International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Survey Paper on Content Moderation

Mrs. Sonali Sonawane,
Assistant Professor,
G.H.Raisoni Institute of Engineering and technology,
Pune,India

Ashish Raj,
Student,
G.H.Raisoni Institute of Engineering and technology,
Pune,India

Hardik Ghuge,
Student,
G.H.Raisoni Institute of Engineering and technology,
Pune,India

Nikhil Agrawal,
Student,
G.H.Raisoni Institute of Engineering and technology,
Pune,India

Vikas Kumar Yadav,
Student,
G.H.Raisoni Institute of Engineering and technology,
Pune,India

Sana Aland,
Student,
G.H.Raisoni Institute of Engineering and technology,
Pune,India

***Abstract*: With the growth and accessibility of the internet to everyone, recognizing pornographic images is of great significance for protecting children's physical and mental health. It's common for children to surf the internet and they are just one click away from getting access to pornographic images. It is the desire of all parents to protect their children from online pornography, cyberbullying, and cyber predators. Several existing methods analyze a limited amount of information from a child's interaction with the respective online section. Some restrict access to sites based on blacklists known as banned URLs, others attempt to scan and analyze media content being exchanged between two parties. However, new URLs can be used to bypass blacklists, and images, videos, and text appear to be safe individually but should be evaluated together. Due to the vast size of social media-generated content, these approaches are insufficiently accurate. Furthermore, the approaches are not discriminative enough on a variety of image properties. The Colour, shadow, and frequency features of the images can vary, even the context is the same according to lumination features. The problem can be solved more accurately with deep learning techniques. Notably, the specific type of deep learning architecture called convolutional neural network is suitable for the problem space. Having a potential solution becomes even more important than detecting and recommending NSFW (Not Safe For Work) content that may be present to the user.**

Keywords: adult content, deep learning, NSFW, cyberbullying, pornographic, convolutional neural network

I. INTRODUCTION

In the era of the internet and increasing demand for technology, recognizing obscene images is very important for protecting children's physical and psychological health. Nowadays many of our children have access to mobile phones for educational purposes and surfing through the internet and our one clicks away from getting

exposed to pornographic or obscene content. Thus it has become very crucial that we protect our children from such harmful content. Due to covid restrictions, children cannot meet with their friends or family members and are limited to communicating via calling platforms or social networking websites. This online life exposes the child to various types of potentially dangerous interactions with unknown parties. Many can, in extreme cases, affect mental and physical safety. Protecting children online means minimizing exposure to online threats such as cyberbullying, pornography, cyber grooming, and self-harm, and providing tools to protect yourself from online threats. For this reason, we are conducting research on various methods and technologies that promote or support content moderation [5].

Today, people are more and more privacy Oriented, they hesitate to upload their personal photos along with some potential NSFW images on the server. Not safe for work(NSFW) is defined as links to content, videos, or website pages the public may not wish to see, formally or controlled environment. In a survey of about 210 people conducted, more than 190 people were interested in the automatic tagging feature on their phone that can detect and allow private photos to move to a safe folder. Private are both pornographic and semi-nude images Examples: bikinis, lingerie, swimwear, etc. This feature should work locally without uploading information to the server.[2]

We investigated whether sufficient performance can be achieved by utilizing solely geometric features of the torso rather than only facial traits in this study. For this paper, Haar filters were used with an SVM classifier to extract features, and then an SVM classifier was used to classify the age group and gender. The outcomes of the acquired results were compared to those of a CNN algorithm [4]. The goal of this research is to classify adult photographs. The above-mentioned deep learning techniques were applied in this situation. In image-based problem spaces, convolutional neural network design in the context of deep learning approaches is showing promising results. In this case, the state-of-the-art convolutional neural network Efficient-Net was utilized using transfer learning [3].

II. RELATED WORK

The above paper focuses on adult content detection based on the percentage of skin exposed in the given images. It uses an approach that completely relies on machine learning. As mentioned the proposed system classifies images into pornographic and non-pornographic based on the amount of skin being exposed in those images. The system uses an SVM algorithm for this classification [1]. This paper presents an on-device solution for detecting NSFW(not safe for work) images. It not only detects complete nude images but also includes semi-nude images content moderation. For this, they have curated a dataset comprising 3 major categories – nude, semi-nude, and safe images. The paper proposes an ensemble that consists of a single shot multibox detector(SSD) as a feature detector with MobileNetV3 as a classifier [2].

This paper explains their framework which identifies objectionable content be it images, videos, audio, or text. The framework is designed to detect explicit images, cyberbullying images and texts, inappropriate audio content, and online sexual grooming. The CASPER (Children's agent for Secure and Privacy Enhanced Reaction) has 3 modules which are audio, video and image, text processing [5].

