IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 2, February 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 2, February 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1002034 |

Content Moderation System Using Machine Learning

Pratap Thakur, Janhavi Patil, Ashwini Thorat, Avinash Ahire, Payal Jadhav, Prof.Vidya Jagtap

Student, Dept. of Information Technology Engineering, JSPM's Bhivarabai Sawant Institute of Technology & Research, Pune, Maharashtra, India

Assistant Professor, Dept. of Information Technology Engineering, JSPM's Bhivarabai Sawant Institute of Technology & Research, Pune, Maharashtra, India

ABSTRACT: With the rapid development of the Internet and smartphone technology, a large number of short videos are shared through social platforms. Therefore, video content analysis may be a vital and popular add machine learning currently. However, it's very difficult to research all aspects of video content originally produced by large-scale users. How to sort bad and illegal content from short videos published by an out-sized number of users, select high quality videos to share with other users, and improve the standard of video on the distribution platform of the whole user is a top priority. supported this background, this paper focuses on optimizing video auditing to supply basic features for algorithm judgment, supporting original content and increasing the distribution of latest content, and strengthening manual intervention which combines algorithm recommendation with manual recommendation

KEYWORDS: Remote auditing, sensitive information hiding, malicious manager preventing

I. INTRODUCTION

In the last 20 years ,online platforms that let users to interact and upload content for others to look at have become integral to the lives of the many people and have provided a benefit to society. However, there's growing awareness among the general public and businesses of the potential damage caused by harmful online material. User-generated content (UGC) adds to the variety and richness of content on the web, but it is not subject to the same editorial constraints as traditional media. This permits some users to post content which could harm others, particularly children. Examples of this type of content include content which is cruel and insensitive to others, which promotes terrorism or depicts maltreatment. Because the amount of UGC that platform users upload is growing at an exponential rate1, traditional human-led moderation systems are no longer able to discover and remove harmful content at the speed and scale required.

This project investigates the capabilities and applications of artificial intelligence (AI) in addressing the issues posed by online content moderation, as well as how advances are likely to reinforce such capabilities over the next five years. Nowadays, all the things are on smartphones, small children also use smartphones. Due to this it is necessary to filter illegal and inappropriate contents from the internet. A content management system, often abbreviated as CMS, is a software that helps users create, manage, and modify content on a website without the need for specialized technical knowledge. We are going to develop a system, which checks/audits all the content on the systems manually and as per the rules of the government of India. We will do operations on such content to filter out illegal content, and pass only legal content.

II. RELATED WORK

Effective moderation of harmful online content may be a challenging problem for several reasons. While many of these difficulties affect both human and automated moderation systems, AI-based automation systems face additional challenges.

There is a wide range of potentially hazardous content, including but not limited to: child abuse material, violent and extreme content, hate speech, graphic content, and sexual content, to name a few. Some dangerous content can be recognized simply by examining it, but other forms of content necessitate an awareness of the environment in which it exists in order to determine whether it is harmful or not. Interpreting this content consistently is challenging for human and automatic systems because it requires a broad and wide understanding of societal, cultural, and historical

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 2, February 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1002034 |

factors. Some of these contextual considerations vary round the world thanks to differences in national laws and what societies deem. acceptable. Content moderation processes therefore be contextually aware and culturally specific to be effective.

Online material can take several forms, some of which are more difficult to investigate and monitor, such as video and memes (which require a mixture of text and image analysis with contextual and cultural understanding). Deep fakes, which are created using machine learning to get fake but convincing images, video, audio or text, have the potential to be extremely harmful and are very difficult to detect by human or AI based methods.

In addition, content could also be posted as a live video stream or live text chat which must be analyzed and moderated in real time. this is often tougher because the extent of harmfulness can escalate quickly and only the previous and current elements of the content are available for consideration.

The vocabulary and structure of online information has rapidly evolved, and some users may choose to circumvent moderation systems by changing the terms and phrases they employ. Moderation systems must therefore adapt to stay pace with these constant changes.

