International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Detecting and Classifying Inappropriate Content in Youtube Videos Using Deep Learning Approach

Sanaboina Chandra Sekhar[1], Yandamuri Eswara Anil[2]

[1]Assistant Professor, Department of Computer Science and Engineering, University College of Engineering Kakinada,Jawaharlal Nehru Technological University Kakinada

[2]Post Graduate Student, Department of Computer Science and Engineering, University College of Engineering Kakinada,Jawaharlal Nehru Technological University Kakinada

Abstract: ***The proliferation of hate speech, pornography, and violence in online platforms is a significant concern, especially in video content on platforms like YouTube. An automated solution can help identify and remove such content, creating a safer and more positive online environment. The main objective of this project is to identify and classify undesirable content in YouTube videos using a variety of deep learning approaches. The suggested method analyzes video frames and audio segments using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. Creating a dataset of YouTube videos, pre-processing them to extract pertinent visual and audio attributes, and then training the CNN-LSTM model to discover the spatial and temporal relationships between the video frames and audio segments are all steps in the procedure. On a test set of YouTube videos that have been flagged as unsuitable or not by human annotators, Using measurements of accuracy, precision, recall, and F1-score, the model's performance will be evaluated.***

Keywords: Detection, Classification, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks

1. Introduction

A Brief Overview of Youtube

Users can publish, share, and watch videos on the well-known web platform YouTube. Three former PayPal workers, Chad Hurley, Steve Chen, and Jawed Karim, Since its creation in February 2005. YouTube revolutionized the way people consume and interact with video content, offering a diverse range of videos, from educational tutorials and music videos. The platform provides a space for content creators, known as YouTubers, to display their abilities, they may communicate with audiences worldwide and express themselves. Anyone can create a YouTube channel and start uploading videos, Consequently, it is a useful tool for individuals, businesses, and organizations to exchange ideas and promote their products or services. It offers, first and foremost, a huge and continuously growing library of videos, covering almost every imaginable topic. Whether you're looking for educational content, entertainment, news updates, or even niche interests, chances are you'll find relevant videos on YouTube. Moreover, commenting, and subscribing to channels they enjoy. This creates a sense of community and enables discussions and interactions between creators and their viewers. YouTube has not only become an entertainment platform but also an influential media channel. Many YouTubers have gained massive followings and have even become celebrities in their own right. The platform has created new possibilities for content producers to monetise their channels through sponsorships, merchandise sales, advertising revenue, and other collaborations. However, it is important to note that YouTube has faced challenges as well, including issues related to copyright infringement, misinformation, and content moderation. The platform has implemented various policies and guidelines to address these concerns and ensure a safer and more inclusive environment for users.

Overall, YouTube has had a profound impact on the way we consume and create video content. It has democratized media, allowing individuals from all walks of life to express themselves, share their stories, and connect with a global audience, making it an integral part of today's digital landscape.

Problem Statement

A unique deep learning-based approach has been created for the identification and categorization of unsuitable video content for children. EfficientNetB7 architecture is used in transfer learning to extract the characteristics of videos.

2. Literature Review

The comprehension of various deep learning algorithms is a key focus of this literature review, along with the identification of suitable deep learning algorithms that can be applied to detection. We took a number of stages in our investigation, including the following:

Hou et al., (2018) The restriction to single vision-modality approaches is addressed by the idea of combining aural and visual information to create a bleeding video recognition system. In the lack of available malicious video data, this effort built a database of gory movies using web crawlers and data augmentation methods. After that, it extracted the spatiotemporal characteristics of the visual channels using the CNN and LSTM algorithms. In the meanwhile, the 1D convolutional network directly retrieved the audio channel properties from the original waveforms.

Sumon et al., (2020) In order to detect violent videos, we looked at a number of ways for judging the importance of the features from various pretrained models. There is a dataset of violent and nonviolent videos from different circumstances. VGG16, VGG19, and ResNet50, three ImageNet models, are used to extract features from the frames of the video. In one of the trials, a fully connected network that identifies

violence at the frame level was fed the retrieved data. Additionally, in a different experiment, we input 30 frames' worth of extracted features to an LSTM network at a time.

Honarjoo et al., (2021) Determine if a violent act has occurred is the aim of violence detection, according to the author. Since there was a need for practical, automatic violence detection techniques that analyzed the visual information obtained from security cameras shared throughout diverse places, this business saw significant growth. In this paper, we provide a low-complexity, pre-trained deep neural network technique for violence detection. To determine if a violent action has taken place, the collected attributes from pre-trained models have been combined and incorporated to a fully connected network. The suggested approach evaluates the ResNet-50 and VGG16 outputs as trained models.

