OBSCENITY BLOCKER SOLUTION

Prof. Veena Bhat[1], Ankitha T O[2], Chaya D[3], Deepthi B[4], Deeptiman[5]

[1] *Assistant Professor, CSE, AMCEC, Karnataka, India*

[2] *Student, CSE, AMCEC, Karnataka, India*

[3] *Student, CSE, AMCEC, Karnataka, India*

[4] *Student, CSE, AMCEC, Karnataka, India*

[5] *Student, CSE, AMCEC, Karnataka, India*

ABSTRACT

The proliferation of obscene media across digital platforms poses significant challenges for maintaining a safe and appropriate online environment. In response to this pressing issue, we present a novel technological solution for identifying and blocking obscene media content at the user's end. Leveraging deep learning frameworks andtechniques, our solution employs ensemble learning, frame-by-frame video analysis, and web scraping to enhance accuracy and coverage of obscene content detection. The solution, implemented as a browser extension, automaticallyscans and classifies media content viewed in web browsers, sending alerts to the concerned nodal agency in case of obscene content detection. Our approach incorporates state-of-the-art models such as InceptionV3, OpenNsfw, and Nudenet, along with data augmentation and web scraping techniques to improve model performance and coverage. Through rigorous testing and evaluation, our solution demonstrates high accuracy and effectiveness in identifying andblocking obscene media, thereby contributing to the creation of a safer online environment.

Keywords: - *Obscene media detection, Content moderation, Deep learning, Ensemble learning, Web scraping, Data augmentation, Browser extension, Image classification, Video classification, Digital content filtering.*

1. INTRODUCTION

The internet is an ever-expanding landscape filled with diverse content ranging from informative to entertaining. However, amidst this vast array of information, there exists a darker side - obscene media. This includes adult images, videos, and other explicit content that can have harmful effects on users, especially children and adolescents. The accessibility of such content poses serious risks, including desensitization, mental health issues, and adverse impacts on social behaviour. Consequently, there is an urgent need for effective solutions to detect and mitigate the spread of obscene media in web browsers, safeguarding users from exposure to harmful content. To address the pressing issue of obscene media, this research proposes an innovative ensemble learning approach tailored specifically for web browsers. Ensemble learning, a powerful technique in machine learning, involves combining the predictions of multiple models to improve overall performance. By leveraging ensemble learning, we aim to enhance the accuracy and reliability of obscene media detection, thereby providing users with a more robust defense against harmful content. This approach represents a significant advancement in the field, offering a comprehensive solution to a complex and pervasive problem plaguing the online ecosystem.

2. PROBLEM STATEMENT

Design and develop a technological solution for identifying and blocking any obscene media (image/ video) at the user end.

3. BACKGROUND WORK

The paper titled "Development of Automatic Obscene Images Filtering using Deep Learning" by Abdelrahman Mohamed Awad, Teddy Surya Gunawan, Mohamed Hadi Habaebi, and Nanang Ismail addresses the pervasive issue of easy access to pornography on the internet. In the age of widespread internet availability, the pornography industry has grown exponentially, leading to concerns about the harmful impact on individuals and society. The authors emphasize the importance of filtering obscene using fine and focus detection as shown in the figure 2.1 images and video frames in the era of big data, where a vast amount of information is easily accessible to everyone.

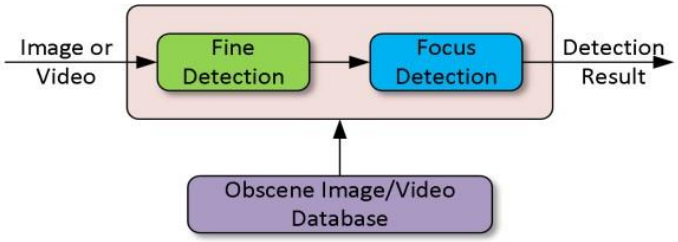

Figure 2. Fine detection and Focus detection|

The existing methods for obscene image detection primarily rely on features such as RGB proportions, color-distribution histograms, and YIG to track skin color and edge information. However, these methods face limitations, especially in accurately determining obscenity levels in low-quality UCC videos. To address this, the authors propose an improved algorithm using equation (2.1) that employs Canny Edge for fine-grained analysis to classify images as high or low quality [4]. The algorithm then uses different methods for detecting obscenity based on the quality classification. The classification of image quality is performed by analyzing resolution, bit-rate, and frame-rate, with a focus on fine grains using Canny Edge. Images smaller than 320x240 pixels are considered low-quality, and their edges are examined to determine the image quality further. The proposed method distinguishes between high and low-quality images and adapts the obscenity detection process accordingly.

