International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | June 2022 www.irjet.net p-ISSN: 2395-0072

INAPPROPRIATE & ABUSIVE CONTENT CENSORSHIP

Rachana P Bennur[1], Samruddhi C Shetty[2], Anagha P[3], Hema G R[4], Prof. Shalini K C[5]

[1,2,3,4] *B.E Student, Dept. of Computer Science Engineering, JSS Science & Technology University, Mysore, India - 570006*

[5] *Faculty, Dept. of Computer Science Engineering, JSS Science & Technology University, Mysore, India - 570006*

Abstract – *With the growth and easy access of internet, kids today have access to content that are not appropriate for their age. Studies have shown that this can impact their mental development negatively. To prevent a kid from watching inappropriate content, while not affecting the eligible users, is important. The current solutions that determine the age of the user based on the login credentials is not efficient in combating the problem. This project is an attempt to provide safe streaming environment by classifying the videos as nsfw or sfw and to actively predicting the age of the user while he tries to watch a nsfw video. If the user is underaged, the video will be blocked. Further, every comment is checked for its toxicity before posting on the platform.*

***Key Words*: Age Prediction, Video Classification, Sentiment Analysis, Non-Safe For Work (NSFW), Safe For Work (SFW).**

1. INTRODUCTION

Social media today is one of the best ways to connect to people, learn and engage in communications. With several social media platforms available, streaming videos has become a primary source of entertainment on social media. However, sites across internet have content varying from age-appropriate to age-inappropriate videos. Such age-inappropriate videos can affect the mental development of kids.

However, it is difficult for the parents to keep a check constantly on the content that children watch. Currently the streaming platforms blocks the NSFW content based on the age that the user has given in the login credentials. However, there is a loophole that the account belonging to an adult can be misused by a kid to watch NSFW content.

In order to combat this issue, we are building an age prediction module, which predicts the age of the user actively by capturing real time facial images when the user clicks on any video that is labelled as “age-inappropriate”. The content is blocked to the user if their age is determined as underaged (below 12). With the help of pre-trained machine learning modules, the video content uploaded on the streaming platform is labelled as “Safe for Work” and “Not Safe for Work”.

User’s safe streaming is further ensured by inclusion of censorship over abusive comments, where all the comments of a video undergo sentiment analysis and if the comments turn out to be toxic or inappropriate, they are either blurred or are not displayed to the other users in the comments section.

2. LITERATURE SURVEY

2.1 EXTRACTIVE TEXT SUMMARISATION USING DEEP NATURAL LANGUAGE FUZZY PROCESSING

The review paper "Extractive Text Summarisation using Deep Natural Language Fuzzy Processing" by Neelima G, Veeramanickam M.R.M, Sergey Gorbachev, and Sandip A. Kale focuses on the summarising of a document into a smaller set of meaningful and usable sentences. It employs fuzzy logic, with the Naïve-Bayes technique being used to identify the most essential sentences in a given document.

2.2 UNSUPERVISED KEYPHRASE EXTRACTION USING MASKED DOCUMENT EMBEDDING

Linhan Zhang, Qian Chen, Wen Wang, Chong Deng, Shiliang Zhang, Bing Li, Wei Wang, and Xin Cao's review paper "A Masked Document Embedding Rank Approach for Unsupervised Key-Extraction" focuses on extracting keywords from a huge document of text. The keywords are generated using a BERT-based model. It extracts essential phrases by selectively masking terms in the manuscript depending on their importance and use.

2.3 GENDER AND AGE DETECTION USING DEEP LEARNING

Prof. Supriya Mandar Khatavkar, Abhinav Banerjee, Hemanka Sarma, and Anurag Sharma's review paper, "Gender and Age Detection Utilizing Deep Learning Techniques," focuses on predicting an individual's age by using Convolution Neural Network algorithms (CNN) to extract face data.

2.4 SENTIMENT ANALYSIS IN SOCIAL MEDIA

The review paper "Sentiment Analysis in Social Media and Its Application: Systematic Literature Review" by Zulfadzli Drus and Haliyana Khalid investigates how to use the high-

density data accessible on social media platforms to do some useful activities by assessing the sentiment underlying the data.

3. PROPOSED WORK

The user logins into the platform using their login credentials. Then the user is directed to the homepage, where they can access the video based on their requirement. Every video present on the homepage will be labelled as age-appropriate or inappropriate based on the content classification algorithm running in the background. If the user tries to access an age-restricted video then the platform captures their image using the web camera and the age prediction algorithm determines whether the user is underaged or an adult.

