A Novel Approach for Image based Cyberbullying Detection and Prevention

Vijayakumar V., PhD
Professor,
Department of Computer Science,
Sri Ramakrishna College of Arts and Science,
Coimbatore, Tamilnadu, India

Hari Prasad D., PhD
Professor,
Department of Computer Applications,
Sri Ramakrishna College of Arts and Science,
Coimbatore-Tamilnadu, India

Adolf P.
Research Assistant,
Sri Ramakrishna College of Arts and Science,
Coimbatore Tamilnadu, India

ABSTRACT
Now a day's social media offer great communication opportunities and also it increases the addiction and raises vulnerability of young people to threatening situations online. The cyberbullying is one of the most affected cybercrime that has gained more attention in recent years. It is an aggressive, intentional behaviour that is carried out by a group or individual, repeatedly and over time against a victim. Most of the current works have focused on detecting cyberbullying based on textual information and very few are based on image cyberbullying. Due to inexpensive of the internet cost, image-based bullying is growing. To detect and prevent, many works have been done by researchers with novel technologies. Chatbot is a really valuable tool to help prevent cyberbullying which simplify the interaction between humans and computers. This paper proposed a Convolutional Neural Network (CNN) deep model to predict the cyber bullying. If a cyber bullying image is detected, it gives the suitable alert messages to the users, parents and caretaker through chatbot interface. The experiments are conducted on publicly available social media datasets and tested with telegram chat communication. This paper aims to direct future research on integrating multimodal data sources such as text and images to prevent cyberbullying issues.

General Terms
Cyber Safety

Keywords
Cyber bullying Image detection, Image based deep learning Models, Cyberbullying prevention

1. INTRODUCTION
With the increasing demand of internet and social media usage, a new form of bullying has emerged. Cyberbullies use these technologies to harass, threaten, or humiliate which can occur anywhere, via any devices like smartphones, and any mode like emails, texts, and social media [14]. The method of cyberbullying can vary from text to images and videos by socializing untrue rumors and disclosure of personal information that are directed to harm and dishonour the victims [15]. People who are cyberbullied will be addicted with use alcohol and drugs, skip school or college, unwilling to do regular work, have lower self-esteem and have more health problems. Due to increase online learning, there is no simple solution to stop cyberbullying, and no fool proof way to handle a bully. One of simple solution is to prevent communication from the cyberbully, by blocking their email address, cell phone number, and deleting them from social media contacts. Cyberbullying Detection implements Natural Language Processing (NLP) and Machine Learning algorithms in finding a negative comment from the messages in text which are received by a user [1].

Most of these studies have mainly focused on analysing textual content, such as comments and text messages for cyber bullying detection. Nowadays, Social networking websites such as Instagram, Telegram, Flickr and Pinterest focus exclusively on image-sharing. These trends make a shift from text-based cyberbullying to cyberbullying content that makes use of image to perpetrate cyberbullying behaviours among victims. So, it is important to systematically investigate cyberbullying in images and understand its factors, based on which automatic detection approaches can be formulated. Cyberbullying images can be identified based on detection techniques that use visual cyberbullying factors such as body-pose, facial emotion, object, gesture, and social factors. The detection of cyberbullying in images is a far more challenging task due to social media data is short, noisy and unstructured [2][3].

As there is evidence of very strong usage of social media by people between the ages of 18 and 29, currently becomes the focus of investigation of additional techniques and strategies to assist students to prevent cyberbullying [16]. The main objective of this paper is to leverage innovative technology to detect and prevent the cyberbullying. Sometimes, the social media user might not report occurrences of cyberbullying to their parents. So, the effective user interface technology is needed to communicate and notify the user to prevent the cyberbullying events. In this paper, a deep learning based system is developed to detect cyberbullying in images with Convolutional Neural Network (CNN). Also telegram notifications are added to Telegram user and automatically reply to messages to prevent the cyberbullying.

