performed binary video genre classification for adult content detection by processing the spatiotemporal features with two types of SVM algorithms: sequential minimal optimization (SMO) and LibSVM.

Jung *et al.* [41] worked with the one dimensional signal of spatiotemporal motion trajectory and skin color. Tang *et al.* [42] proposed a pornography detection system_PornProbe, based on a hierarchical latent Dirichlet allocation (LDA) and SVM algorithm. This system combined an unsupervised clustering in LDA and supervised learning in SVM, and achieved high efficiency than a single SVM classifier. Lee *et al.* [43] presented a multilevel hierarchical framework by taking the multiple features of different temporal domains. Lopes *et al.* [44] worked with the bag-of-visual features (BoVF) for obscenity detection.

Kaushal *et al.* [21] performed supervised learning to identify the child unsafe content and content uploaders by feeding the machine learning classifiers (i.e., random forest, K-nearest neighbor, and decision tree) with video-level, user-level and comment-level metadata of YouTube Reddy *et al.* [45] handled the explicit content problem of videos through text classification of YouTube comments. They applied bigram collocation and fed the features to the naïve Bayes classifier for final classification.

Disadvantages

- An existing system doesn't ANALYSIS OF PRE-TRAINED CNN MODEL VARIANTS.
- An existing system doesn't ANALYSIS OF EFFICIENT-NET FEATURES WITH DIFFERENT CLASSIFIER VARIANTS.

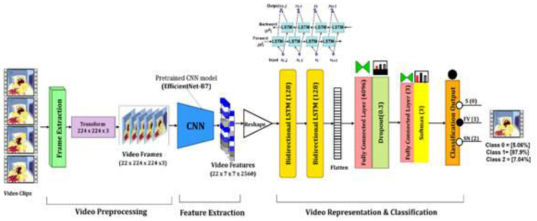
3. PROPOSED SYSTEM

1. The system proposes a novel CNN (EfficientNet-B7) and BiLSTM-based deep learning framework for inappropriate video content detection and classification.
2. The system presents a manually annotated ground truth video dataset of 1860 minutes (111,561 seconds) of cartoon videos for young children (under the age of 13). All videos are collected from YouTube using famous cartoon names as search keywords. Each video clip is annotated for either safe or unsafe class. For the unsafe category, fantasy violence and sexual-nudity explicit content are monitored in videos. We also intend to make this dataset publicly available for the research community.
3. The system evaluates the performance of our proposed CNN-BiLSTM framework. Our multiclass video classifier achieved the validation accuracy of 95.66%. Several other state-of-the-art machine learning and deep learning architectures are also evaluated and compared for the task of inappropriate video content detection.

Advantages

- The most frequent applications of image/video classification employed the convolutional neural networks.
- The EfficientNet model is a convolutional neural network model and scaling method that uniformly scales network depth, width and resolution through compound co efficient.

4. SYSTEM ARCHITECTURE



5. MODULES

To implement this project we have designed following modules

1) **Upload YouTube Normal & Inappropriate Content Dataset:** using this module we will upload YouTube dataset images to application
2) Dataset Preprocessing: using this module we will read all images and then resize all images to equal size and then normalize image pixel values and then shuffle the dataset
3) **Run Propose DL-BILSTM-GRU Algorithm:** using this module we will split dataset into train and test

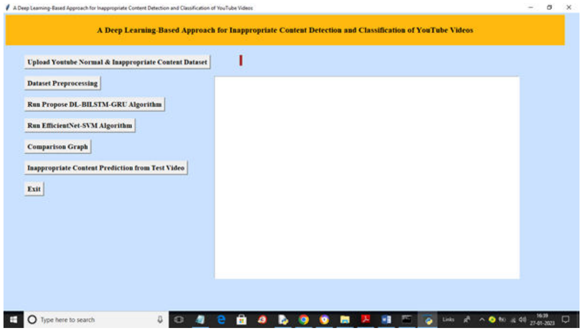
and then input 80% training data to Pre-Trained CNN (EfficientNetB7) algorithm to extract digital content from images and then those features will get retrained with BI-LSTM algorithm to train a model. Trained model will be applied on 20% test data to calculate prediction accuracy

