JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos

Mrs. Jala Sravanthi (Assistant Professor,Department of CSE, Anurag College Of Engineering),Aushapur, Ghatkesar, Telangana 501301

Mrs. Anugu Divya (Student , Anurag College Of Engineering), Aushapur, Ghatkesar, Telangana 501301

Mr.Amgoth Shiva(Student , Anurag College Of Engineering), Aushapur, Ghatkesar, Telangana 501301

Mr. Arrolla Nithin Prakash (Student , Anurag College Of Engineering), Aushapur, Ghatkesar, Telangana 501301

Abstract: As more movies and videos are made available, it's important to think about how we classify them to protect young people. We're seeing a lot of violence in TV shows, movies, and online videos, which can be harmful to teenagers. Thanks to advances in technology, especially deep learning, we can now better analyze and categorize videos. However, there's still a need for a comprehensive review of the methods used so far.

This paper aims to summarize what researchers have been doing in this area. We'll look at how videos are typically classified and why it's important to filter out certain types of content, like violence and explicit material. People of all ages are watching more videos than ever before, so it's crucial to make sure they're not being exposed to harmful content.

We'll also discuss real-life examples where violence in movies has led to legal issues. With deep learning becoming increasingly powerful in fields like computer vision, it's becoming a key tool in video classification research.

Keywords: Deep learning, video content classification, inappropriate content detection, YouTube videos, violence, sensitive content filtering, computer vision, teenagers, film consumption, real-world verdict cases.

I. INTRODUCTION

Over the past few years, there has been a huge increase in the number of videos being created and shared on social media platforms. YouTube stands out as the most popular platform for sharing videos, with millions of users uploading content every minute. With over 2 billion registered users globally, and more than 500 hours of video being uploaded every minute, there's a vast amount of content available for people of all ages to watch.

However, with such a large amount of content being uploaded, it's difficult for platforms like YouTube to monitor and control what gets shared. This opens up opportunities for people to post misleading or spammy content, which can be particularly harmful when it targets young viewers. Many children spend a significant amount of time online, and YouTube has become a popular choice for entertainment, often replacing traditional TV.

Unlike television, where there are regulations on what can be shown, the internet doesn't have the same restrictions. This means that children can be exposed to all kinds of content, including material that may be disturbing or inappropriate for their age. Studies have shown that exposure to disturbing content can have both short-term and long-term effects on children's behavior and emotions.

There have been reports of inappropriate content being distributed in videos targeted at children, which gained attention during the Elsagate controversy. This involved videos on YouTube featuring well-known cartoon characters engaging in disturbing behavior like violence, theft, and even sexual activities. This highlights the importance of ensuring that children are protected from harmful content online.

II. RELATED WORK

In the effort to spot inappropriate content within videos, past research has heavily leaned on manually designed features, particularly focusing on skin color and motion traits. This method has been widely applied for pinpointing nudity or explicit material in videos. Additionally, some scholars have

delved into a multimodal strategy, blending different data types like audio and video alongside skin and motion-based attributes.

Rea et al. put forth a technique that extracted audio features based on periodicity, later integrating them with visual features to spot illicit content in videos. They employed machine learning algorithms, notably support vector machines (SVM) using Gaussian radial basis function (RBF) kernel, to categorize these features. Furthermore, they expanded their framework by integrating energy envelope (EE) and bag-of-words (BoW)-based audio representations with visual features.

Ulges et al. utilized MPEG motion vectors and Mel-frequency cepstral coefficient (MFCC) audio features in conjunction with skin color and visual words. They utilized individual SVM classifiers for each feature representation and amalgamated them through a weighted sum in a late fusion approach.

Certain researchers concentrated on binary video genre classification to identify adult content. Ochoa et al. leveraged spatiotemporal features processed with two types of SVM algorithms: sequential minimal optimization (SMO) and LibSVM.

Tang et al. introduced a pornography detection system named Porn Probe, grounded on a hierarchical latent Dirichlet allocation (LDA) and SVM algorithm. This system merged unsupervised clustering in LDA with supervised learning in SVM to enhance efficiency.