2. REVIEW OF LITERATURE
The negative effects of cyberbullying are abnormally growing and a large number of studies have been dedicated to detect and prevent the cyberbullying. Most of the technical studies concentrate on text analysis. There are three main directions in detecting cyberbullying events such as Natural Language Processing, Machine Learning and Deep Learning. Traditional text analysis methods cannot fulfil the variety of bullying data in social networks. N. Tahmasbi et al. [4] discussed the parental monitoring software that is used for prediction and prevention of cyberbullying. Machine Learning and deep learning algorithms help researchers to understand data and provide an opportunity to effectively predict and

detect cyberbullying in social media communication. Machine learning approaches like Decision Tree (DT), Random Forest, Support Vector Machine, and Naïve Bayes are used to detect the bullying message and text [17].

Deep learning has recently attracted the attention of many researchers in different fields such as Natural Language understanding. Now, deep neural based models have shown substantial enhancement over traditional models in detecting cyberbullying. Deep learning architectures are useful in various detection and preventive tasks. S. V. Georgeakapoulos et al. [6], M. Dadvar et al [9] and M. A. Al-Ajlan et al. [8]., used the Deep learning based convolutional neural network (CNN) for text based cyberbullying detection. It is mostly used for image recognition, Electromyography (EMG) recognition, Video analysis, Natural Language Processing, Anomaly Detection, Drug discovery, Health risk assessment and biomarkers of aging discovery, Time series forecasting etc. Y. S. M. V. Tanmayee Patange et al. proposed a deep learning method of Cyberbullying detection from social media [5]. A. Gangwar et al. used NSFW dataset to detect abuse content [10]. The CNN and NSFW data set is used by Q. H. Nguyen et al. for detection of pornographic content [11]. Nishant Vishwamitra et al., discussed classifier models and a multimodal classifier to detect visual cyberbullying based highly contextual visual factors such as body-pose, facial emotion, gesture, object and social factors [12]. K. Wang et al., proposed multi-modal detection framework that takes into image, video, comments, time (multi-modal information) on social networks [13]. Pradheep.T. et. al., proposed a multimodal cyberbullying detection approach from audio, video and image. The cyberbully image was detected using the computer vision algorithm and cyberbully video was detected using the Shot Boundary detection algorithm [7].

The detailed review of literature is presented in the Table 1.

Table. 1. Image based Cyberbullying detection

Authors	Year	Algorithm	Outcome	Limitations
Q. H. Nguyen, et. al., [11]	2020	Mask R-CNN	Cyberbullying Detection in Images and video	Not supported for Image Texts and writings
Mhd Wesam Al-Nabki et al., [19]	2020	CNN	Pornography Detection	Not supported for Texts and writings
N. Tahmasbi and A. Fuchsberger [4]	2019	NLP, Machine learning	Cyberbullying detection for helping parents	Targeted detection only
Tanmaye e Patange et al., [5]	2019	CNN, Word2 Vec, Offensivene ss	Multi input cyberbullying detection	Instagram only
S. V. Georgako poulos et al., [6]	2018	CNN	Cyberbullying Detection	Not supported for Multilingual & Multimodal
M. A. Al-Ajlan and M. Ykhlef [8]	2018	CNN	Cyberbullying Detection	Not supported for Multilingual & Multimodal
M. Dadvar and K. Eckert [9]	2018	CNN, LSTM, BLSTM and BLSTM with attention	Cyberbullying Detection	Not supported for Multilingual & Multimodal
KaiLong Zhou et al., [22]	2016	CNN	Pornography Detection	Targeted detection only
Seyed Mostafa et al., [21]	2014	Multi-Layer Perceptron	Pornography Detection	Old method
André Tabone et al., [20]	2020	Mobile Net	Pornography Detection	Text is not detected

There are still not many studies dedicated to automatically prevent cyberbullying but the number is increasing. Telegram is one of the biggest messaging apps in the world, with 400 million active monthly users. It gives a great opportunity to engage with users on a platform where they feel safe and comfortable. New generations of chatbots, will not only perform services but also engage in more complex and diverse conversations. Creating an intelligent Telegram chatbot that can automate the conversation. It can be created and customized easily. R.Cohen created a customized chatbot CyBully using Rebot.me to respond to most user inputs with insulting remarks [18]. Vijayakumar V and Hari Prasad D presented a deep learning based LSTM algorithm to detect and prevent the cyberbullying incident interfaced with chatbot [24]. A major challenge is alerting a user about the potentially sensitive comments and informing them about the cyberbullying event in the social media images. This paper proposed the image based cyberbullying classification, prediction and prevention model developed based on deep learning technique. The bot interface is integrated with the telegram to notify the bullying images.