4) **Run EfficientNet-SVM Algorithm:** EfficientNetB7 features will get retrained with existing SVM algorithm and then calculate prediction accuracy
5) **Comparison Graph:** using this module we will plot accuracy comparison graph between propose EfficientNetB7-BILSTM and EfficientNetB7-SVM.
6) **Inappropriate Content Prediction from Test Video:** using this module we will upload any YouTube and if video contains fighting or violence then application will predict as 'Inappropriate Content' otherwise will predict SAFE content.

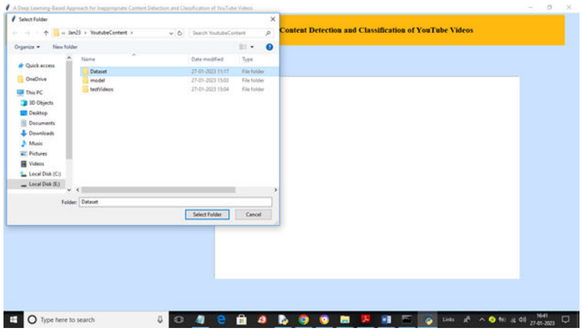
6. SCREEN SHOTS

In above screen we have two folders and juts go inside any folder to view training images

To run project double click on 'run.bat' file to get below screen


A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
Upload Youtube Normal & Inappropriate Content Dataset
Dataset Preprocessing
Run Propose DL-BILSTM-GRU Algorithm
Run EfficientNet-SVM Algorithm
Comparison Graph
Inappropriate Content Prediction from Test Video
Exit
Type here to search

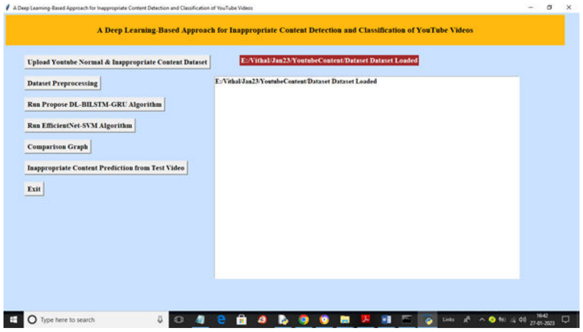
In above screen click on 'Upload YouTube Normal & Inappropriate Content Dataset' button to upload dataset and get below output



Name	Date modified	Type
Dataset	27-01-2023 11:17	File folder
model	27-01-2023 15:03	File folder
testVideos	27-01-2023 15:04	File folder

Folder: Dataset

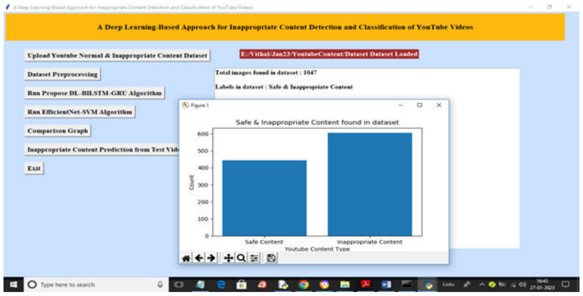
In above screen selecting and uploading entire 'Dataset' folder and then click on 'Select Folder' button to load dataset and get below output


A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
Upload Youtube Normal & Inappropriate Content Dataset
Dataset Preprocessing
Run Propose DL-BILSTM-GRU Algorithm
Run EfficientNet-SVM Algorithm
Comparison Graph
Inappropriate Content Prediction from Test Video
Exit
Type here to search

E:/Vithal/Jan23/YoutubeContent/Dataset Dataset Loaded

E:/Vithal/Jan23/YoutubeContent/Dataset Dataset Loaded

In above screen dataset loaded and now click on 'Dataset Preprocessing' button to read all images and then processes those images for training and get below output


