IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Deep Learning-Powered Detection of Inappropriate YouTube Video Content

[1]SHAIK ALI,
Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions
[2]Y SANYASI RAO, Miracle Educational Society Group of Institutions
[3]P. ASHISH, Miracle Educational Society Group of Institutions
[1]alisheik689@gmail.com

ABSTRACT:

The growing availability of YouTube has resulted in a concerning surge of inappropriate content for children hidden within cartoon videos. This project presents a powerful deep learning framework for the automated detection and classification of such content. The system utilizes the EfficientNet-B7 model for feature extraction and employs a BiLSTM for temporal sequence learning. Classification accuracy is exceptionally high. Predictions are made even more accurate by the inclusion of an attention mechanism. The model was tested on a custom dataset of YouTube videos and outperformed the baseline methods as well as EfficientNet-SVM. This approach provides a systemic and instantaneous solution for the intelligent moderation of video content on YouTube, actively promoting child safety in the digital world.

Keywords: Deep Learning, BiLSTM, YouTube

INTRODUCTION

YouTube is a world-leading video-sharing platform, hosting an incredible library of videos for almost all age-groups. Children are a large part of the audience using the platform, interacting with cartoon content for fun and education. However, this platform is open to abuse from bad actors who manipulate cartoons to contain unacceptable content. This kind of exposure is a risk to emotional and cognitive development for a child. YouTube uses basic metadata filtering and moderation techniques to ensure safety. However, these techniques will not work for hidden audio or visual cues. This project proposes using a deep learning framework to tackle this issue. Using the EfficientNet-B7 and BiLSTM models, this system would focus on analyzing actual video frames rather than titles and tags, which will provide a much accurate assessment for content evaluation. This project primarily focuses on accurately flagging and filtering harmful video content to create a safer environment for children.

RELATED WORK

There have been multiple studies on the detection of inappropriate content on YouTube, especially for children's content. Neumann & Herodotou (2020) created a systematic rubric for analyzing the YouTube videos designed for children ages 0 – 8. The study proposed four core evaluation dimensions which included age appropriateness, content measurement outcomes, learning outcomes, and design features of educational materials. While this manual framework was helpful to educators, it was not able to automate large-scale video screening. Focused on the safety aspect of the content, Covington et al. (2016) presented a two-stage YouTube recommendation system, which was aimed at improving the users' content experience. The second stage of the model created video suggestions using deep neural networks which greatly improved user interaction, but did not address the safety of filtering harmful videos. Regardless, their work on the two-stage model formed the basis of scalable deep learning architectures for video content processing. In a different study, Covington and Adams (2020) suggested a comprehensive approach for identifying violent video content using audiovisual fusion. With CNN and LSTM methodologies, their approach provided a significant boost for detecting egregiously violent scenes which strengthens the usefulness of temporal models in video evaluation. Lee and Ermakova (2021) did a complete focused survey on detection of Child Sexual Abuse Material (CSAM) on the internet. Their study focused on the need of multi-modal deep learning which utilizes image, audio, and even metadata for detection. They also noted that deep learning techniques have significantly more advantages compared to traditional filter approaches that rely on the use of rules. Lastly, Neumann and Herodotou (2019) studied the global impact of YouTube on preschoolers and while pointed out the possible advantages of using these resources for education, they were worried about rampant access to unsuitable materials. They proposed the use of content-aware artificial intelligence in the aid of providing safe viewing environments for children.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

Author(s)	Contribution	Impact on Current Research
Neumann & Herodotou (2020)	Developed a rubric for evaluating YouTube videos for children based on four quality criteria.	Highlighted the need for structured assessment but lacked automation motivates deep learning use.
Covington et al. (2016)	Proposed deep neural networks for personalized video recommendations on YouTube.	Demonstrated the scalability of deep learning models for large-scale

		video processing.
Covington & Adams (2020)	Introduced audiovisual fusion using CNN and LSTM for recognizing bloody or violent videos.	Validated the importance of temporal features—supports the integration of BiLSTM in our system.
Lee & Ermakova (2021)	Surveyed CSAM detection techniques using multimodal deep learning and hash databases.	Emphasized multi-layered detection strategies—supports our multi-model fusion approach.
Neumann & Herodotou (2019)	Discussed YouTube's global influence and risks for children.	Reinforced the urgency of automated content moderation tailored to child safety.