The given paper focuses on the identification of explicit Twitter or linked website messages that may promote illegal services and exploit minors, by using natural language processing. As mentioned the work has two phases. In the first stage, NLP techniques are used in order to identify messages on Twitter that promote illicit services provided by minors. In the second phase, from the websites categorized as suspects, images are extracted in order to perform image processing and gender recognition of two age

groups[4]. An image-based classifier is used for adult content detection. The classifier is based on convolutional neural networks which is a kind of deep learning method which got improved with transfer learning methods. In this, we measured the effect of different image processing methods on the results and determined how we can create improved classifiers for an adult content domain. Methods used are Image processing, Deep Learning, and Histogram [3].

III. DATA OVERVIEW

The dataset Provided for the system is of non-linear nature therefore the system uses non-linear SVM. The model was initially trained with 4K pornographic images and 4K normal images. Some images cannot be differentiated from pornographic or not [1].

The MobileNetV3 is pretrained with the ImageNet Dataset COCO dataset for body part detection.NPDI dataset consists of around 80 hours of videos 40 pornographic and 40 non-pornographic videos. All these videos are segmented into a total of 16,727 images out of which images are pornographic, 333 are hard-non pornographic images and the remaining 333 images are easy nonpornographic images. The authors produced their own customized dataset NSFW_16k dataset (total number of images is 16142 out of which 10933 are SFW and 5209 NSFW) [2].

Scraping is the process used for websites that have a massive download of explicit content and also data cleaning is done. Geometric feature extraction is used to detect faces and the upper body of the collection from suspicious websites. Image Classification is done via two different classification processes. Support Vector Machine-SVM is used to predict two classes like yes or no. The convolutional neural network is a supervised machine learning model that requires a big image dataset to build a classification model after some iterations [4]. Around 32,000 photos have been collected in total. Data should be validated and cleansed by hand after the gathering phase. The dataset was reduced to 27, 673 after human supervision. All of the classes have lost 5% of their former volume. Dimensional preprocessing of collected data is required. The data was scaled to 224 by 224 pixels with three RGB channels after the procedure [3].

This paper presents an on-device solution for detecting NSFW(not safe for work) images. It not only detects complete nude images but also includes semi-nude images content moderation. For this, they have curated a dataset comprising 3 major categories – nude, semi-nude, and safe images. The paper proposes an ensemble that consists of a single shot multibox detector(SSD) as a feature detector with MobileNetV3 as a classifier. The detector provides a human localization portion which is catered to the classifier for further predictions. Initially, the model was also tested with MobileNetV2 but found the best performance with MobileNetV3.The MobileNetV3 is pretrained with the ImageNet Dataset COCO dataset for body part detection. For training SSD COCO dataset is used and thus images are RESIZED to

IV. METHODOLOGY

In the SVM algorithm, the hyperplane separates data points into two classes. The dataset Provided for the system is of non-linear nature therefore the system uses non-linear SVM and kernel function. The kernel function converts low dimensional feature to high dimensional feature i.e. non-separable data to separable data so that the data could be classified by SVM. The model was initially trained with 4K pornographic images and 4K normal images. Though some images cannot be differentiated into pornographic or not, in such cases non-linear SVM and kernel function comes into role. Once the image is loaded to the model and classified as unsafe or obscene, the unsafe area is colored black [1].

300x300. For training, the classifier images are resized to 224X224.SSD with MobileNetV3 provides a faster model with a low execution time. Thus SSD with MobileNetV3 is used as a feature extractor for a lightweight detector [2]. The framework is designed to detect explicit images, cyberbullying images and texts, inappropriate audio content, and online sexual grooming. The CASPER (Children's agent for Secure and Privacy Enhanced Reaction) has 3 modules. The first module is the audio module – Kaldi is used for audio transcription as it controls all parts of speech-to-text conversion and easily adapts to different noisy environments by integrating different acoustic modeling scripts at the Operating System level. The second module is the image processing module.CASPER uses screenshots taken to detect image or video content and feeds it to the developed HCI tracker software. The images are free to screen the region classifier which finds the region for classification. After classification using NSFW Data Scraper and MobileNetV3 if the content is offensive, it is censored or blocked on the screen. The third module is the text processing module. The text from the messages or images or videos is extracted using XLM-R which can process up to 100 different languages. The xlm-Roberta model is pre-trained with 10 databases of cyberbullying detection. As soon as cyberbullying is detected the system sends an email to the child's parents containing the detected paragraph [5].