Ali and Senan et al.,(2018) In order to investigate the impact of the number of hidden layers and hidden nodes on the classification performance, alternative designs of hidden layers and hidden nodes in DNN have been created in this study utilizing the try-error technique and equation-based method. The outcomes show that the try-and-error strategy has an accuracy rate of 53%, while the equation-based method has a 51% accuracy rate.

Kaushal et al.,(2016) The prevalence of inappropriate content for youngsters and users who support it are the main topics of this essay. We employ two different techniques for the detection of kid-unsafe content and its promoters: the first is based on supervised classification and makes extensive use of video- level, user-level, and comment-level characteristics; the second is based on convolutional neural networks and relies on the use of video frames. 85.7% detection accuracy is achieved, and this information can be used to create a system that will give youngsters a secure YouTube experience.

Objectives

1) To Study the Deep learning models can automate the content moderation process, making it more efficient and scalable.
2) To make sure that marketers' brands are not connected to objectionable or dangerous content requires the identification and classification of improper content.
3) Users may have a safer and more satisfying experience on the site by flagging and classifying inappropriate content.

3. Research Methodology

Existing System

The current content moderation system on YouTube primarily relies on a combination of automated algorithms and human reviewers. Copyrighted content is identified using YouTube's Content ID system, while community flagging helps users report potentially inappropriate content. Additionally, YouTube employs human moderators who review flagged content and apply community guidelines.

However, this system has certain limitations. First, it heavily relies on user reports, making it reactive rather than proactive. Secondly, human moderators can be resource-intensive and may not always maintain a consistent standard across all content. These factors result in delays in content removal, and sometimes, inappropriate content can go unnoticed.

4. Proposed System

The suggested solution uses deep learning to identify and categorize improper content in videos with the goal of improving content moderation on YouTube. The technology will make use of a dataset of movies that has been carefully selected and spans a wide variety of unsuitable content, from graphic images to hate speech. Convolutional neural networks (CNN) and recurrent neural networks (RNN), two cutting-edge deep learning models, will be used to train the system to recognize detailed elements and patterns in unsuitable content. The model will be taught to classify videos into predetermined categories including nudity, hate speech, violence, and more through supervised learning. To prevent such content many deep learning algorithms and machine learning are introduced but their inappropriate content detection and classification accuracy is not up to the mark. To overcome from this issue author of this paper employing combination of CNN and BI-LSTM algorithm to detect inappropriate content. To extract features from the YouTube video pictures, the pre-trained EfficientNetB7 CNN method is used. It is then retrained with the BI- LSTM technique to improve prediction accuracy.

System Architecture

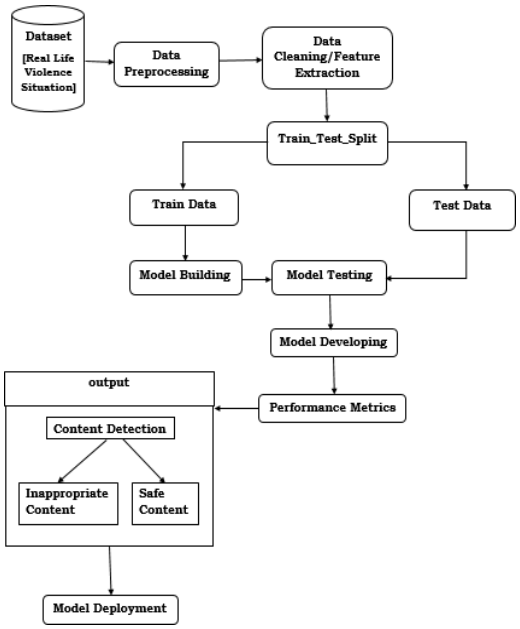

Figure 4.1: System Architecture

Real Life Violence Dataset:

The first module is where we gather the data. The actual process for developing a deep learning model and accumulating data establishes now. Datasets are gathered for detections in this study. They are Violence and Non-Violence Videos. Both of these data sets were gathered from kaggle.com. Both 1000 violence and 1000 non- violence videos are added in the Real Life Violence Situation dataset.

Data Preprocessing

Data Collection:
Gather a Real Life Violence dataset from kaggle.com

Data Labelling:
Annotate the dataset by labelling each video with appropriate categories such as “Safe content”, “Inappropriate content”.

Data Cleaning:
Remove duplicate Videos, irrelevant videos, or those not related to the task. Ensure your dataset is well-structured and consistent.

Feature Extraction:
In Feature Extraction for each frame, extract relevant features that can help the model differentiate between Safe content and Inappropirate content.

Train_Test_Split:

Train Data:
In order to be enhanced for a better solution to the problem in the following stage, the model needs to be trained. We use datasets to train the model using various deep learning approaches. In order for a model to comprehend the various patterns, laws, and features, it must be taught.In this procedure, training is allocated for 80% of the data.

Test data:
In this stage, we assess the accuracy of our model using a test dataset. Testing the model evaluates its accuracy percentage in accordance with the requirements of the project or challenge. 20% of the data in this process is set aside for testing.