3. PROPOSED METHOD: IMAGE BASED CYBERBULLYING DETECTION

The proposed image based cyberbullying detection with deep learning technique and notification based prevention system is developed as shown the Fig.1 and discussed in the following steps:

Step 1: Collect and Prepare Dataset

The proposed method collected the scrapped data from NSFW (Not Safe For Work) publicly available dataset and pre-process the data to avoid the possibility of images which contains wrongly tagged. The pre-process avoids model corruption. The size of the image is refined using Open-CV pre-processing function library. Each image is converted into an array. Neural networks process inputs uses small weight values, and inputs with large integer values disrupt or slow down the learning process. So, all pixels values divided by the largest pixel value 255. It will be converted into all values range between 0 to 1.

Step 2: Train a deep neural network model

The Convolutional Neural Network (CNN) with RELU activation is proposed to build a deep neural network which

effectively identify the cyberbullying images. CNN contains multiple layers of perceptron in a manner that each neuron in one layer is connected with all neurons in the next layer. All neurons apply convolution operation to the input in convolutional layer. A convolution involves the multiplication of a set of weights with the input. After convolution operation, an activation function Rectified Linear Unit (RLU) is applied which takes the output of max_pool. The function returns 0 if it receives any negative input, but for any positive value x, it returns that value back. So, it can be written as f(x)=max (0, x).

Steps:

1. Add Convolutional Layer, Kernel size = (5,5), activation function = ReLU and then Max Pooling two dimensional size = (2,2).
2. Add Convolutional Layer, Kernel size = (3,3), activation function = ReLU and then Max Pooling two dimensional size = (2,2)
3. Repeat STEP 2
4. Add Flatten layer
5. Add Dense layer with activation function "ReLU"
6. Repeat STEP 5
7. Add Dense layer = No. of output classes, activation function = softmax

The model is trained on the dataset with NSFW images as positive and SFW (Safe For Work) images as negative. The convolutional neural network is used to train a deep learning model. This model will get a clarity to the inputs on detection of bully content.

Step 3: Test the created model

The model is tested with Telegram user data.

Step 4: Deploy the model and prevention

The model is integrated with telegram plugin interface. If any cyberbullying event occurs, it notifies the user.

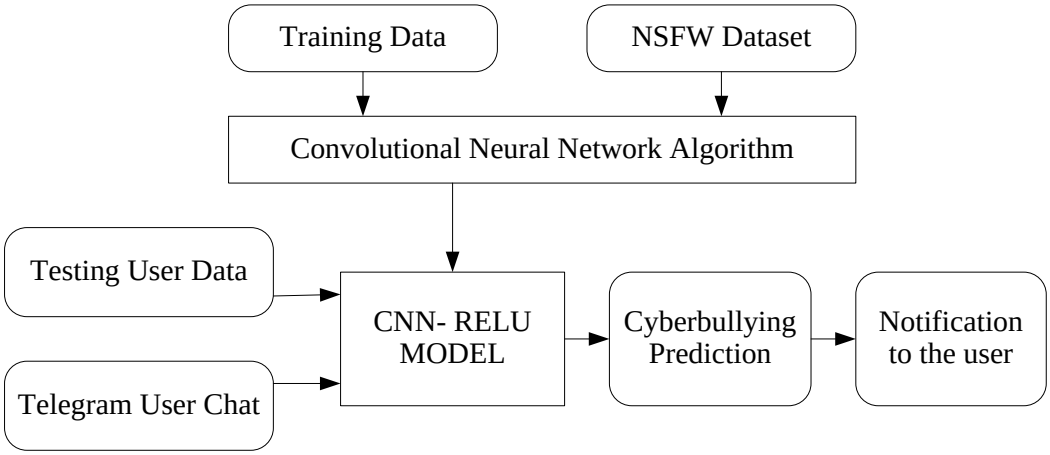

Fig 1: Image Based Cyberbullying detection

4. EXPERIMENTAL RESULTS & DISCUSSION

A system with Intel i5 processor and 8 GB RAM with internet speed of 1 Mbps is used for implementing and testing the proposed method. Google Colab environment with GPU support is used for program execution which helps for performing the deep learning algorithms very quickly. The python libraries cv2 and numpy are used for better performance. Here three different layers of convolutional neural network are used with RELU activation function. A total of 11,155,454 parameters are observed. Each parameter is uniquely important for detection. The CNN Model summary is shown in the Fig 2.