If the user is underaged, the video is blocked and the user is redirected back to the homepage. Else, the user is allowed to watch the video. Additionally, the toxic comment classifier algorithm is run against all the videos on the platform to segregate toxic comments. Thus, ensuring every video of the platform is free from toxic comments.

3.1 SYSTEM ARCHITECTURE

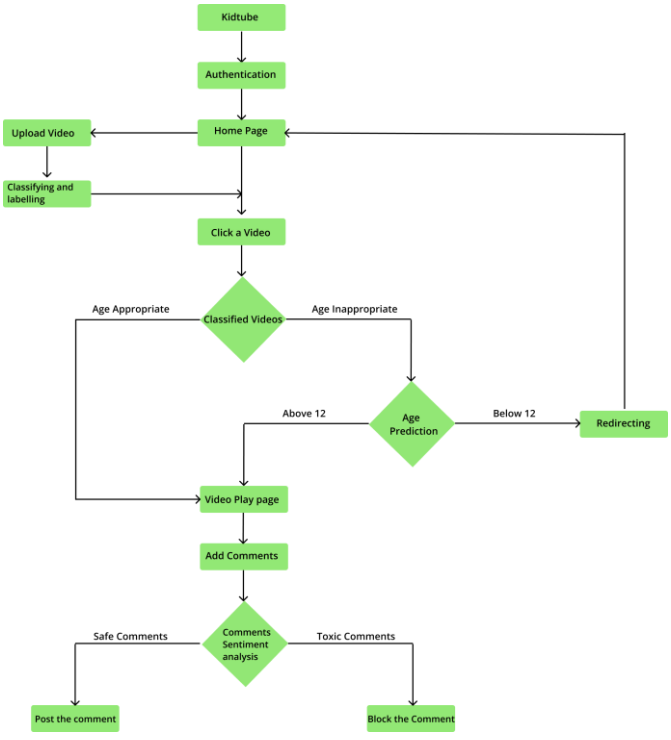

Figure 1 Flow Chart

A. Content Analysis of the video: The content analysis of the video predicts whether the video contains inappropriate content or not. It is achieved by extracting the frames of the videos and analyzing the frames using pre-trained module. To train this model, we choose Convolution Neural Network (CNN).

B. Automatic Age Estimation: The automatic age estimation module predicts whether the user is underage or appropriate aged to watch content that is recognized as "adult content". The module captures real-time facial images of the user through a web camera or the selfie camera and employs CNN architecture to classify the age of the user.

C. Automated Immodest Comment Classifier: Automated immodest comment classifier segregates toxic comments from non-toxic using NLP and RNN architectures. The toxic comments are censored on the platform.

3.2 ML MODELS AND PARAMETERS

In the solution we are using pretrained ML models for age-detection, video classification and comment classification. Age Prediction is achieved by using Caffenet model, which is a CNN (convolutional Neural Network) model that works by taking a set of DAGS. Yahoo-NSFW model classifies a video by capturing random frames in a video and analyzing it.

3.3 METHODOLOGY

3.3.1 AGE PREDICTION

Facelytics or analytics of face, is the process of extracting physical features of a face, to determine characteristic features like gender and age. Using deep learning architecture CNN model Caffenet, which has major layers such as Convolution, Relu, SoftMax, Pooling, LRN, Inner Product of varying kernel size and stride. Caffenet model assigns weights to the region of interest in the captured facial image to perform feature extraction. Following the same, it is passed as input to another pre-trained caffenet convolution model which contains multiple classes of ages, defined in prototxt which are predicted by the model with a sigmoid value.

3.3.2 SENTIMENT ANALYSIS

Sentiment analysis is a subset of text analysis where each text is classified as positive, negative or neutral based on the toxicity of the text. Before analysis, the text is preprocessed to remove noise, summarized if the text is large. Then parts of speech tagging used to identify structural words. Following this, toxicity is determined by assigning weights to each word and determining the weighted average of the whole text. Thus, the text is classified.

3.3.3 VIDEO CLASSIFICATION

Every uploaded video is classified as NSFW (Not Safe for Work) or SFW (Safe for Work) using Yahoo's Open Source NSFW TensorFlow model. It is performed on multiple stages, starting with accessing multiple frames from the video to perform classification on content type from the image. The model uses a general purpose caffe deep neural network. The model is fed in with an input image and receives a value(probability) to detect NSFW content. The value generated is safe to view if it is below 0.2, and is considered neutral up to 0.8, above which it is considered NSFW.