Model Building:
Deep learning models are potent instruments used to successfully complete important jobs and resolve challenging issues. Building a CNN(Convolutional Neural Network) model for detecting and classifying inappropriate content in youtube videos.

Model Testing:
"Model testing" is the term used in deep learning to describe the assessment of a fully trained model's performance on a testing set. With the same probability distribution as the training and validation sets, the testing set should be kept apart from them.It is made up for several test samples. The target value is known for each testing sample.

Performance Metrics:
The effectiveness or quality of the model is evaluated using a number of indicators, also referred to as performance metrics or evaluation metrics. These performance metrics allow us to evaluate how effectively our model processed the given data.

Performance Metrics for Classification:
- Accuracy
- Precision
- Recall
- F1-Score

Output:
Content Detection:
In this stage, we are showing the project's outcomes. They are
- Inappropriate Content
- Safe Content

2. Results

Confusion Matrix Representation for Efficient Net-BiLSTM:

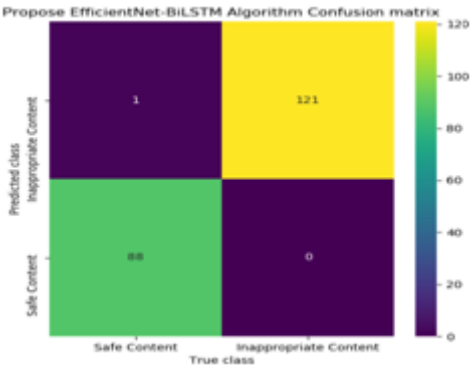

Figure 5.1: Confusion matrix for EfficientNet- BiLSTM

The confusion matrix is depicted in Figure 3 between the predicted class labels to the actual class labels of a dataset. In confusion matrix

True Positive
True Positive (TP): 121 – This represents how many instances were correctly anticipated as being positive.

False Negative
False Negative (FN): 0 – This indicates situations where the outcome was projected to be positive but turned out to be negative.

False Positive
False Positive (FP): 1 – This is the number of cases where a negative outcome was mistakenly expected to be a positive outcome.

True Negative
True Negative (TN): 88 – The number of accurately anticipated negative instances.

Confusion Matrix Representation for Efficient Net-SVM

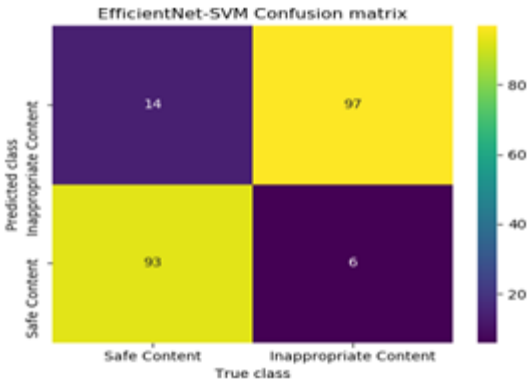

Figure 5.2: Confusion matrix for Efficient Net-SVM

True Positive
True Positive (TP): 97 – This represents how many instances were correctly anticipated as being positive.

False Negative
False Negative (FN): 6 – This indicates situations where the outcome was projected to be positive but turned out to be negative.

False Positive
False Positive (FP): 14 – This is the number of cases where a negative outcome was mistakenly expected to be a positive outcome.

True Negative
True Negative (TN): 93 – The number of accurately anticipated negative instances.

Comparison Graph
In this Comparison Graph is drawn with the help of Propose EfficientNet BiLSTM and EfficientNet SVM Accuracy, F1-Score, Precision, Recall Values.

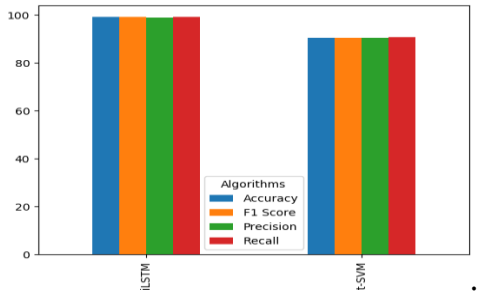

Figure 5.3.1: Graphical representation between Efficient Net-BiLSTM and Efficient Net-SVM Algorithm

Accuracy, F1-Score, Precision, and Recall Values were shown in Fig. 6.2.1 Shows EfficientNet- BilLSTM and EfficientNet-SVM Algorithms employing the distinct hues Blue, Orange, Green, and Red.