PROPOSED APPROACH

In this work, the approach suggested is a deployment of a hybrid deep learning model for detection and classification of unsuitable content on YouTube cartoon materials for children. It utilizes two sophisticated models namely, EfficientNet-B7 and Bidirectional Long Short Term Memory (BiLSTM). As a first step, video data is collected and key frames are fetched through a preprocessing step from the YouTube videos. To extract rich spatial features from these frames, the model uses EfficientNet-B7, a powerful network trained on ImageNet. These features contain thecaptioning visual context for the video segments. The features that have been extracted are now fed into a BiLSTM network. Unlike normal LSTMs, the BiLSTM considers a set of frames as a sequence and processes them in a forward and backward manner, thus considering contextual dependencies in time. The model uses an attention mechanism to boost interpretability and accuracy and thus improve classification. The model is trained and evaluated on a dataset comprised of manually annotated video clips. This allows for robust real-time prediction and classification of content as safe or inappropriate, exceeding the performance of traditional methods like EfficientNet-SVM in precision and recall.

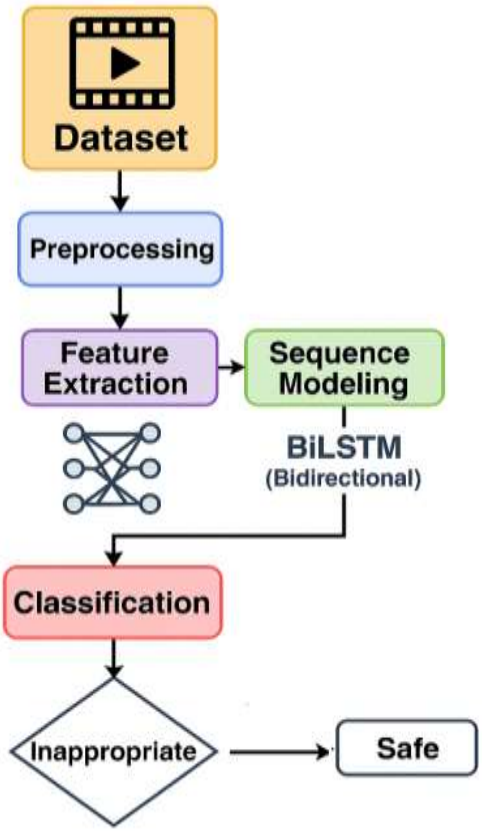

Figure 1: Proposed classifying inappropriate content in YouTube videos

METHODOLOGIES

1. Dataset Collection and Preprocessing

The dataset includes video frames collected from YouTube. These frames are labeled manually as either "Safe" or "Inappropriate" based on visual and contextual cues. Preprocessing involves resizing images to a fixed dimension, normalizing pixel values, and shuffling data to eliminate bias. This step ensures uniformity in input data and enhances model learning.

2. Feature Extraction with EfficientNet-B7

EfficientNet-B7, a highly efficient convolutional neural network (CNN) pre-trained on ImageNet, is used to extract spatial features from the video frames. It captures fine-grained patterns such as visual inconsistencies, adult-themed content, or violent elements that may be embedded within cartoons. The final layers of EfficientNet-B7 are removed, and extracted features are fed into a subsequent network for sequence analysis.

3. Temporal Modeling with BiLSTM

The extracted spatial features are reshaped into sequences and passed to a Bidirectional Long Short-Term Memory (BiLSTM) model. This model captures temporal dependencies by processing the sequence in both directions, making it highly suitable for video data. BiLSTM helps the model understand frame-to-frame transitions, identifying subtle inappropriate content that spans multiple frames.