To reduce the risk of exposing its users to hazardous information and to lessen the organization's reputation risk, online platforms may censor the third-party content that they host. Removing content that isn't commonly recognize to be detrimental, on the other hand, might affect a company's reputation and limit users' freedom of expression. Facebook, for example, has been criticized for removing an image of a statue of Neptune in Bologna, Italy for being sexually explicit

III. METHODOLOGY

The aim of the proposed system is to build a machine learning model that can filter out content which is not suitable for children. To make the web safer for children, many social media sites like Tumblr, Facebook, and Instagram are curbing NSFW (Not Suitable for Work) Content and making a safer community.

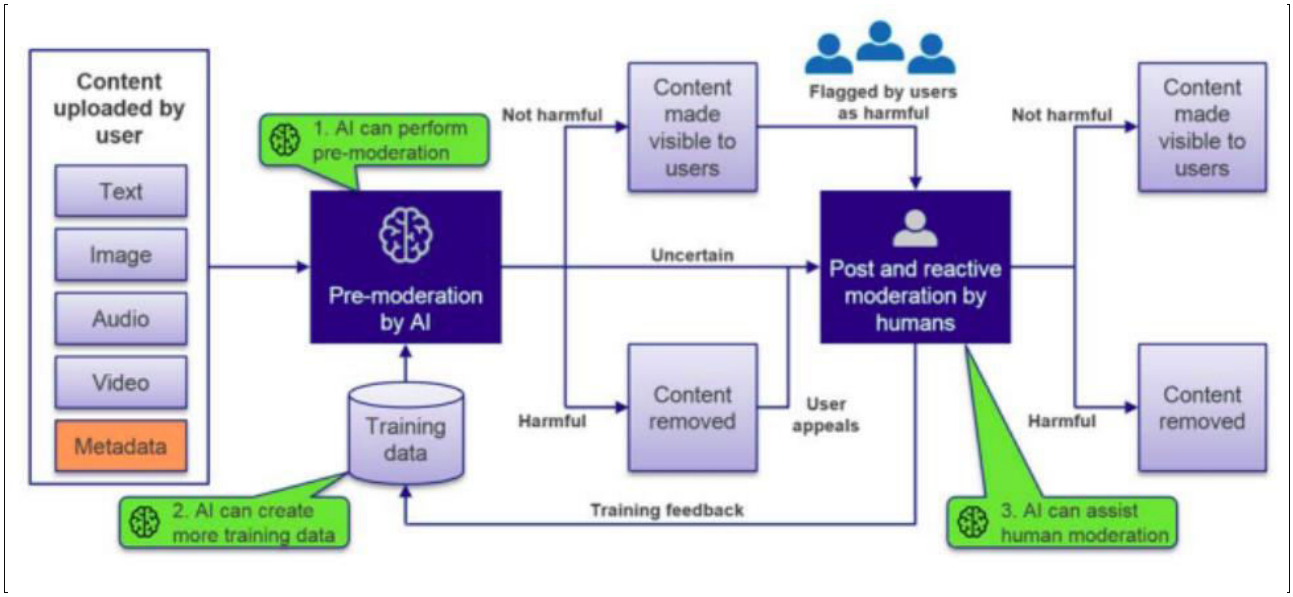


Fig.1 Methodology Of the system

There are 4 processes in proposed work:

1. GET DATA:

Write a set of scraping script to extract data from the internet, The code is simple and already comes with labeled data categories. This means that simply adopting the defaults of this data scraper will offer us 5 categories from hundreds of

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 2, February 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1002034 |

subreddits. Because most subreddits are slightly policed by humans to be on goal for that subreddit, Reddit is a terrific source of content from throughout the web. The instructions are quite simple, you'll simply run the 6 friendly scripts. Pay attention to them as you'll plan to change things up. If you want to add more subreddits, you need update the source URLs before proceeding to step 1.

2. LABEL AND CLEAN THE DATA

The information obtained from the NSFW data scraper has already been tagged! But expect some errors. Especially since Reddit isn't perfectly curated. Duplication is additionally quite common, but fixable without slow human comparison. The duplicate-file-finder, which is the fastest exact file match and deleter is the first item we prefer to execute. It's powered in Python. With this command, we can usually get rid of the majority of duplicates. This does not, however, catch photos that are 'basically' the same. For that, we advocate using a Macpaw tool called "Gemini 2".

