A Safer Web Experience: Deep Learning-

Enhanced Obscene Content Filtering Plugin

Pratham Taneja  
*Electronics and Communication Engineering*  
*Dr. Akhilesh Das Gupta Institute of Technology & Management*New Delhi, India  
pratham2014taneja@gmail.com

Divyansh Singh  
*Information Technology*   
*Dr. Akhilesh Das Gupta Institute of Technology & Management*New Delhi, India  
singhdivyansh0110@gmail.com

Tushar Rajora  
*Computer Science and Engineering*  
*Dr. Akhilesh Das Gupta Institute of Technology & Management*New Delhi, India  
tusharrajora72@gmail.com

***Abstract***— **Our lives have benefited greatly from the internet's rapid development, but it has also presented new challenges. Obscene content and advertising on websites is one such issue that users find highly offensive and upsetting. The language or imagery used in the content, as well as any graphic or personal values, may be explicit. When obscene content is readily available, it can pose a major risk to youngsters who are unaware of it, negatively affecting their development. Pictures, videos, or words that are inappropriate for minors' emotional development may be found in this type of content. Users are distressed when obscene content is present on web pages, which has a detrimental effect on their online experience. Users may mistakenly click on obscene adverts or other content that directs them to dangerous websites, endangering their safety and online security. Our study report suggests a novel solution—installing a web plugin or extension in the web browser—to address this crucial issue. This plugin will examine or scan the webpage's elements and advertisements. It will be able to tell the difference between NSFW (Not Safe For Work) and non-NSFW content and use a blocking mechanism to obfuscate the NSFW or offensive material.**

Keywords— Web Extension, Obscene Content Blocker, Deep Learning Techniques, MongoDB, Web Plugin, HTML, CSS

# INTRODUCTION

In present-day society, the widespread integration of an economy that has readily embraced an excess of sexualization has resulted in a pervasive saturation of explicit themes, imagery, and content across multiple facets of our existence. Although acknowledging the significance of discussing such subjects for educational purposes, it is crucial to recognize that an undue emphasis on sexual content can result in severe adverse consequences. Emerging research indicates a correlation between exposure to sexually explicit content and the manifestation of antisocial behaviour [1]. With a substantial portion of the global population, including both adults and children, spending extensive periods online consuming diverse forms of media and information, the imperative of distinguishing between obscene content and educational material is evident, particularly in the context of safeguarding young minds. The ramifications of exposure to violent and pornographic content, even when emotionally unrelatable for kids and teens, can have enduring impacts on their psychological well-being [2].

In recent times, reports have surfaced where users have unexpectedly encountered sexual or obscene visuals while browsing websites that should ostensibly remain devoid of such material [3]. This unsettling phenomenon can arise due to deliberate placements or unintended vectors, including web advertisements [4]. These occurrences disrupt the overall browsing experience, dissuading visitors from revisiting websites and subsequently causing reputational and traffic losses for the implicated sites.

The endeavour at hand aims to develop a pioneering web extension that harnesses the potential of deep learning technology. This extension will be trained to discern various forms of content, such as images, videos, and pop-up ads, and subsequently suppress or censor such content during the user's browsing session. Moreover, the gathered data will be shared with multiple internet safety organisations, contributing to a collective effort to maintain online decorum. The primary objective here is to create an adept classification model that categorises websites into three color-coded groups based on the extent of displayed obscene content.

The pervasive expansion of the internet has democratised access for a diverse spectrum of demographics. The fabric of modern society is now intricately woven through the integration of the Internet into homes, workplaces, libraries, and portable devices. However, ubiquitous accessibility has precipitated an exponential surge in the generation and dissemination of online content, including advertisements. This prevalence of explicit content not only raises ethical concerns but also elicits psychological, societal, and legal implications that necessitate thoughtful intervention.

In response to these evolving challenges, the development of obscenity blockers has emerged as a pivotal technological advancement. Often referred to as content filters or parental control systems, these tools are designed to pre-empt unintended exposure, especially among minors, to graphic, inappropriate, or objectionable content. These blockers operate on a combination of algorithms, heuristics, and predefined criteria, functioning to automatically detect and restrict potentially detrimental material.