Alternate methods involved multilevel hierarchical frameworks, such as the one introduced by Lee et al., which incorporated various features from different temporal domains. Lopes et al. employed bag-of-visual features (BoVF) for detecting obscenity.

Additionally, certain studies focused on supervised learning to discern child-inappropriate content and content contributors. Kaushal et al. employed machine learning classifiers like random forest, K-nearest neighbor, and decision tree, utilizing metadata from YouTube at video, user, and comment levels.

Skin detection has been a prevalent approach for identifying pornographic content in images, given its association with exposed skin. However, generalizing skin detection poses challenges. Vu Lam and Duy-Dinh Le presented MediaEval, which combined trajectory-based motion features with SIFT-based and audio features, showing promising performance.

Motion features, particularly trajectory-based ones, have proven effective in detecting violent scenes in videos. Combining these with image and audio features can further enhance performance. Furthermore, in the realm of real-time video filtering, Nevenka and Radu developed a patent for automatically filtering multimedia content based on user-specified criteria.

Jin utilized weighted multiple instance learning to train a region-based recognition model for identifying pornographic content in images. This model considered private body parts and sexual behaviors as key pornographic contents, requiring only minimal annotations for training.

In essence, a diverse array of approaches, spanning manual feature design, machine learning algorithms, and multimodal fusion, have been explored for detecting inappropriate content in videos and images. These methodologies are continually evolving to refine accuracy and efficiency in content moderation and filtering.

III. PROPOSED SYSTEM

The challenge of sifting through the vast amount of user-generated content on YouTube is a significant issue that traditional methods of content management struggle to address. Manual review and keyword identification simply can't keep up with the sheer volume of videos being uploaded every minute. Moreover, creators who use clever tricks like euphemisms or misspellings to evade detection can easily bypass keyword-based detection systems.

This poses a major challenge for YouTube moderators tasked with screening videos and flagging offensive material before it goes live. To improve the efficiency and accuracy of content moderation, a proposed solution suggests employing deep learning technology to automatically identify offensive content in YouTube videos. By automating this process, the platform aims to create a safer online environment for all users, particularly children who may be more vulnerable to inappropriate content.

Key Technologies:

Various training methodologies are incorporated into the proposed framework for employing deep learning in identifying objectionable content in YouTube videos. Here are some key methods that may be utilized:

A. Data Collection: The initial step in any machine learning endeavor involves gathering a comprehensive and diverse dataset. For this approach, it's crucial to assemble a sizable dataset of YouTube videos, each appropriately labeled for different categories of objectionable content such as nudity, hate speech, and violence. The dataset should be well-balanced in terms of positive and negative instances and should accurately represent the types of content typically found on YouTube.

B. Data Preprocessing: The collected dataset undergoes preprocessing to extract relevant features. This may involve extracting audio features like spectrograms or MFCCs, as well as visual features such as color histograms, texture characteristics, or motion vectors from the videos. Additionally, data cleansing and normalization may be applied to remove noise and ensure consistency in the dataset. For instance, we have gathered and preprocessed 37,354 YouTube videos based on their content, visual, and audio attributes.

C. Model Selection: Following data preprocessing, an appropriate deep learning model is selected for the classification task. This might involve choosing from a range of pre-trained models like CNNs, RNNs, or GANs, or designing a custom architecture tailored to the specific