3. USE KERA'S AND TRANSFER LEARNING

TensorFlow, Py-torch, and raw Python have all been considered as options for creating a machine learning model from the ground up. But We are not looking to discover something new, we want to effectively do something preexisting. After a little research, we chose Inception v3 weighted with ImageNet.

```
conv_base=InceptionV3(
 weights='imagenet',
 include_top=False,
 input_shape=(height,width,num_channels)
)
```

With the model in place, we added 3 more layers. A 256-neuron hidden layer is followed by a 128-neuron hidden layer, and finally a 5-neuron layer. The latter being the last word classification into the five final classes moderated by softmax.

```
#Add256
x=Dense(256,activation='relu',kernel_initializer=initializers.he_normal(seed=None),
kernel_regularizer=regularizers.l2(.0005))(x)
x=Dropout(0.5)(x)
#Add128
x=Dense(128,activation='relu',kernel_initializer=initializers.he_normal(seed=None))(x)
x=Dropout(0.25)(x)
#Add5
predictions=Dense(5, kernel_initializer="glorot_uniform", activation='softmax')(x)
```

We're employing dropout, which removes brain connections at random so that no single feature dominates the model. In addition, L2 regulation has been introduced to the first layer.. Now that the model is done, we augmented our data-set with some generated agitation. We rotated, shifted, cropped, sheared, zoomed, flipped, and channel shifted our training images. This helps with assuring the pictures are trained through common noise. All the above systems are meant to stop over-fitting the model on the training data. Even if it is a ton of data, we want to keep the model as generalize to new data as possible. We got around 87% accuracy on the model.

4. REFINE YOUR MODEL

Once the new layers are trained up, you'll unlock some deeper layers in your Inception model for retraining. As of the layer conv2d 56, the following code unlocks everything.

```
set_trainable=False
forlayerinconv_base.layers:
  iflayer.name=='conv2d_56':
     set_trainable=True
  ifset_trainable:
     layer.trainable=True
```

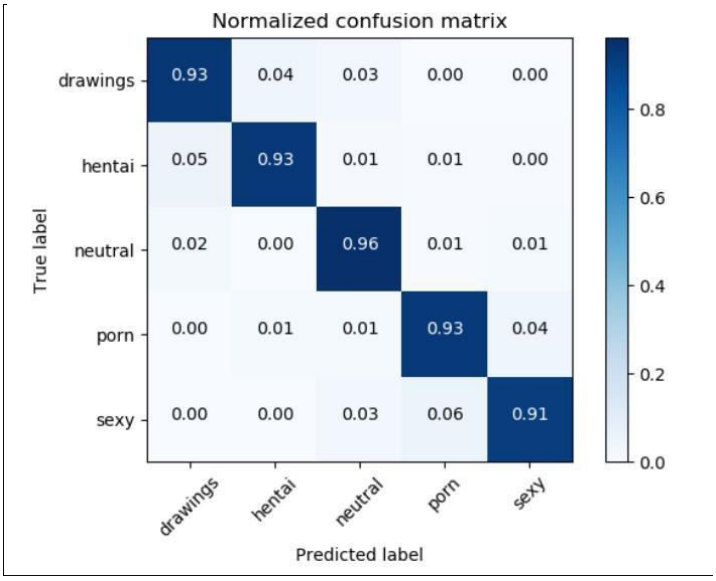
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 2, February 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1002034 |

```
else:
    layer.trainable = False
```

With these newly unlocked layers, we ran the model for a long time, and after adding exponential decay (via a planned learning rate), the model converged on a 91 percent accuracy on our test data. With 300,000 images, finding mistakes in the training data was impossible. But with a model with only 9% error, we could break down the errors by category, and then We could look at only around 5,400 images! Essentially, we could use the model to help us find misclassifications and clean the data set.