In light of the dynamically evolving technology landscape, the relevance of obscenity blockers has only amplified. This research paper introduces the proposition of a plugin capable of scrutinising webpage content for obscene material. By leveraging the capabilities of Deep Learning, the plugin effectively categorizes and subsequently suppresses such content. Furthermore, records of websites containing obscene content are meticulously maintained in a comprehensive database accessible to pertinent authorities. Regular intervals witness the generation and sharing of analysis reports, delineating the percentage of obscenity and the cumulative count of recorded alerts, via email and SMS to the concerned agencies. This paper delves into the intricate details of this proposed solution, aiming to present a robust strategy for mitigating the impact of obscene content in the digital age.

# BACKGROUND AND LITERATURE REVIEW

Mr. Shubham D. Naik et al. [5] introduced a method to filter skin regions, the YCbCr color model technique employs an orthogonal color space. It distinguishes between luminance and chrominance components, improving skin color identification in a variety of lighting circumstances. The equation for converting RGB to YCbCr is Y = 0.299R + 0.587G + 0.114B, Cb = R minus Y, and Cr = B minus Y. Cb and Cr denote color variations, while Y represents brightness. Since it has unique components, this technique is efficient for skin detection since it divides RGB into Y (luminance), Cb (chrominance blue), and Cr (chrominance red). Because of its clarity and separation of luminance and chrominance, YCbCr is commonly used for skin detection.

In their work, K. V. Chandrinos et al. [6] introduced a novel approach for identifying explicit content in web pages.A binary vector is used to represent each web page, indicating the presence or absence of specified features such as words or graphics. The technique employs mutual information to identify significant features and a Naive Bayesian classifier to distinguish between pages that contain explicit content and those that do not. Detecting skin tones is an important aspect of the process, and unlike previous methods that used color-space transformations, this new method prioritizes speed and accuracy by using only RGB data. Recognising the importance of visuals, the system considers both text and image attributes for classification. However, simply having potentially graphic photos on a page isn't enough to classify it as improper; accompanying language also plays a role in the assessment. It is possible to integrate image attributes.

In the work conducted by W.H. Ho et al. [7] the objective revolves around the effective filtration of online content, for successful content filtering, the study identifies distinguishing characteristics between pornographic and non-pornographic online pages. It creates personalised filters using Bayes' rule modifications, outperforming simple association rules. Classifier adaptations show potential for minimising false positives and negatives. The applicability of the technique to additional content categories, the difficulties in integrating text-based algorithms with multimedia content, and the need for enhanced solutions are all acknowledged. In conclusion, Bayes' rule modifications demonstrate content filtering potential, emphasising integration issues and advanced techniques.

Jau-Ling Shih et al. [8] work focuses on spotting adult photos using image retrieval This work describes an approach for recognising adult photos using a content-based image retrieval technique. The first step is to remove the backdrop to isolate the region of interest, after which non-skin-like pixels are removed. The solution collects the 100 most comparable images from a pre-established image library for each input image using MPEG-7's SCD, EHD, and a newly introduced CD feature. When the number of detected adult photos in these recovered images exceeds a particular threshold labelled Tad," the searched image is classified as an adult image.

André Tabone et al. [9] is proposed in a paper to help law enforcement detect and classify adult and child pornography. This paper proposes an obscenity organ detector to help law enforcement identify and classify pornography. This approach aims to reduce false alarms by directly detecting explicit content. This framework detects genitals with high accuracy and minimal false alarms. In particular, the detector's learning capabilities include child sexual abuse (CSA)-related content. The Accuracy score is 87.8, and these traits distinguish between multiple porn scenarios. Various domains can be addressed through this concept and an efficient multitasking model. Additionally, the authors plan to work with the Missing Persons Foundation to refine the model for the CSA material.

In this study conducted by Chung, Myoungbeom & Ko[10], a high- and low-quality segmented algorithm for the detection of explicit images was suggested. This explicit image identification serves as the cornerstone of a larger system for filtering pornographic content. The method discussed here examines elements including image resolution, fine details (through Canny Edge analysis), and textural properties. This technology is especially useful for locating explicit material in low-resolution movies or photographs taken with mobile phones. The incorporation of factors that affect video quality, such as bit rate or frame rate, is projected to result in higher accuracy levels. Particularly in cases involving movies of various quality, this comprehensive technique has the potential to greatly improve the detection accuracy across explicit content.