The work has two phases. In the first stage, NLP techniques are used in order to identify messages on Twitter that promote illicit services provided by minors. In the second phase, from the websites categorized as suspects, images are extracted in order to perform image processing and gender recognition of two age groups i.e. under 14 years and above 14 years. There are several processes that are used for the working of the proposed system in two phases. For the first phase tweet extraction, processing and classification are used to determine whether there are signs of human trafficking. Firstly there is a Harvesting process in which data for each harvested day is stored on a JSON file that has information regarding the tweet post. Secondly Cleaning and preprocessing are used in which all hashtags and messages mentioned in a given table are stored locally in a JSON file. The information is processed and cleaned using a Python application, and tweets are deleted. Thirdly there is Text Normalisation in which incorrectly spelled words are detected, words with repetition of characters, detect the correct context of the words and foreign languages. Lastly, a semi-supervised learning technique with Naive Bayes and SVM algorithms is used in order to classify the tweets as "suspicious" or "not suspicious" of being related to sex trafficking. For the second phase image extraction and processing to determine if the image is a person under 14 years old or not. Scraping is the process used for websites that have a massive download of explicit content and also data cleaning is done. Geometric feature extraction is used to detect faces and the upper body of the collection from suspicious websites. Image Classification is done via two different classification processes. Support Vector Machine-SVM is used to predict two classes like yes or no. A convolutional neural network is a supervised machine learning model that requires a big image dataset to build a classification model after some iterations. Results for both SVM and CNN compared with experiments depend on the quality of the image provided.SVM had better accuracy in age, and gender classification [4].

The image-based classifier is used for adult content detection. The classifier is based on convolutional neural networks which is a kind of deep learning method which got improved with transfer learning methods. In this, we measured the effect of different image processing methods on the results and determined how we can create improved classifiers for the adult content domain. Methods used: A) Image processing: Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal

processing in which input is an image and output may be an image or characteristics/features associated with that image. B) Deep Learning: Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised. Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to andIn some circumstances, it outperforms human experts. C) Histogram Equalization: Histogram Equalization is a computer image processing technique used to improve contrast in images. It achieves this by effectively distributing the most common intensity values., i.e. stretching out the intensity range of the image. In this study, we found that the color and edge feature of the image is discriminative for the problem space of adult content classification. Although, applying of histogram equalization technique to the color has a negative effect on model training. The main reason behind this is that Histogram Equalization equalizes the color distribution because reducing the color differences that are important to the classifier will bring the images between the different classes closer together. Standalone sharpness enhancement has given the most accurate results. According to this observation, the sharpness of the image is especially important for the problem domain. With the integration of the model we have presented here to a social platform, any content uploaded by the user can be filtered and automatically deleted if it contains adult content. Besides, filtering can be provided in the content of the videos by taking the frames from the uploaded videos at certain intervals and classifying them with this model. For future works, different image sharpness enhancement methods can be tested with edge enhancement methods to improve the accuracy of the model [3].

V. RESULT

This model gave an accuracy of 94% on the training dataset and 91% while testing the model. It also gave 4% false-positive results [1]. The model has achieved an F1 score of 0.91 with 95% accuracy and 88% recall on the customized dataset produced by the authors ie the NSFW_16k dataset (total number of images is 16142 out of which 10933 are SFW and 5209 NSFW) and 0.92 MAP on NPDI dataset. It has also achieved an average 0.002 false-positive rate. The average inference time of the proposed system is 85ms for one image(60ms by body detector,25ms by classifier [2]. The CASPER demonstrates an average accuracy of 88% and an F1 score of 0.85 when classifying text and 95% accuracy when classifying pornography [5]. When only upper body features in the photos are evaluated, the results show that the SVM model has a greater classification accuracy than the CNN model. On the one hand, SVM's gender and age group classification accuracy values are 81.6 % and 82.1 %, respectively. CNN's gender and age group classification accuracy values, on the other hand, are 64.2 % and 51.3 %, respectively [4].

	Train Accuracy	Test Accuracy	Precision	Recall
Base Model	0.923	0.875	0.878	0.875
Histogram Equalization	0.922	0.860	0.867	0.860
Colour Enhancement	0.927	0.886	0.888	0.886
Sharpness Enhancement	0.950	**0.937**	**0.943**	**0.937**
Colour & Sharpness En.	0.963	0.908	0.911	0.908
All of three	0.941	0.873	0.876	0.874

Result Table[3]