IV. SIMULATION RESULTS

- **MAIN GUI SNAPSHOT**

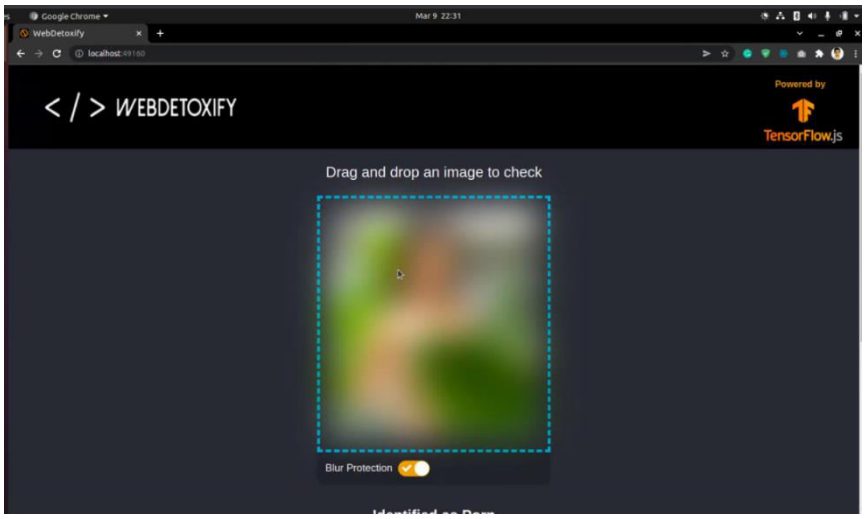
In the proposed system, we have to drag and drop the image/video into the drop box. Image/video then will be scanned and processed pixel by pixel by the CNN model.

After completion of the process, result will be shared on the screen in word or graph form. If it is detected that image/video is an adult content then that content gets 'blurred' to hide it from displaying on the screen.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 2, February 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1002034 |



V. CONCLUSION AND FUTURE WORK

We proposed a secure and efficient auditing scheme that supports sensitive information hiding and malicious manager prevention. It allows the file-owner to share data with sensitive information and guarantees the integrity and authenticity of shared data to be fully trusted by the owner and researchers. The proposed technique has an advantage over earlier work because of the innovative system model and mechanism for sensitive information concealing. Meanwhile, we gave the security analysis in detail to guarantee the robustness and soundness of our scheme

1. Automation Engine that will automatically delete inappropriate images. Set the parameters and keep manual labour to a minimal.
2. Admin Panel where all images in need of moderation are stacked up in a beautiful interface, which allows you to make decisions with just a click.
3. Launching google chrome extension.
4. It aims to expand the capacity of filtration of the content.

REFERENCES

1. S. Tavakoli, M. Shahid, K. Brunnström, B. Lövström and N. García, "Effect of content characteristics on quality of experience of adaptive streaming," .
2. F. Berečićc, M. Vranješ, D. Stefanović and Z. Kaprocki, "Design and Implementation of Live Video Analysis Information Logging Module,"
3. Y. Wu, P. Zheng, J. Guo, W. Zhang and J. Huang, "A Controllable Efficient Content Distribution Framework Based on Blockchain and ISODATA,"
4. S. Tavakoli, K. Brunnstrom, K. Wang, B. Andren, M. Shahid, and N. Garcia, “Subjective quality assessment of an adaptive video streaming model,”
5. N. Cranley and L. Murphy, “Incorporating User Perception in Adaptive Video Streaming Systems,”
6. “Subjective video quality assessment methods for multimedia applications,”
7. S. Tavakoli, J. Gutierrez, and N. Garcia, “Subjective quality study of adaptive streaming of monoscopic and stereoscopic video,”
8. A. Doan, N. England and T. Vitello, "Online Review Content Moderation Using Natural Language Processing and Machine Learning Methods : 2021 Systems and Information Engineering Design Symposium (SIEDS),"
9. A. Pandey, S. Moharana, D. P. Mohanty, A. Panwar, D. Agarwal and S. P. Thota, "On-Device Content Moderation,"
10. S Kühn and J Gallinat, "Brain structure and functional connectivity associated with pornography consumption: the brain on porn"
11. D. Ganguly, M. H. Mofrad and A. Kovashka, "Detecting Sexually Provocative Images", 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)
12. Y. Xu, B. Li, X. Xue and H. Lu, "Region-based pornographic image detection", IEEE 7th Workshop on Multimedia Signal Processing (MMSP)

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 2, February 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1002034 |

13. Q. Zhu, C.-T. Wu, K.-T. Cheng and Y.-L. Wu, "An adaptive skin model and its application to objectionable image filtering"
14. J.-S. Lee, Y.-M. Kuo and P.-C. Chung, "The adult image identification based on online sampling"
15. L. Duan, G. Cui, W. Gao and H. Zhang, "Adult image detection method base-on skin color model and support vector machine"

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details