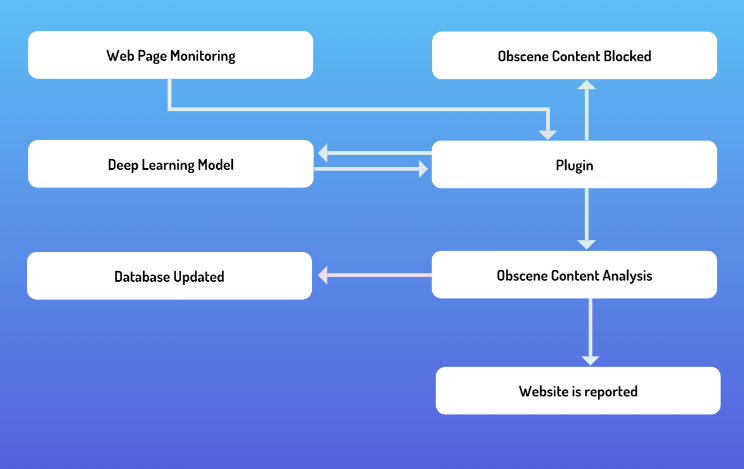
# METHODOLOGY

## WEB PLUGIN

The heart of this system is an innovative web extension written in JavaScript[11]. Users will install this extension to make data collecting easier, and it will be given the rights it needs to keep track of their browsing habits. The user's interactions with web content, such as photos, advertisements, gifs, and pop-up windows, will be recorded by the extension once it has been activated. Each of these separated components will be individually assigned a special serial number that will serve as one of their distinctive identification identifiers.

The web extension will manage the transfer of these recognised elements to a central server after the extraction and tagging processes are completed. This server holds the machine learning model that is in charge of detecting whether the material is possibly NSFW. The transmission will be aided by a streamlined data pipeline, which will improve the efficiency and speed of information transfer.

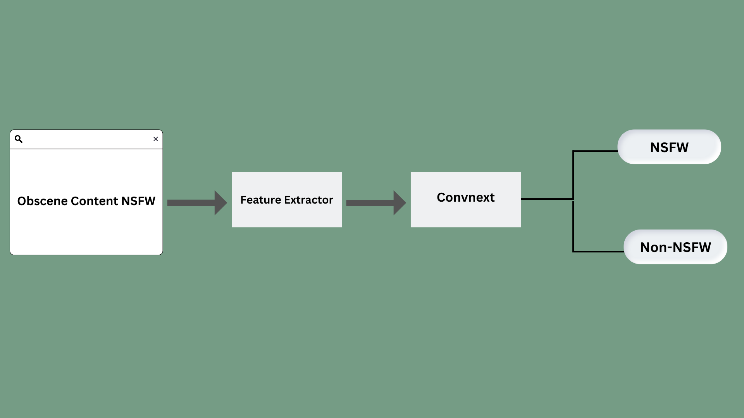
This design combines the web extension's real-time tracking capabilities with the machine learning model's analytical power. It guarantees that each extracted element is appropriately labelled before being submitted to the server for classification and, if considered NSFW, may be subjected to content blurring. Through the controlled extraction and processing of web material, the technology contributes to a safer browsing experience while preserving user privacy.



1. Plugin linkage description

## DATA PRE-PROCESSING

The collected data will be formatted to ensure consistency. Videos and gifs will be examined frame by frame, allowing for careful examination of dynamic content [12]. Images will be downsized to 240 by 240 pixels at the same time, promoting consistency across the collection. Additionally, photos will be converted to the RGB color standard, which includes red, green, and blue colors for accurate visual representation. These pre-processing stages ensure that the data is coherent and ready for model application. The rigorous frame analysis of movies and gifs, as well as image resizing and color conversion, improves the dataset's quality. The plugin will meticulously monitor the data from each website visited, noting the percentage of identified NSFW content. This data will be carefully maintained in a dedicated database, giving a thorough record of the prevalence of explicit content on each site.



1. Classification of visual content on site

## DATABASE

The data will be stored in a MongoDB database, it will share the accumulated data with organisations focused on ensuring secure internet use and safeguarding online experiences on a regular basis, often weekly or monthly. Websites will be classified into three unique color-coded categories: red, yellow, and white, using a categorization algorithm. Each color represents a different level of content nature. The "red" category will denote websites with extremely obscene content. Comprehensive censorship may be difficult due to the large amount of explicit material. The "yellow" category will reflect websites with moderate degrees of obscene content, ranging from extreme to safe. Finally, the "white" category will include websites that have been considered completely free of any obscene content, offering a safe and appropriate browsing experience.

This classification approach assists users in determining the type of content they may encounter on various websites, allowing them to make educated decisions regarding their online interactions. By providing clear indicators of material intensity, it provides a proactive way to regulate internet exposure and build a safer online environment.