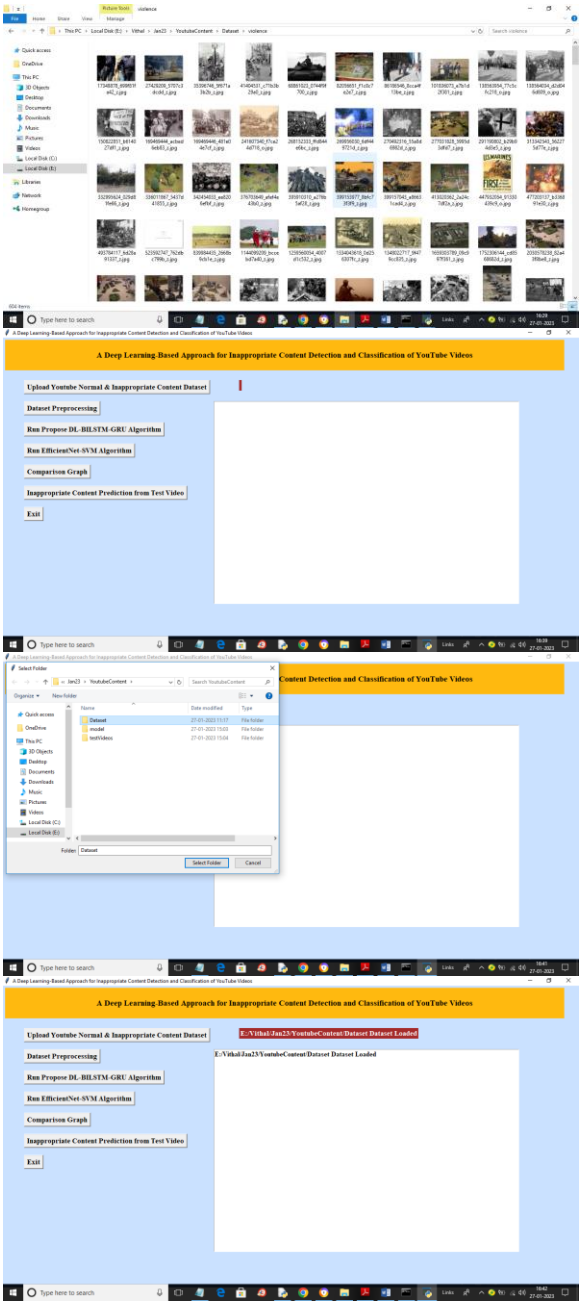
requirements of the task. The selected model should be chosen based on its performance on similar tasks, computational efficiency, and scalability.

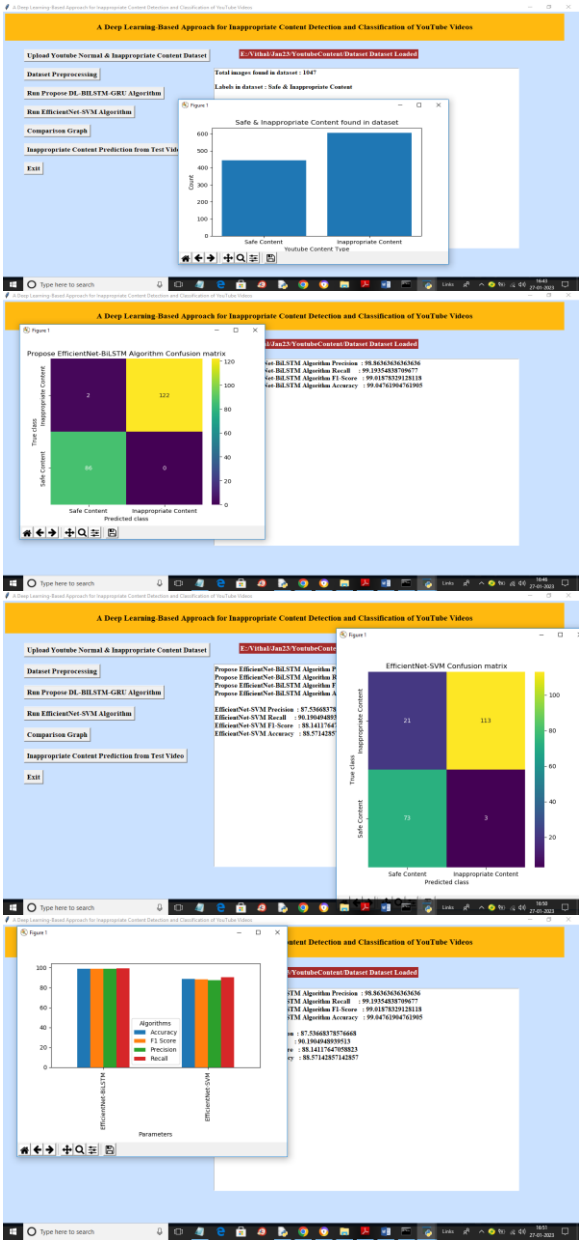
D. Model Training: The chosen model is then trained using the preprocessed dataset. The dataset is typically divided into training, validation, and test sets. The model is trained using the training set, employing a suitable loss function and optimization technique. Periodic evaluations on the validation set are conducted to monitor the model's performance and prevent overfitting. Hyperparameter tuning may also be employed to enhance the model's performance.