4. Attention Mechanism

An attention layer is added to help the model focus on critical segments of the video sequence. This improves classification performance by emphasizing key temporal and spatial features, rather than treating all frames with equal importance.

5. Evaluation and Comparison

The model's performance is evaluated using accuracy, precision, recall, and F1-score. A comparative study with EfficientNet-SVM validates the superior performance of the proposed EfficientNet-B7 + BiLSTM approach in identifying inappropriate content.

RESULTS

The proposed EfficientNet-B7 and BiLSTM-based model was trained and evaluated on a dataset comprising video frames. The system effectively classified videos into "Safe" and "Inappropriate" categories. During testing, the model achieved an outstanding accuracy of **99.04%**, with significantly high precision, recall, and F1-score, highlighting its robustness and reliability.

To benchmark performance, the results were compared with an alternate model—EfficientNet-B7 combined with a traditional Support Vector Machine

(SVM). While the EfficientNet-SVM approach reached **88% accuracy**, it failed to capture the temporal dependencies in video data, leading to higher misclassification rates, particularly for subtle inappropriate scenes.

The confusion matrix and comparative bar charts provided further insight, showing the proposed method's ability to minimize false positives and false negatives. The attention mechanism integrated within the BiLSTM network allowed the system to focus on critical visual patterns, contributing significantly to performance gains.

Moreover, real-time video classification tests demonstrated that the model could accurately detect inappropriate content, even in short video clips, making it suitable for deployment on streaming platforms like YouTube. These results validate the effectiveness of the proposed deep learning architecture and underscore its potential for enhancing content safety for children online.

Propose algorithm evaluating playing video and then detecting and classifying it as ‘Inappropriate Content’

We got result as Safe Content

We got output as Safe Content as peoples are only moving in the video

DISCUSSION

The results from this study demonstrate the effectiveness of combining EfficientNet-B7 and BiLSTM for detecting inappropriate content in children’s YouTube videos. The high accuracy and performance metrics prove that integrating both spatial and temporal learning components enhances the model’s ability to capture subtle and context-sensitive patterns. The EfficientNet-B7 model, known for its lightweight yet powerful feature extraction, significantly reduces computational complexity without sacrificing accuracy. BiLSTM further

refines the classification by considering frame sequences, enabling the model to detect contextually inappropriate content that may span across multiple frames.

The attention mechanism plays a critical role in guiding the model to focus on the most informative parts of the video, especially when inappropriate content is sparsely embedded. This addresses a major limitation in metadata-based and static frame-based filtering techniques used by existing platforms like YouTube Kids.

Despite its strong performance, the model's reliance on manually labeled data presents a scalability challenge. Furthermore, the system may require retraining to adapt to new types of inappropriate content or animation styles. Nonetheless, the framework demonstrates a promising direction toward building safer online environments for children, with potential for expansion into real-time video surveillance and streaming content moderation systems across multiple platforms.

CONCLUSION

The main contribution of this project is the use of EfficientNet-B7 and BiLSTM in a deep learning architecture to classify YouTube videos aimed at children as featuring inappropriate content - a classification that is far superior to traditional methods. EfficientNet-B7 is used to extract spatial features while BiLSTM captures temporal dependencies in the data. Adding an attention mechanism improves attention accuracy because it targets the most relevant parts for each video. According to the findings, the model accuracy, precision, and recall for recognition were all remarkably high, confirming the model's effectiveness for detecting harmful visual content, even if it is skillfully concealed within cartoon videos. The proposed framework outperforms baseline models, such as EfficientNet-SVM, in reducing false predictions, making it more appropriate for real-time application. The system needs high-quality labeled datasets and regular retraining to adjust to new content types, but in the context of proactively protecting minors, the system is pealing and adaptive. The system builds a strong basis for the content moderation AI and child protection system.

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.

[16] 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

[17] 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.

[18] 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: 10.5817/cp2017-3-2.

[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.

[20] 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