This proactive strategy enables internet security agencies to take informed action against websites that contain too explicit information. Corrective actions can be implemented to mitigate the appearance of unsuitable material by collecting and sharing data, boosting the overall safety of online browsing for users. This combined effort by the plugin, database, and protective agencies is a positive step towards creating a more safe and respectful online ecology.

## MODEL WORKING

The current study used an experimental approach to examine ConvNeXt's [13] efficacy in categorising NSFW and non-NSFW photos. A dataset of over 50,000 NSFW photos and 40,000 non-NSFW photographs was gathered. The photos were scaled to a resolution of 240x240 pixels to ensure homogeneity in the input data, resulting in a pixel value range of [0-255].

A Normalisation layer was then used, which is included in the ConvNeXt model architecture, to construct a preprocessing strategy. This enabled conversion of the input data into a format that met the model's needs. The Normalisation layer, in particular, normalises the pixel values of the input tensor to have zero mean and unit variance, allowing the model to learn more successfully.

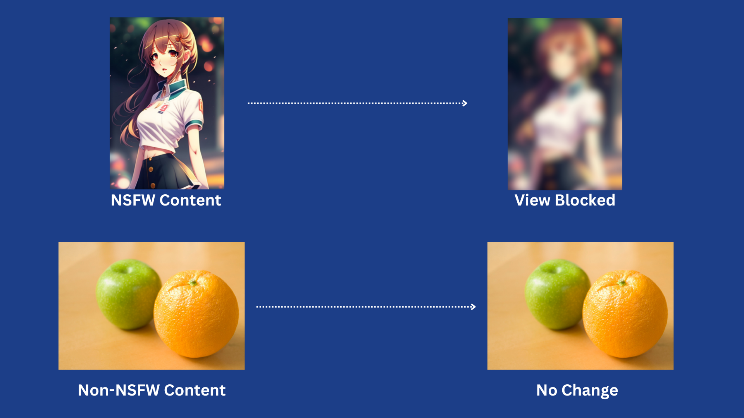
An approach of freezing the model weights during training to improve the performance of the ConvNeXt model was implemented. By doing so, training time was greatly reduced, allowing to attain satisfactory outcomes in a fair amount of time. In this experiment, the model was trained for 10 epochs before evaluating its performance on both the training and validation datasets.

The results were quite positive, with the model obtaining 98% training accuracy and 95% validation accuracy. When given with a large and diverse dataset, these findings show that ConvNeXt can easily identify between NSFW and non-NSFW photos. Furthermore, the excellent validation accuracy implies that the model generalised well to previously unreported data, bolstering its robustness.

Finally, the methods used in this study confirmed ConvNeXt's effectiveness in categorising NSFW and non-NSFW photos. Outstanding accuracy scores were achieved by combining a large dataset, suitable preprocessing, and ideal model architecture. These findings provide vital insights into ConvNeXt's capabilities and lay the groundwork for future image categorization research.

# RESULT ANALYSIS

We conducted an empirical study to evaluate the effectiveness of our approach. I used a website made with HTML and CSS. The core functionality of our model was tested by integrating a web plugin. This comprehensive evaluation included a variety of visual content and strategically timed presentations of pop-ups to accurately assess model responsiveness.



1. Response to obscene content vs. non-obscene content

It is important to note that the images featured are of a relatively bland nature by the standards of reputable magazines. However, it is important to emphasize that our model's performance extends equally to content with varying degrees of explicitness. This is consistent with the expected results after the discrimination process for both obscene and non-obscene visual stimuli.

Our methodology serves as a robust framework for evaluating a model's ability to detect objectionable content. Using real-world scenarios across different manifestations increases the validity and applicability of our results. This study not only demonstrates the capabilities of the model but also contributes to the broader discussion of content filtering and moderation. The database stores aggregate data about each website's obscene content composition and shares it with Internet user protection agencies on a weekly or monthly basis.

# CONCLUSION

This study leverages advanced computer vision web plugins and machine learning technology to present an innovative strategy to address the pressing problem of obscene Internet content. The main goal is to seamlessly integrate visual data from websites with robust machine learning models to provide a holistic solution facilitated by an optimized data pipeline.