A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
Upload Youtube Normal & Inappropriate Content Dataset
Dataset Preprocessing
Run Propose DL-BILSTM-GRU Algorithm
Run EfficientNet-SVM Algorithm
Comparison Graph
Inappropriate Content Prediction from Test Vid
Exit
Figure 1
Safe & Inappropriate Content found in dataset
600
500
400
300
200
100
0
Count
Safe Content
Inappropriate Content
Youtube Content Type
Type here to search

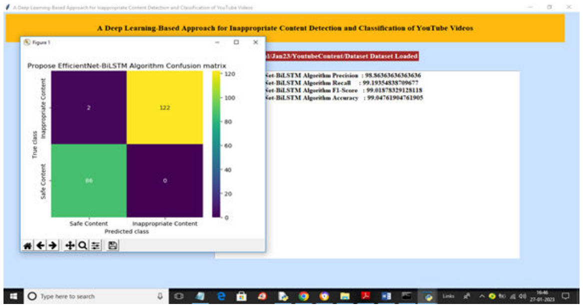
E:/Vithal/Jan23/YoutubeContent/Dataset Dataset Loaded

Total images found in dataset : 1047

Labels in dataset : Safe & Inappropriate Content

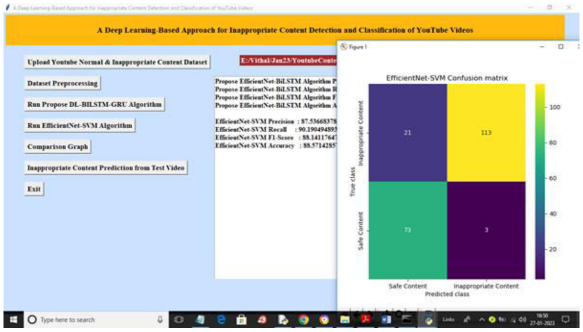
In above screen we can see dataset contains 1047 images and then in graph x-axis represents type of images such as 'Safe and Inappropriate' and y-axis represents count of those images. Now dataset processing completed and now click on 'Run Propose

DL-BILSTM-GRU Algorithm' button to train propose algorithm and get below output


A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
Figure 1
Propose EfficientNet-BiLSTM Algorithm Confusion matrix
True class
Inappropriate Content
Safe Content
Predicted class
2
122
86
0
120
100
80
60
40
20
0

```
l/Jan23/YoutubeContent/Dataset Dataset Loaded
Net-BiLSTM Algorithm Precision  : 98.86363636363636
Net-BiLSTM Algorithm Recall     : 99.19354838709677
Net-BiLSTM Algorithm F1-Score   : 99.01878329128118
Net-BiLSTM Algorithm Accuracy   : 99.04761904761905
```

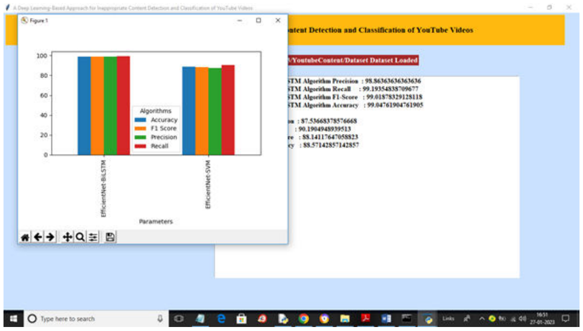
In above screen with propose EfficientNetB7-BI-LSTM we got 99.04% accuracy and in confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels and green and yellow boxes contains correct prediction count and blue boxes contains incorrect prediction count which is 2 only. Now close above graph and then click on 'Run EfficientNet-SVM Algorithm' button to get below output