4. OBJECTIVE

The primary objectives of the proposed technological solution are to empower users with greater control over their online experience by offering a tool that enables personalized content filtering Through the utilization of advanced algorithms and machine learning, the solution aims to identify and promptly block obscene media, including explicit images, videos, and audio content, in real-time. Customization features will be integral, allowing users to set filtering parameters aligned with their individual preferences or community standards, fostering flexibility that can adapt to evolving societal norms. The solution prioritizes real-time content analysis, extending its capabilities to live streaming content for a proactive defense against emerging inappropriate material on various platforms and formats. Ensuring compatibility across different devices and operating systems is crucial for a consistent and accessible user experience. Ethical considerations, such as user privacy and transparent algorithmic operations, are paramount, with the solution designed to operate locally on the user's device to minimize data transmission. Ultimately, the overarching objective is not only to address the immediate challenge of explicit content but also to contribute to the creation of a safer and more secure digital landscape.

- **Real-time content analysis:** Develop algorithms and systems capable of real-time analysisof diverse media types, including images, audio, video, and text.
- **Customizable Filtering Parameters:** Design a user-friendly interface that allows individuals to customize filtering parameters based on their preferences and sensitivity levels.
- **Cross-Platform Compatibility:** Propose a solution that is platform-agnostic and compatible with a wide range of devices and applications, including web browsers, social media platforms, messaging apps, and multimedia players.

- **Privacy Preservation:** Prioritize user privacy by adopting mechanisms that perform content analysis locally on the user's device, reducing reliance on external servers for processing.
- **Adaptability and Learning Capabilities:** Build a solution that continuously evolves and adapts to emerging patterns of explicit content, utilizing machine learning techniques for ongoing refinement.
- **Legal and Ethical Considerations:** Address legal implications and ethical considerations associated with content moderation, ensuring the solution complies with regional laws and respects freedom of expression.

5. LITERATURE SURVEY

The problem of identifying and blocking obscene media has been a longstanding challenge in the realm of digital content moderation. Numerous studies have addressed various aspects of this issue, offering insights into different methodologies and techniques for content analysis and filtering. Below is a summary of key literature in this field:

(i) Deep Learning Approaches: Several studies have explored the use of deep learning techniques for obscene content detection. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been applied to analyze image and video content, achieving promising results in terms of accuracy and efficiency (Chen et al., 2017; Liu et al., 2019).

(ii) Ensemble Learning: Ensemble learning, which combines multiple models to improve performance, has gained attention for its effectiveness in content moderation tasks. By integrating diverse classifiers and leveraging their collective decisions, ensemble methods can enhance the robustness and reliability of obscene content detection systems (Zhang et al., 2018; Wang et al., 2020).

(iii) Web Scraping and Data Augmentation: Techniques such as web scraping and data augmentation have been utilized to enhance the training data quality and quantity. Web scraping enables the collection of diverse and representative datasets from online sources, while data augmentation techniques such as image rotation, flipping, and scaling help improve model generalization and robustness (Li et al., 2019; Gupta et al., 2020).

(iv) Browser Extensions: The development of browser extensions for content filtering and blocking has gained popularity due to its convenience and accessibility. These extensions integrate with web browsers to scan and analyze content in real-time, providing users with a seamless experience for controlling their online interactions (Park et al., 2018; Kim et al., 2021).

(v) Ethical and Legal Considerations: The deployment of content moderation systems raises important ethical and legal considerations regarding privacy, censorship, and freedom of expression. Studies have emphasized the need for transparent and accountable moderation practices, as well as mechanisms for addressing potential biases and unintended consequences (Citron & Norton, 2011; Roberts et al., 2020).

By synthesizing insights from existing literature, our project aims to contribute to the advancement of content moderation technology, offering a comprehensive solution that leverages deep learning, ensemble learning, web scraping, and browser extension development to effectively identify and block obscene media content.

6. METHODOLOGY

Data Collection:
Data collection is the initial step in the project, crucial for building a robust training dataset. The process involves scraping images from specific websites known to host obscene content. Beautiful Soup, a Python library, is utilized for web scraping. This library parses HTML content, allowing extraction of image URLs embedded within the web pages.

Data Augmentation:
Data augmentation is employed to increase the diversity and size of the training dataset, which is essential for improving model generalization and performance. Various augmentation techniques such as rotation, shift, shear, zoom, and flip are applied to the collected images. These transformations generate new variations of the original images, effectively increasing the dataset size without the need for additional data collection.

