A DEEP LEARNING-BASED APPROACH FOR INAPPROPRIATE CONTENT DETECTION AND CLASSIFICATION OF YOUTUBE VIDEOS

[1] Dr.K. Sailaja.
[1]Associate Professor & HOD
[1]Department of Computer Applications
[1]Chadalawada Ramanamma Engineering College (Autonomous),Tirupathi

[2] Chinthamakula Jayasree
[2] Student
[2] Department of Computer Applications
[2] Chadalawada Ramanamma Engineering College (Autonomous), Tirupathi

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 = 95.66%) performs better than attention mechanism-based EfficientNet-BiLSTM (accuracy = 95.30%) framework

KEYWORDS : Videos ,Feature extraction, Deep learning ,Support vector machines, Convolutional neural networks, Visualization, Classification algorithms

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 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 .Considering such a large-scale crowd sourced 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) The YouTube press release 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 .

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 cyber bullying, cyber predators, hate Bushman and Huesmann 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 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 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 Regardless of YouTube's efforts in controlling the unsafe content phenomena, disturbing videos still appear even in YouTube Kids 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 _ags). Hence, altering videos based on the metadata and community _agging is not sufficient to assure the safety of children 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 up loaders. 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 identi_ed 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

Prior techniques addressed the challenge of identifying disturbing content (i.e., violence, pornography, etc.) from videos by using traditional hand-crafted features on frame-level data. In recent years, the state-of-the-art performance of deep learning has motivated researchers to employ it in image and video processing. The most frequent applications of image/video classification employed the convolutional neural networks. Apart from that, the long-short term memory (LSTM), a special type of recurrent neural network (RNN) architecture, has proven to be an effective deep learning model in time-series data analysis Hence, this study targets the YouTube multiclass video classification problem by leveraging CNN (Ef_cientNet-B7) and LSTM to learn video effective representations for detection and classi_cation of inappropriate content. We targeted two types of objectionable content geared towards young viewers, one, which contains violence and the second, which includes sexual nudity connotations.

The main contributions of this study are threefold:

1. We propose a novel CNN (Ef_cientNet-B7) and BiLSTM-based deep learning framework for inappropriate video content detection and classification.

2. We present 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. We evaluate 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. To summarize, this work can assist any video sharing platform to either remove unsafe video or blur/hide any portion of video involving unsafe content. Secondly, it may also help in the development of parental control solutions on the web via plugins or browser extensions where children inappropriate content filters automatically. The upcoming sections of the article are outlined as follows: Section II covers the related work in this research area. The methodology of our proposed system is explained in Section III. The experimental setup of the proposed system is presented in Section IV. The results obtained from the experimental setup are analyzed and discussed in Section V, and finally, Section VI concludes the work and directs some future scope for improvements.

2. LITERATURE SURVEY

2.1 DIFFERENT AUTHORS DISCUSSION:

We propose a novel CNN (EffcientNet-B7) and BiLSTM-based deep learning framework for inappropriate video content detection and classification.

We present 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.

2.2 DOMAIN DESCRIPTION:

The system presents a manually annotated ground truth video dataset of 1860 minutes of cartoon videos for young children .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. PROBLEM STATEMENT

3.1 EXISTING SYSTEM

Rea 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 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 by applying the energy envelope (EE) and bag-of-words (BoW)-based audio representations and visual features.

Ulges 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 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. ith the one dimensional signal of spatiotemporal motion trajectory and skin color. Tang *et al.* 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 presented a multilevel hierarchical framework by taking the multiple features of different temporal domains. Lopes worked with the bag-of-visual features (BoVF) for obscenity detection.

Kaushal 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 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.

3.2 DISADVANTAGE OF EXISTINTG SYSTEM:

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.

4.PROPOSED SYSTEM

4.1 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.

4.2 ADVANTAGE OF PROPOSED SYSTEM:

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.

5.IMPLEMENTATION

5.1 Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Attack Status, View Attack Status Ratio, Download Trained Data Sets, View Attack Status Ratio Results, View All Remote Users.

5.2 View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

5.3 Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT ATTACK STATUS TYPE, VIEW YOUR PROFILE.

6.CONCLUSION

In this paper, 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 D 128) exhibits higher performance (accuracyD95.66%) 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 D 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 Ef_cientNet-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 up loaders to deceive the audiences. In the future, we intend to combine the temporal stream using optical low 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.