A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
Upload Youtube Normal & Inappropriate Content Dataset
Dataset Preprocessing
Run Propose DL-BILSTM-GRU Algorithm
Run EfficientNet-SVM Algorithm
Comparison Graph
Inappropriate Content Prediction from Test Video
Exit
Figure 1
EfficientNet-SVM Confusion matrix
True class
Inappropriate Content
Safe Content
Predicted class
21
113
73
3
100
80
60
40
20

```
E:/Vithal/Jan23/YoutubeConte
Propose EfficientNet-BiLSTM Algorithm P
Propose EfficientNet-BiLSTM Algorithm R
Propose EfficientNet-BiLSTM Algorithm F
Propose EfficientNet-BiLSTM Algorithm A

EfficientNet-SVM Precision  : 87.53668378
EfficientNet-SVM Recall     : 90.19049489
EfficientNet-SVM F1-Score   : 88.14117647
EfficientNet-SVM Accuracy   : 88.57142857
```

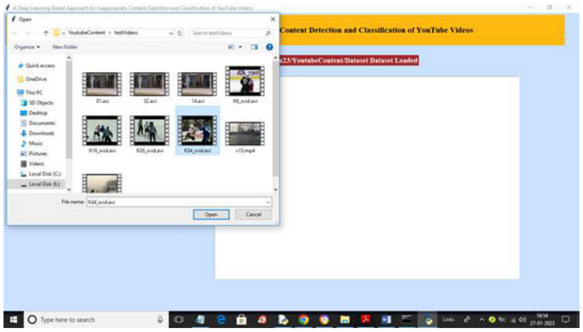
In above screen with EfficientNetB7-SVM we got 88% accuracy and in confusion matrix graph we can see in blue boxes that SVM predicted total 24 incorrect prediction so its accuracy is less. Now close above graph and then click on 'Comparison Graph' button to get below output



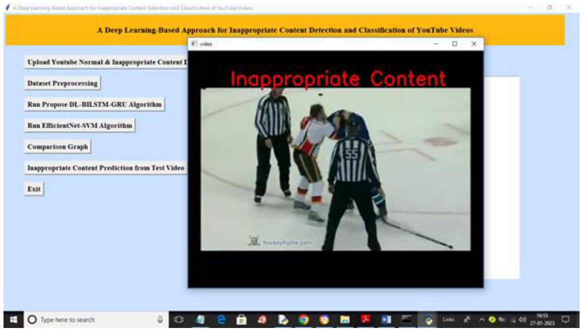
```
/YoutubeContent/Dataset Dataset Loaded
STM Algorithm Precision  : 98.86363636363636
STM Algorithm Recall     : 99.19354838709677
STM Algorithm F1-Score   : 99.01878329128118
STM Algorithm Accuracy   : 99.04761904761905

on  : 87.53668378576668
    : 90.19049489939513
re  : 88.14117647058823
cy  : 88.57142857142857
```

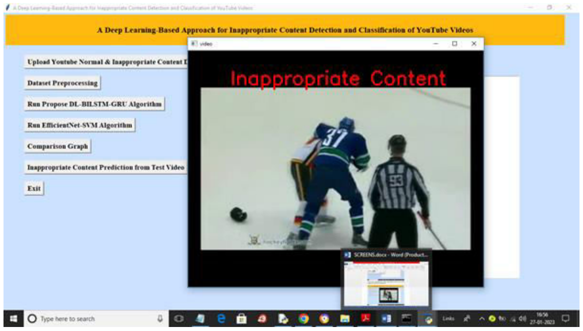
In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars. In both algorithms propose EfficientNetB7-BI-LSTM got high accuracy. Now close above graph and then click on 'Inappropriate Content Prediction from Test Video' button to upload test video and classify it as Safe or inappropriate.