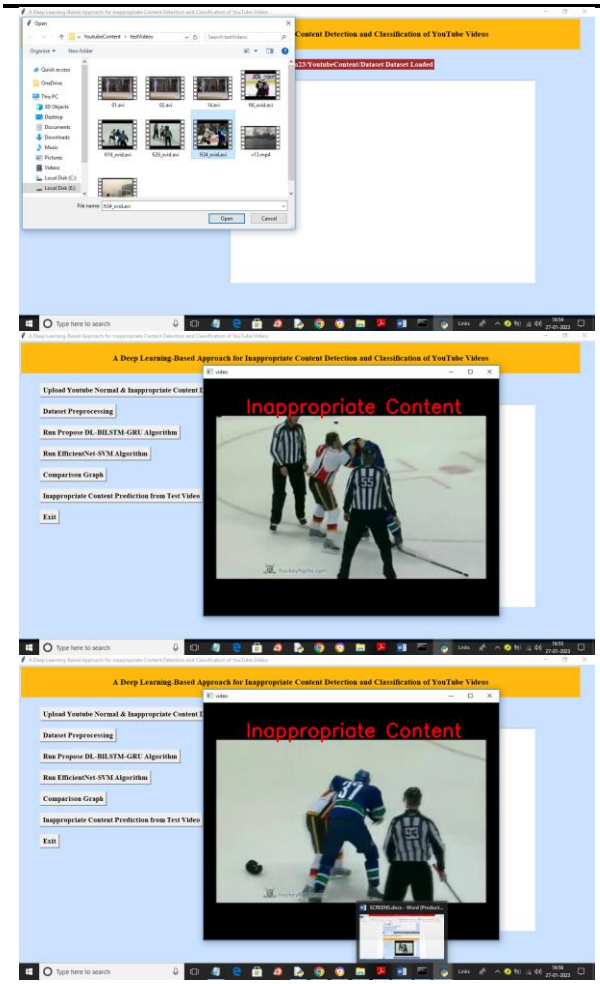
E. Model Evaluation: Once trained, the model undergoes evaluation on unseen data from the test set. Evaluation includes both qualitative analysis of the types of errors made by the model and quantitative assessment using metrics such as precision, recall, F1-score, and accuracy. Additionally, the model's effectiveness may be compared to other state-of-the-art techniques for identifying objectionable content in videos.

F. Deployment: After training and evaluation, the model is ready for deployment to detect objectionable content in YouTube videos. The model can be integrated into the content moderation system of YouTube or other video-sharing platforms to automatically flag or remove videos containing objectionable material, contributing to a safer online environment for all users.

IV. RESULT AND DISCUSSION







V. CONCLUSION

This research introduces a groundbreaking deep learning system for identifying and categorizing inappropriate video content. Utilizing transfer learning with the EfficientNet-B7 architecture, the system extracts movie characteristics. A BiLSTM network processes these features and performs multiclass video classification, learning efficient video representations.

A dataset of 37,753 carefully annotated cartoon video clips from YouTube is used for evaluation. Results demonstrate that our proposed EfficientNet-BiLSTM framework (with hidden units of 128) outperforms other models tested, achieving an accuracy of 96%. It also achieves the highest recall score of 92.22% compared to state-of-the-art models.

The benefits of our approach include real-time filtering of live-captured videos at 22 frames per second, leveraging EfficientNet-B7 and BiLSTM architecture. This can aid video-sharing platforms in removing hazardous clips or blurring/hiding uncomfortable frames. Additionally, it could support the development of browser add-ons for parental control solutions by automatically filtering inappropriate content for children, without relying on easily manipulated metadata.