The CNN model is built as it gives output as two class NSFW and SFW. NSFW datasets are scrapped from GitHub collection by EBazarov [19] [23]. 1100 images which contains both NSFW and SFW content are chosen for this project. After filtering unnecessary images, 950 images were selected. It contains 475 NSFW images and 475 SFW images. The system detects NFSW AND SFW images using deep neural network CNN algorithm. Cyberbullying images are identified from the test images. The model also tested with individual images to detect cyberbullying.

The performance of the proposed framework was measured in terms of the quality measures, namely precision, F1 score and Recall. Precision is measured with the ratio between the true positive (correct predictions) and the total predictions. Recall is calculated with the proportion of the correct predictions and the total number of correct items in the set. F1 score is measured with the weighted harmonic mean of precision and recall. Accuracy calculates the proportion of correctly identified cyberbully words. The accuracy of the model is 82% working efficiency. The detail of the prediction performance is shown in the Fig. 3. Fig.4 shows the Predicted Output of testing samples.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 224, 224, 32)      2432
_________________________________________________________________
activation (Activation)      (None, 224, 224, 32)      0
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 112, 112, 32)      0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 110, 110, 64)      18496
_________________________________________________________________
activation_1 (Activation)    (None, 110, 110, 64)      0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 55, 55, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 53, 53, 64)        36928
_________________________________________________________________
activation_2 (Activation)    (None, 53, 53, 64)        0
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 26, 26, 64)        0
_________________________________________________________________
flatten (Flatten)            (None, 43264)             0
_________________________________________________________________
dense (Dense)                (None, 256)               11075840
_________________________________________________________________
activation_3 (Activation)    (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 84)                21588
_________________________________________________________________
activation_4 (Activation)    (None, 84)                0
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 170
=================================================================
Total params: 11,155,454
Trainable params: 11,155,454
Non-trainable params: 0
_________________________________________________________________
```

Fig. 2. Image based Prediction Model Summary

	precision	recall	f1-score	support
SFW	0.78	0.88	0.83	94
NSFW	0.87	0.75	0.80	96
accuracy			0.82	190
macro avg	0.82	0.82	0.82	190
weighted avg	0.82	0.82	0.82	190

Fig. 3. Prediction Accuracy

Telegram bots are a new popular feature allowing third-party apps to run within the platform. It enhances the messaging experience and can set up to carry out specific tasks. The telethon library is used in python to get access to all the messages of a user. After the user's permission, the images are observed.