VI. RESULT TABLE

Paper No.	Technique	Advantages	Disadvantages	Accuracy
[1]	Support Vector Machine Algorithm	It can be used to protect children from getting exposed to widely available adult content on internet	Time taken to blur the images is more.	94% on training dataset. 91% on testing dataset. 4% false positive
[2]	Deep Learning, Single Shot Multibox Detector (SSD), MobileNetV3	Average interface time for mobile is 85 sec for image	Interface time increases as we send more images consecutively	88% on NSFW_16K dataset
[3]	Deep Learning-CNN, Histogram Equalization technique	Average F1 score recorded is most efficient among all the models i.e., 0.95	The efficiency decreases as up to 0.92 when more than 180 images send at a time	The training accuracy is 0.941, test accuracy is 0.873 and precision is 0.876
[4]	HAAR face detection algorithm, LaGrange's optimization Parameters, Confusion matrix for gender, SVM, CNN	The results of the paper can be used for human trafficking, disappearance, kidnapping and among others.	1.Study of some characteristic related to ethic and racial features. 2.Different video formats are not supported 3.Detection of medical issues.	SVM's gender and age group classification accuracy values are 81.6 % and 82.1 %, respectively. CNN's gender and age group classification accuracy values, on the other hand, are 64.2 % and 51.3 %, respectively
[5]	Deep learning specifically GANs, Tesseract-optical character recognition (OCR) for deriving semantic text meaning, CNN, k-NN classifier for audio	We have shown that comprehensive, modular method of our framework is identifying objectionable online content. It performs well given limited hardware resources and a limited number of categories of harmful content to detect.	Incorporating online grooming, and self-harm detection modules is the main focus of our future work. These new Resources modules present a challenge in terms of accuracy of detection, especially for online grooming as this is more lengthy attacking process that requires knowledge of previous conversations between parties.	The CASPER demonstrates an average accuracy of 88% and an F1 score of 0.85 when classifying text and 95% accuracy when classifying pornography

VI.CONCLUSION

Adult content detection models are critical for both kids and adults to safeguard their privacy and security while surfing the web, as no one wants a random advertisement popping up and presenting undesirable and unwelcome information at any time. So, we can boost efficiency even more by minimizing the blurred time of an image from the time it first appears on the screen [1]. With the help of the application of the paradigm, we've described here a social situation. Any content uploaded by the user can be screened by the platform. If it contains explicit content, it will be automatically destroyed. Besides, By taking into account the content of the movies, filtering can be provided. the frames from the videos that have been uploaded at regular intervals and This methodology can be used to categorize them [5].

We offer a deep learning technique for detecting nudity and semi-nudity contents in this paper. For training and evaluation of our approach, we created a dataset encompassing a wide mix of acceptable and hazardous photos, including semi-nudity images. We presented a deep learning ensemble with a MobileNetV3 classifier and an SSD with a MobileNetV3 feature extractor for this purpose. SSD detects potentially dangerous body components while also supplying a human-located portion of the image, which is then supplied to the classifier for classification [2]. The best option to detect a possible case of human

trafficking of minors is using the SVM algorithm. The picture classification is based on the upper body to forecast the age group to detect human trafficking is the key contribution in this work [4].

VII. REFERENCES

[1] Gajula, Ganesh; Hundiwale, Ajinkya; Mujumdar, Shreyas; Saritha, L.R (2020). 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom) - A Machine Learning-Based Adult Content Detection Using Support Vector Machine.

[2] Anchal Pandey, Sukumar Mohrana, Debi Prasad Mohanty, Archit Panwar, Dewang Agarwal and Siwa Prasad Thota. 2021 International Joint Conference on Neural Networks (IJCNN) - On-Device Content Moderation

[3] Kalkan, Soner Can; Gozutok, Burak; Nahas, Abdullah Al; Kulunk, Aysenur; Erdinc, Hakki Yagiz (2020). *2020* International Conference on Innovations in Intelligent Systems and Applications (INISTA) - Image Enhancement Effects On Adult Content Classification.

[4] Yue, Hongzhou; He, Shuilong; Liu, Zhenghui (2020). Social Media Users Send Promotional Links to Strangers: Legitimate Promotion or Security Vulnerability

[5] Granizo, Sergio L.; Hernandez-Alvarez, Myriam; Lopez, Lorena Isabel Barona; Caraguay, Angel Leonardo Valdivieso (2020). Detection of Possible Illicit Messages Using Natural Language Processing and Computer Vision on Twitter and Linked Websites. IEEE Access.

[6] Aleksandar Jevremovic; Mladen Veinovic; Milan Cabarkapa; Marko Krstic; Ivan Chorbev; Ivica Dimitrovski; Nuno Garcia; Nuno Pombo; Milos Stojme(2021).Keeping Children Safe Online With Limited Resources: Analyzing What is Seen and Heard.IEEE Access.