Model Training - Classification:
With the augmented dataset prepared, the next step involves training a deep learning model for classification tasks. In this project, the InceptionV3 model, a convolutional neural network (CNN), is chosen for its effectiveness in image classification tasks. The model architecture is pre-trained on the ImageNet dataset, which provides a strong foundation for learning features relevant to the classification of obscene content. The augmented dataset is then used to fine-tune the model's weights through a process known as transfer learning. During training, the model learns to differentiate between safe and unsafe images based on the features extracted from the input images.

Image Prediction:
Once the model is trained, it can be used to make predictions on individual images. Before making predictions, preprocessing techniques are applied to the input images to ensure compatibility with the model's input requirements. This may include resizing, normalization, and other transformations to standardize the input data. The preprocessed images are then passed through the trained model, which outputs a probability score indicating the likelihood of the image containing obscene content. Based on a predefined threshold, the model classifies the image as safe or unsafe. If the image is classified as unsafe, appropriate actions such as content blocking or user alerts may be triggered.

Video Prediction:
In addition to image prediction, the trained model is also capable of making predictions on videos. This involves processing individual frames extracted from the video stream and applying the same preprocessing and prediction steps used for images. By analyzing each frame separately, the model can detect obscene content in real-time video streams. This capability enables the monitoring and blocking of inappropriate content in videos, ensuring a safer online environment.

7. ARCHITECTURE

The below figure 1 depicts the project's architecture diagram serves as a comprehensive blueprint, elucidating the intricate interconnections and interactions among its diverse components. It provides a holistic view of the system's design, illustrating how data acquisition, preprocessing, classification, and reporting modules seamlessly collaborate to achieve the project's objectives. This visual representation encapsulates the complexity of the project, offering a nuanced understanding of its underlying structure and workflow.