The CNN Deep learning model listen the user. When any image is sent to the user, a copy sent to the deep learning model server. The server analyse the content. The results will be sent back to the user. The telegram image based communication, detection and notification is shown in the Fig.5.

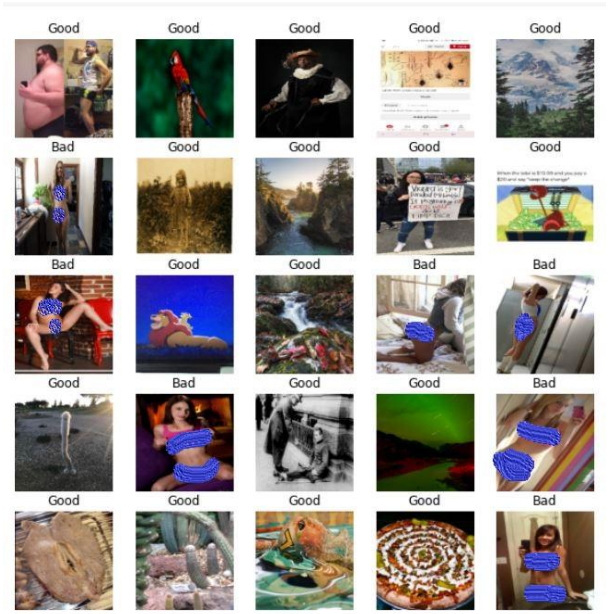

Fig 4. Predicted Output for Testing Samples

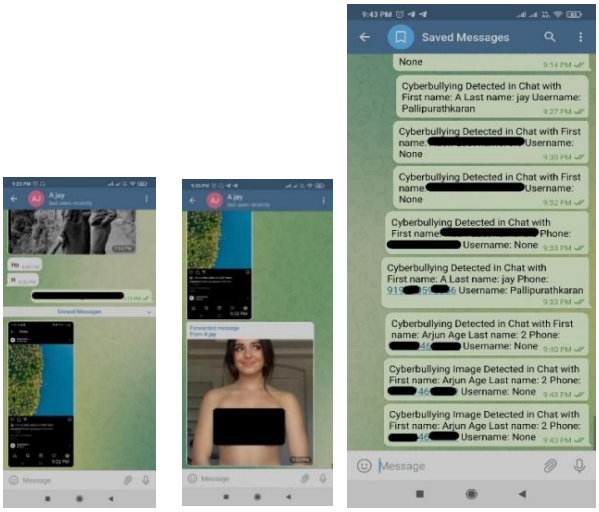

Fig5. Telegram Message Communication

If no bully content detected, it neglects the content. If bully content detected, the server will send a notification to the telegram user through telethon. The details of the bullying detection in the server is shown in the Fig. 6a and Fig.6b..

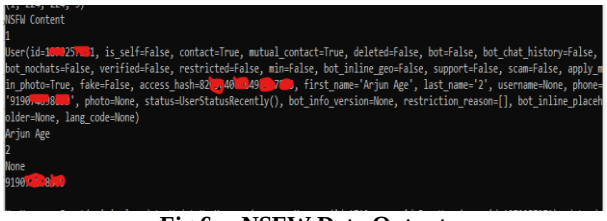

Fig 6.a. NSFW Data Outputs

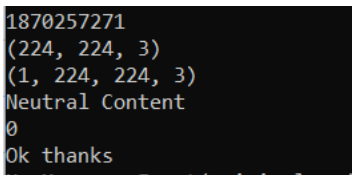

Fig 6.b. NSFW Data Outputs

5. CONCLUSION AND FUTURE WORK

The detection and prevention of cyberbullying is much needed due to the increasing social media users day by day. Most of systems uses Machine Learning and Natural Language Processing classification models to detect and reduce the cyberbullying. Recently Deep Neural Network Based (DNN) models have also been applied for detection of cyberbullying to improve the detection accuracy. The proposed system has a convolutional neural network model with Relu activation function to detect cyberbullying. The message communication interface system gives the suitable notification to prevent the bullying event. It shows a great advantage to the world, to fight against cyberbullying. In future, cyberbullying detection can be improving with image text detection and handwritten detection from cyberbullying images. It can also extended for multi-model cyberbully detection from image, text, video and other media.

6. ACKNOWLEDGMENTS

This research work was funded by the ICSSR under IMPRESS Scheme. The author also thanks the ICSSR and Sri Ramakrishna College of Arts and Science who provided the support for doing the research.

7. REFERENCES

[1] S. Salawu, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A Survey," IEEE Trans. Affect. Comput., vol. 11, no. 1, pp. 3–24, 2020, doi: 10.1109/TAFFC.2017.2761757.

[2] A. Singh and M. Kaur, "Content-based cybercrime detection: A concise review," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 8, pp. 1193–1207, 2019.

[3] H. Rosa et al., "Automatic cyberbullying detection: A systematic review," Comput. Human Behav., vol. 93, no. October 2018, pp. 333–345, 2019, doi: 10.1016/j.chb.2018.12.021.

[4] N. Tahmasbi and A. Fuchsberger, "ChatterShield – A multi-platform cyberbullying detection system for parents," 25th Am. Conf. Inf. Syst. AMCIS 2019, pp. 1–5, 2019.

[5] Y. S. M. V. Tanmayee Patange, Jigyasa Singh, Aishwarya Thorve, "DETECTION OF CYBERHECTORING ON INSTAGRAM," pp. 5–8, 2019.

[6] S. V. Georgakopoulos, A. G. Vrahatis, S. K. Tasoulis, and V. P. Plagianakos, "Convolutional neural networks for toxic comment classification," ACM Int. Conf. Proceeding Ser., no. April, 2018, doi: 10.1145/3200947.3208069.

[7] T. Pradheep, J. . Sheeba, T. Yogeshwaran, and S. Pradeep Devaneyan, "Automatic Multi Model Cyber Bullying Detection from Social Networks," SSRN Electron. J., pp. 248–254, 2018, doi: 10.2139/ssrn.3123710.