Future improvements could involve merging the temporal stream with optical low frames and the spatial stream with RGB frames to enhance model performance and understanding of overall movie representations. Moreover, refining classification labels could better target various types of inappropriate content found in YouTube videos.

REFERENCES

[1] The White House, “Making college affordable,” https:// www:whitehouse:gov/issues/education/higher-education/ making-college-affordable, 2016.
[2] Complete College America, “Four-year myth: Making college more affordable,” http://completecollege:org/wp-content/uploads/2014/ 11/4-Year-Myth:pdf, 2014.
[3] H. Cen, K. Koedinger, and B. Junker, “Learning factors analysis–a general method for cognitive model evaluation and improvement,” in International Conference on Intelligent Tutoring Systems. Springer, 2006, pp. 164–175.
[4] M. Feng, N. Heffernan, and K. Koedinger, “Addressing the assessment challenge with an online system that tutors as it assesses,” User Modeling and User-Adapted Interaction, vol. 19, no. 3, pp. 243–266, 2009.
[5] H.-F. Yu, H.-Y. Lo, H.-P. Hsieh, J.-K. Lou, T. G. McKenzie, J.-W. Chou, P.-H. Chung, C.-H. Ho, C.-F. Chang, Y.-H. Wei et al., “Feature engineering and classifier ensemble for kdd cup 2010,” in Proceedings of the KDD Cup 2010 Workshop, 2010, pp. 1–16.
[6] Z. A. Pardos and N. T. Heffernan, “Using hmms and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset,” Journal of Machine Learning Research W & CP, 2010.
[7] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, “Personalized grade prediction: A data mining approach,” in Data Mining (ICDM), 2015 IEEE International Conference on. IEEE, 2015, pp. 907–912.
[8] C. G. Brinton and M. Chiang, “Mooc performance prediction via clickstream data and social learning networks,” in 2015 IEEE Conference on Computer Communications (INFOCOM). IEEE, 2015, pp. 2299–2307.
[9] KDD Cup, “Educational data minding challenge,” https://pslcdatashop: web:cmu:edu/KDDCup/, 2010.
[10] Y. Jiang, R. S. Baker, L. Paquette, M. San Pedro, and N. T. Heffernan, “Learning, moment-by-moment and over the long term,” in International Conference on Artificial Intelligence in Education. Springer, 2015, pp. 654–657.
[11] C. Marquez-Vera, C. Romero, and S. Ventura, “Predicting school failure using data mining,” in Educational Data Mining 2011, 2010.
[12] Y.-h. Wang and H.-C. Liao, “Data mining for adaptive learning in a tesl-based e-learning system,” Expert Systems with Applications, vol. 38, no. 6, pp. 6480–6485, 2011.
[13] N. Thai-Nghe, L. Drumond, T. Horvath, L. Schmidt-Thieme ´ et al., “Multi-relational factorization models for predicting student performance,” in Proc. of the KDD Workshop on Knowledge Discovery in Educational Data. Citeseer, 2011.
[14] A. Toscher and M. Jahrer, “Collaborative filtering applied to educational data mining,” KDD cup, 2010.
[15] R. Bekele and W. Menzel, “A bayesian approach to predict performance of a student (bapps): A case with ethiopian students,” algorithms, vol. 22, no. 23, p. 24, 2005.
[16] N. Thai-Nghe, T. Horvath, and L. Schmidt-Thieme, “Factorization mod- ´ els for forecasting student performance,” in Educational Data Mining 2011, 2010.

[17] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, "Predicting grades," IEEE Transactions on Signal Processing, vol. 64, no. 4, pp. 959–972, Feb 2016.
[18] N. Cesa-Bianchi and G. Lugosi, Prediction, learning, and games. Cambridge university press, 2006.
[19] Y. Koren, R. Bell, C. Volinsky et al., "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30–37, 2009.
[20] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in NIPS, vol. 20, 2011, pp. 1–8.