7. FUTURE ENHANCEMENT

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

8.REFERENCES

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/

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.

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.

J. Marsh, L. Law, J. Lahmar, D. Yamada-Rice, B. Parry, and F. Scott, *Social Media, Television and Children*. Shef_eld, U.K.: Univ. Shef_eld, 2019. [Online]. Available: https://www.stac-study.org/downloads/ STAC_Full_Report.pdf

L. Ceci. *YouTube_Statistics & Facts*. Accessed: Sep. 01, 2021. [Online]. Available: https://www.statista.com/topics/2019/youtube/

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.

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

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.

S. Maheshwari. (2017). *On YouTube Kids, Startling Videos Slip Past Fil- ters*. 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.

A. Ali and N. Senan, ``Violence video classi_cation 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.

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

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

B. Burroughs, ``YouTube kids: The app economy and mobile parenting," *Soc. media*C *Soc.*, vol. 3, May 2017, Art. no. 2056305117707189, doi: 10.1177/2056305117707189.

H. Wilson, ``YouTube is unsafe for children: YouTube's safeguards and the current legal framework are inadequate to protect children from disturbing content," *Seattle J. Technol., Environ. Innov. Law*, vol. 10, no. 1, p. 8, 2020. [Online]. Available: https://digitalcommons. law.seattleu.edu/sjteil/vol10/iss1/8

S. Alshamrani, A. Abusnaina, M. Abuhamad, D. Nyang, and D. Mohaisen, ``Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in YouTube," in *Proc. Companion Proc. Web Conf.*, Apr. 2021, pp. 508_515, doi: 10.1145/3442442. 3452314.

N. Elias and I. Sulkin, ``YouTube viewers in diapers: An exploration of factors associated with amount of toddlers' online viewing," *Cyberpsychol., J. Psychosoc. Res. Cyberspace*, vol. 11, no. 3, p. 2, Nov. 2017, doi: [19] D. Craig and S. Cunningham, ``Toy unboxing: Living in a (n unregulated) material world," *Media Int. Aust.*, vol. 163, no. 1, pp. 77_86, May 2017, doi: 10.1177/1329878X17693700.

K. Papadamou, A. Papasavva, S. Zannettou, J. Blackburn, N. Kourtellis, I. Leontiadis, G. Stringhini, and M. Sirivianos, ``Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children," in *Proc. Int. AAAI Conf. Web Soc. Media*, 2020, pp. 522_533. [Online].

Available: https://ojs. aaai.org/index.php/ICWSM/article/view/7320/7174

R. Kaushal, S. Saha, P. Bajaj, and P. Kumaraguru, ``KidsTube: Detection, characterization and analysis of child unsafe content & promoters on YouTube," in *Proc. 14th Annu. Conf. Privacy, Secur. Trust (PST)*, Dec. 2016, pp. 157_164, doi: 10.1109/pst.2016. 7906950.

R. Tahir, F. Ahmed, H. Saeed, S. Ali, F. Zaffar, and C. Wilson, ``Bringing the kid back into YouTube kids: Detecting inappropriate content on video streaming platforms," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining*, Aug. 2019, pp. 464_469, doi: 10.1145/3341161. 3342913. [23] A. Ulges, C. Schulze, D. Borth, and A. Stahl, ``Pornography detection in video bene_ts (a lot) from a multi-modal approach," in *Proc. ACM Int. Workshop Audio Multimedia Methods Large-Scale Video Anal.*, 2012, pp. 21_26, doi: 10.1145/2390214.2390222.

C. Caetano, S. Avila, S. Guimaraes, and A. D. A. Araújo, ``Pornography detection using BossaNova video descriptor," in *Proc. 22nd Eur. Signal Process. Conf.*, 2014, pp. 1681_1685. [Online]. Available: https://ieeexplore.ieee.org/document/6952616

L. Duan, G. Cui, W. Gao, and H. Zhang, ``Adult image detection method base-on skin color model and support vector machine," in *Proc. Asian Conf. Comput. Vis.*, 2002, pp. 797_800. [Online]. Available: http://aprs.dictaconference.org/accv2002/accv2002_proceedings/Duan797.pdf