[8] M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbullying detection," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 9, pp. 199–205, 2018, doi: 10.14569/ijacsa.2018.090927.

[9] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," pp. 1–13, 2018, [Online]. Available: http://arxiv.org/abs/1812.08046.

[10] A. Gangwar, E. Fidalgo, E. Alegre, and V. González-Castro, "Pornography and child sexual abuse detection in image and video: A comparative evaluation," IET Semin. Dig., vol. 2017, no. 5, pp. 37–42, 2017, doi: 10.1049/ic.2017.0046.

[11] Q. H. Nguyen, K. N. K. Nguyen, H. L. Tran, T. T. Nguyen, D. D. Phan, and D. L. Vu, "Multi-level detector for pornographic content using CNN models," Proc. - 2020 RIVF Int. Conf. Comput. Commun. Technol. RIVF 2020, 2020, doi: 10.1109/RIVF48685.2020.9140734.

[12] Nishant Vishwamitra, Hongxin Hu, Feng Luo and Long Cheng, "Towards Understanding and Detecting Cyberbullying in Real-world Images ," In Proceedings of the 28th Network and Distributed System Security Symposium (NDSS 2021), February 21-24, 2021.

[13] K. Wang, Q. Xiong, C. Wu, M. Gao and Y. Yu, "Multi-modal cyberbullying detection on social networks," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206663.

[14] https://www.helpguide.org/articles/abuse/bullying-and-cyberbullying.htm#

[15] https://www.stopbullying.gov/cyberbullying/what-is-it

[16] https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

[17] Islam, Manowarul & Uddin, Md Ashraf & Islam, Linta & Akhter, Arnisha & Sharmin, Selina & Acharjee, Uzzal. (2021). Cyberbullying Detection on Social Networks Using Machine Learning Approaches. 10.1109/CSDE50874.2020.9411601.

[18] R.Cohen, N.Mathiarasu, R.Aarif, S.Ansari, D.Fraser, M.Hegde, J.Henderson, I.Kajic, A.Khan, Z.Liao, A.Mancisidor, S.Nagpal, A.Pham, A.Saini, J.Shen, H.Singh, C.Tavares and S.Thandra, "An education-based approach to aid in the prevention of cyberbullying," ACM Computers & Society, Volume 47, Issue 4, July 2018, 17–28.

[19] M. W. Al-Nabki, E. Fidalgo, R. A. Vasco-Carofilis, F. Jañez-Martino, and J. Velasco-Mata, "Evaluating Performance of an Adult Pornography Classifier for Child Sexual Abuse Detection," 2020, [Online]. Available: http://arxiv.org/abs/2005.08766.

[20] A. Tabone, A. Bonnici, S. Cristina, R. Farrugia, and K. Camilleri, "Private body part detection using deep learning," ICPRAM 2020 - Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods, pp. 205–211, 2020, doi: 10.5220/0009101502050211.

[21] S. M. Kia, H. Rahmani, R. Mortezaei, M. E. Moghaddam, and A. Namazi, "A Novel Scheme for Intelligent Recognition of Pornographic Images," 2014, [Online]. Available: http://arxiv.org/abs/1402.5792.

[22] K. Zhou, L. Zhuo, Z. Geng, J. Zhang, and X. G. Li, "Convolutional neural networks based pornographic image classification," Proc. - 2016 IEEE 2nd Int. Conf. Multimed. Big Data, BigMM 2016, pp. 206–209, 2016, doi: 10.1109/BigMM.2016.29.

[23] https://github.com/EBazarov

[24] Vijayakumar V and Hari Prasad D, "Intelligent Chatbot Development for Text based Cyberbullying Prevention," International Journal of New Innovations in Engineering and Technology, Volume 17, Issue 1, pp. 73-81, June 2021.