Content Detection

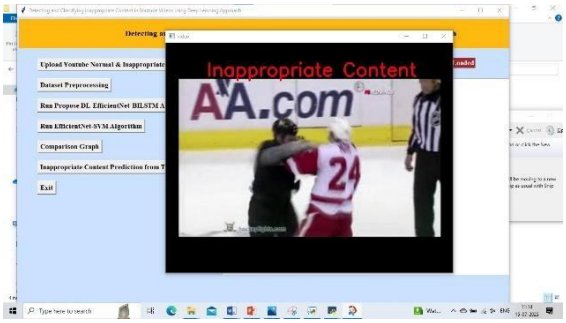

Figure 5.4.1: Detects Inappropriate content from Test Video

Using the Efficient Net-BiLSTM Algorithm, which has a greater accuracy value than the EfficientNet- SVM Algorithm. Fig.5.4.1 detects inappropriate content from the test video.

Figure 5.4.2: Detects Safe content from Test Video

Using the EfficientNet-BiLSTM Algorithm, which has a greater accuracy value than the EfficientNet- SVM Algorithm. Fig.5.4.2 detects inappropriate content from the test video.

5. Conclusion and Future Scope

5.1 Conclusion

A unique deep learning-based method for categorizing and identifying objectionable YouTube video material is presented in this paper. Advanced deep learning approaches include convolutional neural networks (CNNs) and recurrent neural networks (RNNs), this system has demonstrated its effectiveness in accurately identifying and categorizing inappropriate content, including violence, hate speech, nudity, and other explicit material. Adaptive learning EfficientNet-B7 architecture is employed to extract the characteristics of videos. The BiLSTM network is used to handle the collected video features. Here, the model conducts multiclass video classification and discovers effective video representations. Furthermore, Achieving the greatest recall score in a performance comparison with existing state-of-the-art models, our BiLSTM- based framework beat other contemporary models and approaches. The EfficientNet-B7 architecture for adaptive learning is used to extract characteristics of videos. The BiLSTM network is used to handle the collected video features. Here, the model conducts multiclass video classification and discovers effective video representations.

5.2 Future Scope

In The ability to detect and classify inappropriate content in real-time is crucial for platforms like YouTube. Balancing content moderation with user privacy is an ongoing challenge.

The development of shared datasets, benchmarks, and collaborative research efforts among different platforms and researchers can accelerate advancements in deep learning-based inappropriate content detection.

References

[1] C. Hou, X. Wu, and G. Wang, “End-to-end bloody video recognition by audio-visual feature fusion,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2018, pp. 501–510. doi: 10.1007/978-3-030-03398-9_43.

[2] S. A. Sumon, R. Goni, N. Bin Hashem, T. Shahria, and R. M. Rahman, "Violence Detection by Pretrained Modules with Different Deep Learning Approaches," Vietnam Journal of Computer Science, vol. 7, no. 1, pp. 19–40, Feb.2020,doi:10.1142/S2196 88820500013.

[3] N. Honarjoo, A. Abdari, and A. Mansouri, "Violence detection using pre-trained models," in Proceedings of the 5th International Conference on Pattern Recognition and Image Analysis, IPRIA 2021, Institute of Electrical and Electronics Engineers Inc., Apr.2021. doi: 10.1109/IPRIA53 572.2021.948355

[4] A. Ali and N. Senan, "Violence video classification performance using deep neural networks," in Advances in Intelligent Systems and Computing, Springer Verlag, 2018, pp. 225–233. doi: 10.1007/978-3-319-72550-5_22.

[5] R. Kaushal, S. Saha, P. Bajaj, and P. Kumaraguru, "KidsTube: Detection, Characterization and Analysis of Child Unsafe Content & Promoters on YouTube."[Online].Available:https:// developers.google.com/youtube/v3/

[6] W. Afandi, S. M. A. H. Bukhari, M. U. S. Khan, T. Maqsood, and S. U. Khan, "Fingerprinting Technique for YouTube Videos Identification in Network Traffic," *IEEE Access*, vol. 10, pp. 76731 76741,2022,doi:10.1109/ACCESS.2022. 3192458.

[7] P. Rajpurkar *et al.*, "AppendiXNet: Deep Learning for Diagnosis of Appendicitis from A Small Dataset of CT Exams Using Video Pretraining," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-61055-6.

[8] V. Anand, R. Shukla, A. Gupta, and A. Kumar, "Customized video filtering on Youtube."

[9] S. Singh, A. B. Buduru, R. Kaushal, and P. Kumaraguru, "KidsGUARD: Fine grained approach for child unsafe video representation and detection," in *Proceedings of the ACM Symposium on Applied Computing*, Association for Computing Machinery, 2019, pp. 2104– 2111. doi: 10.1145/3297280.3297487.

[10] S. I. Alqahtani, W. M. S. Yafooz, A. Alsaeedi, L. Syed, and R. Alluhaibi, "Children's Safety on YouTube: A Systematic Review," *Applied Sciences (Switzerland)*, vol. 13, no. 6. MDPI, Mar. 01, 2023. doi: 10.3390/app13064044.