Additionally, audio is also extracted from the video and transcribed. The text obtained after transcribing is analyzed using parallel dots API to check for offensive language.

Thus, the scores generated from both the modules (audio-text, video-image) is aggregated to classify the video overall as NSFW or SFW.

4. RESULTS

The confusion matrix generated to represent ambiguity among different classes of age for a particular face below:

	0 -2	4 -6	8 - 12	15 - 20	25 - 32	38 - 43	48 - 53	60 +
0 -2	69.9	14.7	2.8	0.6	0.5	0.8	0.7	0.9
4 -6	25.6	57.3	16.6	2.3	1	1.1	1	0.5
8 - 12	2.7	22.3	55.2	15	9.1	6.8	5.5	6.1
15 - 20	0.3	1.9	8.1	51	10.6	5.5	4.9	2.8
25 - 32	0.6	2.9	13.8	23.9	61.3	29.3	26	10.8
38 - 43	0.4	0.7	2.3	5.8	14.9	46.1	14.6	26.8
48 - 53	0.2	0.1	0.4	0.7	1.7	5.3	33.9	16.5
60 +	0.1	0.1	0.8	0.7	0.9	5	13.4	35.7

Fig.2 Confusion matrix

The model accuracy in classifying the age group from the face of a user is 51.3%. The model accuracy on the concentrated class group [0-2, 4-6, 8-12] is 69.8%.

5. CONCLUSIONS

The access to age-restricted content for the kids can be effectively prevented by employing a combination of automatic video-classification based on frames of the videos, age-prediction by active capturing of facial image of the user and sentiment analysis to restrict and block toxic comments on a video-streaming platform. These methods ensure that the kids have access to only useful and appropriate content. The effectiveness of the technique is established based on the accuracy of the individual models.

The model can be improved by providing biometric age prediction for the compatible devices to achieve higher accuracy of age prediction. Further, the model can be trained to completely censor NSWF scenes in the videos, rather than blocking the video fully.

ACKNOWLEDGEMENT

It gives us immense pleasure to write an acknowledgement to this project, a contribution of all the people who helped to realize it. We extend our deep regards to Dr. S.B. Kivade, Honourable Principal of JSS Science and Technology University, for providing an excellent environment for our education and his encouragement throughout our stay in college. We would like to convey our heartfelt thanks to our HOD, Dr. M.P. Pushpalatha, for giving us the opportunity to embark on this topic. We would like to thank our project guide, Prof. Shalini K C for their invaluable guidance and enthusiastic assistance and for providing us support and constructive suggestions for the betterment of the project, without which this project would not have been possible. We appreciate the timely help and kind cooperation of our lecturers, other staff members of the department and our seniors, with whom we have come up all the way during our project work without whose support this project would not have been a success. Finally, we would like to thank our friends for providing numerous insightful suggestions. We also convey our sincere thanks to all those who have contributed to this learning opportunity at every step of this project.

REFERENCES

[1] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini and Michael Sirivianos, "Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children", Published at the 14th International Conference on Web and social media (ICWSM 2020) R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[2] Alice OTHMANI, Abdul Rahman TALEB, Hazem ABDELKAWY, Abdenour HADID, "Age estimation from faces using deep learning: a comparative analysis", Published in 2020 by Elsevier.

[3] Ashwin Geet D'Sa, Irina Illina, Dominique Fohr, "Towards Non-Toxic Landscapes: Automatic Toxic Comment Detection Using DNN".

[4] Adult Content Detection in Videos with Convolutional and Recurrent Neural Networks Jˆonatas Wehrmann, Gabriel S. Sim˜oes, Rodrigo C. Barros

BIOGRAPHIES

Prof. Shalini K C,

Professor at JSS Science and Technology University, Dept of Computer Science Engineering.

Rachana P Bennur,

Student of JSS Science and Technology University, Dept of Computer Science Engineering.

Samruddhi C Shetty,

Student of JSS Science and Technology University, Dept of Computer Science Engineering.

Anagha P,

Student of JSS Science and Technology University, Dept of Computer Science Engineering.

Hema G R,

Student of JSS Science and Technology University, Dept of Computer Science Engineering.