One of the biggest challenges we are trying to address is the rapid proliferation of explicit content on the web. Traditional methods of content moderation often lag behind the speed with which such material is disseminated and can put users, including minors, at risk. Our approach aims to fill this gap by enabling real-time content evaluation through a combination of web plugins and machine learning algorithms. By integrating web plugins with machine learning models, our system has the ability to quickly and accurately assess content types and identify potentially hazardous work (NSFW) images. When explicit content is detected, these elements are immediately masked by automatic responses, ensuring immediate user protection and promoting a safer browsing experience.

A key aspect of our methodology is data collection and analysis. This plugin carefully monitors and collects his website data to help build a comprehensive database that quantifies his percentage of NSFW content per website. It uses a color-coded classification system to make it easier for users to understand. Red represents extreme content, yellow represents moderate content, and white represents safe content. In addition to technical implementation, this research focuses on Internet security collaboration and advocacy. Regularly sharing data with internet safety organizations facilitates concerted efforts to limit explicit online content. This collaborative approach will help us fix websites with overly explicit content and ensure a more respectful and safer browsing experience for all users.

Essentially, this research pioneers the convergence of computer vision, machine learning, and cybersecurity. Our solution takes an important step towards maintaining a safe and responsible online ecosystem by providing a way to quickly rate, classify and moderate explicit content. Users can confidently navigate digital environments, whether pursuing educational pursuits, professional endeavors, or leisurely exploration.

##### References

1. Mundorf, N., Allen, M., D'Alessio, D., & Emmers-Sommer, T. M. (2007). Effects of Sexually Explicit Media. In R. W. Preiss, B. M. Gayle, N. Burrell, M. Allen, & J. Bryant (Eds.), Mass media effects research: Advances through meta-analysis (pp. 181–198). Lawrence Erlbaum Associates Publishers.
2. Kumar, Sanjeev & Pallathadka, Harikumar. (2022). EXPOSURE TO PORNOGRAPHY AND HARMFUL MATERIALS AGAINST CHILDREN: AN ANALYSIS. Journal of Critical Reviews.
3. Matthews, M.F. (2007), "Virtually Obscene: The Case for an Uncensored internet", Online Information Review, Vol. 31 No. 5, pp. 715-716. doi.org/10.1108/14684520710832450.
4. Sood, Aditya & Enbody, R.J.. (2011). Malvertising - Exploiting web advertising. Computer Fraud & Security. 2011. 11–16. 10.1016/S1361-3723(11)70041-0.
5. Naik, Mr., & Jaiswal, Prof. (2020, June 6). *Analysis And Development of Obscene Image Detection Algorithm*. Retrieved August 30, 2023, from ijcrt.org/papers/IJCRT2006511.pdf
6. Chandrinos, K.V., Androutsopoulos, I., Paliouras, G., Spyropoulos, C.D. (2000). Automatic Web Rating: Filtering Obscene Content on the Web. In: Borbinha, J., Baker, T. (eds) Research and Advanced Technology for Digital Libraries. ECDL 2000. Lecture Notes in Computer Science, vol 1923. Springer, Berlin, Heidelberg. doi.org/10.1007/3-540-45268-0\_50.
7. W. H. Ho and P. A. Watters, "Identifying and Blocking Pornographic Content," 21st International Conference on Data Engineering Workshops (ICDEW'05), Tokyo, Japan, 2005, pp. 1181-1181, doi: 10.1109/ICDE.2005.227.
8. Jau-Ling Shih, Chang-Hsing Lee, Chang-Shen Yang,An adult image identification system employing image retrieval technique,Pattern Recognition Letters,Volume 28, Issue 16,2007,Pages 2367-2374,ISSN 0167-8655,doi.org/10.1016/j.patrec.2007.08.002.
9. Tabone, Andre & Camilleri, Kenneth & Bonnici, Alexandra & Cristina, Stefania & Farrugia, Reuben & Borg, Mark. (2021). Pornographic content classification using deep-learning. 1-10. 10.1145/3469096.3469867.
10. Chung, Myoungbeom & Ko, IlJu & Jang, Daesik. (2010). Obscene Image Detection Algorithm Using High-and Low-Quality Images. IJEI. 2. 522 - 527. 10.4156/ijei.vol2.issue1.2.
11. S. Ndichu, S. Ozawa, T. Misu and K. Okada, "A Machine Learning Approach to Malicious JavaScript Detection using Fixed Length Vector Representation," 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489414.
12. Kim, Inhwa & Kuljis, Jasna. (2010). Applying Content Analysis to Web-based Content. CIT. 18. 10.2498/cit.1001924.
13. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T. and Xie, S., 2022. A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11976-11986).