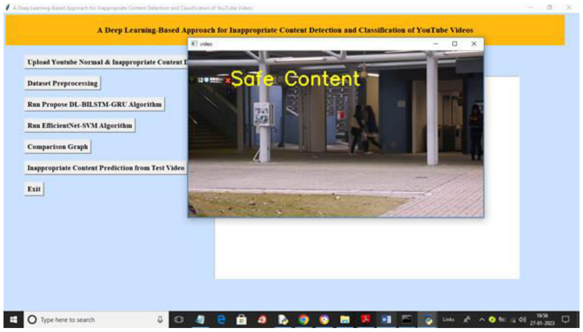
In above screen selecting and uploading video and then click on "Open' button to play video and perform classification


A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
video
Upload Youtube Normal & Inappropriate Content D
Dataset Preprocessing
Run Propose DL-BILSTM-GRU Algorithm
Run EfficientNet-SVM Algorithm
Comparison Graph
Inappropriate Content Prediction from Test Video
Exit
Inappropriate Content

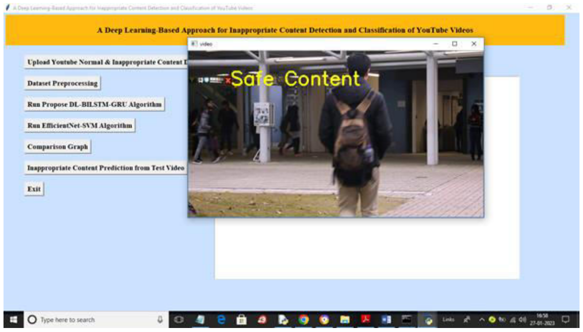
In above screen propose algorithm evaluating playing video and then detecting and classifying it as 'Inappropriate Content'


A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
video
Upload Youtube Normal & Inappropriate Content D
Dataset Preprocessing
Run Propose DL-BILSTM-GRU Algorithm
Run EfficientNet-SVM Algorithm
Comparison Graph
Inappropriate Content Prediction from Test Video
Exit
Inappropriate Content
SCREENS.docx - Word (Product...
Type here to search
Links
16:56
27-01-2023

In above video also we got classification output


A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
video
Upload Youtube Normal & Inappropriate Content D
Dataset Preprocessing
Run Propose DL-BILSTM-GRU Algorithm
Run EfficientNet-SVM Algorithm
Comparison Graph
Inappropriate Content Prediction from Test Video
Exit
Safe Content
Type here to search

In above we got result as Safe Content


A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos
video
Upload Youtube Normal & Inappropriate Content D
Dataset Preprocessing
Run Propose DL-BILSTM-GRU Algorithm
Run EfficientNet-SVM Algorithm
Comparison Graph
Inappropriate Content Prediction from Test Video
Exit
Safe Content
Type here to search

In above screen we got output as Safe Content as peoples are only moving in the video.

7. CONCLUSION

A novel deep learning-based framework is proposed for child inappropriate video content detection and classification. Transfer learning using efficientnet-b7 architecture is employed to extract the features of videos.

The extracted video features are processed through the bilstm network, where the model learns the effective video representations and performs multiclass video classification. All evaluation experiments are performed by using a manually annotated cartoon video dataset of 111,156 video clips collected from youtube. The evaluation results indicated that proposed framework of efficient net bilstm (with hidden units = 128) exhibits higher performance (accuracy = 95.66%) than other experimented models including efficient net-fc, efficient net-svm, efficient net-knn, efficient net-random forest, and efficient net-bilstm with attention mechanism-based models (with hidden units = 64, 128, 256, and 512). Moreover, the performance comparison with existing state-of-the-art models also demonstrated that our bilstm-based framework surpassed other existing models and methods by achieving the highest recall score of 92.22%. The advantages of the proposed deep learning-based children inappropriate video content detection system are as follows:

1) it works by considering the real-time conditions by processing the video with a speed of 22 fps using efficientnet-b7 and bilstm-based deep learning framework, which helps in filtering the live-captured videos.

2) it can assist any video sharing platform to either remove the video containing unsafe clips or blur/hide any portion with unsettling frames.

3) it may also help in the development of parental control solutions on the internet through plugins or browser extensions where child unsafe content can be filtered automatically.