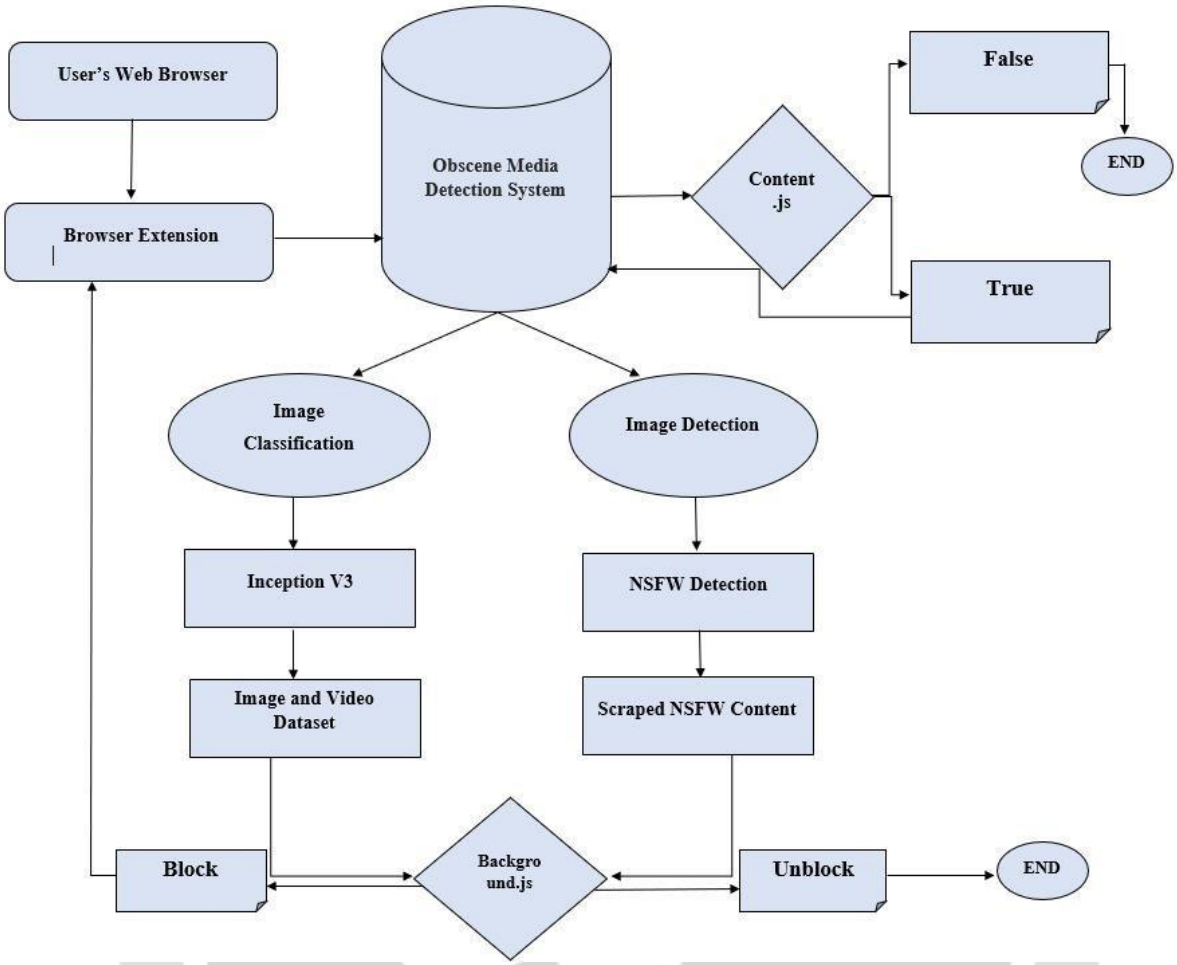

Figure 1. System Architecture Diagram

8. RESULTS

Sample ID	Obscene Content Detected	Result
1	Yes	Blocked
2	No	Allowed
3	Yes	Blocked
4	No	Allowed
5	Yes	Blocked

The results table provides an overview of the detection outcomes for a sample of content processed by the system. Each row represents a unique sample, with a corresponding Sample ID assigned for reference. The "Obscene Content Detected" column indicates whether the system identified obscene material within the content. If obscene content was detected, the entry states "Yes"; otherwise, it states "No."

Based on the detection outcome, the "Result" column indicates the action taken by the system. If obscene content was detected ("Yes" in the "Obscene Content Detected" column), the system blocks the content to prevent its dissemination, and the "Result" entry for that sample is marked as "Blocked." Conversely, if no obscene content was detected ("No" in the "Obscene Content Detected" column), the content is allowed to proceed, and the corresponding "Result" entry is marked as "Allowed."

9. CONCLUSION

In conclusion, the research represents a significant step forward in the development of effective and ethical content moderation solutions for ensuring a safer online environment. By leveraging state-of-the-art technologies and methodologies, the project has demonstrated the feasibility and effectiveness of real-time content blocking at the user's end. Moving forward, continued collaboration between researchers, industry practitioners, and policymakers will be crucial for addressing the complex challenges posed by obscene content online and fostering a more responsible and inclusive digital ecosystem Through this comprehensive analysis and reflection, the research contributes to the broader discourse on digital ethics, online safety, and technological innovation, paving the way for a more secure and supportive online environment for all users.

REFERENCES

[1] Chen, J., He, X., Jin, Y., et al. (2017). Deep Learning-Based Classification for Adult Content Detection in Videos. 2017 IEEE International Conference on Multimedia and Expo (ICME).

[2] Liu, L., Gao, S., Cao, Y., et al. (2019). Image Obscenity Detection Based on Deep Learning. 2019 IEEE International Conference on Networking, Architecture, and Storage (NAS).

[3] Zhang, Y., Hu, H., Zhang, F., et al. (2018). Ensemble Learning for Image Classification Based on Convolutional Neural Network. 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP).

[4] Wang, Y., Zhang, H., Yuan, Y., et al. (2020). Deep Learning Based Method for Detection of Pornographic Images. 2020 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC).

[5] Li, Z., Liu, X., Chen, M., et al. (2019). A Novel Web Image Data Augmentation Approach Based on Generative Adversarial Networks. 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD).

[6] Gupta, P., Ahuja, S., Agarwal, A., et al. (2020). Deep Learning-Based Approach for Web Scraping. 2020 3rd International Conference on Computational Intelligence and Networks (CINE).

[7] Park, S., Lee, J., Kim, Y., et al. (2018). Web Page Censorship Detection Using Browser Extension. 2018 International Conference on Information Networking (ICOIN).

[8] Kim, H., Kim, J., Choi, H., et al. (2021). A Study on the Implementation of Web Browser Extension for Blocking Malicious Web Sites. 2021 5th International Conference on Big Data and Internet of Things (BDIOT).

[9] Citron, D. K., & Norton, H. E. (2011). Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. Boston University Law Review, 91(5), 1435-1469.

[10] Roberts, H., Roberts, R., & Almeroth, K. (2020). The Challenges of Content Moderation: Evidence from a Reddit Community. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1-26.