Furthermore, our methodology to detect inappropriate children content from youtube is independent of youtube video metadata which can easily be altered by malicious uploaders to deceive the audiences. In the future, we intend to combine the temporal stream using optical flow frames with the spatial stream of the rgb frames to further improve the model performance by better understanding the global representations of videos. We also aim to increase the classification labels to target the different types of inappropriate children content of youtube videos.

REFERENCES

[1] L. Ceci. YouTube Usage Penetration in the United States 2020, by Age Group. Accessed: Nov. 1, 2021. [Online]. Available: https://www.statista.com/statistics/296227/us-youtube-reach-age-gender/

[2] P. Covington, J. Adams, and E. Sargin, ‘‘Deep neural networks for YouTube recommendations,’’ in Proc. 10th ACM Conf. Recommender Syst., Sep. 2016, pp. 191–198, doi: 10.1145/2959100.2959190.

[3] M. M. Neumann and C. Herodotou, ‘‘Evaluating YouTube videos for young children,’’ Educ. Inf. Technol., vol. 25, no. 5, pp. 4459–4475, Sep. 2020, doi: 10.1007/s10639-020-10183-7. [4] J. Marsh, L. Law, J. Lahmar, D. Yamada-Rice, B. Parry, and F. Scott, Social Media, Television and Children. Sheffield, U.K.: Univ. Sheffield, 2019. [Online]. Available: https://www.stac-study.org/downloads/ STAC_Full_Report.pdf

[5] L. Ceci. YouTube—Statistics & Facts. Accessed: Sep. 01, 2021. [Online]. Available: https://www.statista.com/topics/2019/youtube/ [6] M. M. Neumann and C. Herodotou, ‘‘Young children and YouTube: A global phenomenon,’’ Childhood Educ., vol. 96, no. 4, pp. 72–77, Jul. 2020, doi: 10.1080/00094056.2020.1796459.

[7] S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson, Risks and Safety on the Internet: The Perspective of European Children: Full Findings and Policy Implications From the EU Kids Online Survey of 9-16 Year Olds and Their Parents in 25 Countries. London, U.K.: EU Kids Online, 2011. [Online]. Available: http://eprints.lse.ac.U.K./id/eprint/33731

[8] B. J. Bushman and L. R. Huesmann, ‘‘Short-term and long-term effects of violent media on aggression in children and adults,’’ Arch. Pediatrics Adolescent Med., vol. 160, no. 4, pp. 348–352, 2006, doi: 10.1001/archpedi.160.4.348.

[9] S. Maheshwari. (2017). On YouTube Kids, Startling Videos Slip Past Filters. The New York Times. [Online]. Available: https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html

[10] C. Hou, X. Wu, and G. Wang, ‘‘End-to-end bloody video recognition by audio-visual feature fusion,’’ in Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV), 2018, pp. 501–510, doi: 10.1007/978-3-030-03398- 9_43.

[11] A. Ali and N. Senan, ‘‘Violence video classification performance using deep neural networks,’’ in Proc. Int. Conf. Soft Comput. Data Mining, 2018, pp. 225–233, doi: 10.1007/978-3-319-72550-5_22.

[12] H.-E. Lee, T. Ermakova, V. Ververis, and B. Fabian, ‘‘Detecting child sexual abuse material: A comprehensive survey,’’ Forensic Sci. Int., Digit. Invest., vol. 34, Sep. 2020, Art. no. 301022, doi: 10.1016/j.fsidi. 2020.301022.

[13] R. Brandom. (2017). Inside Elsagate, The Conspiracy Fueled War on Creepy YouTube Kids Videos. [Online]. Available: https://www.theverge. com/2017/12/8/16751206/elsagate-youtube-kids-creepy-conspiracytheory

[14] Reddit. What is ElsaGate? Accessed: Dec. 14, 2020. [Online]. Available: https://www.reddit.com/r/ElsaGate/comments/6o6baf/

[15] B. Burroughs, ‘‘YouTube kids: The app economy and mobile parenting,’’ Soc. media+ Soc., vol. 3, May 2017, Art. no. 2056305117707189, doi: 10.1177/2056305117707189.