iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59476

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Sensitive and Obscene Content Blocker (SOCB)

Manish Shingre[1], Sahaya Lebisha Nadar[2], Sahil Godse[3], Imran Mirza[4]
Don Bosco Institute Of Technology, Dept of Computer Engineering, Mumbai, India

Abstract: ***In today's digital age, the internet is a vast and diverse platform that offers a wealth of information and entertainment. However, this openness also exposes users to sensitive and obscene content, which can have a detrimental impact on individuals, particularly children and vulnerable populations. To mitigate the exposure to such content, a Sensitive and Obscene Content Blocker has been developed and tested. This paper explores the design, development, and evaluation of a comprehensive content filtering system that is aimed at identifying and blocking sensitive and obscene content across various online platforms. The system employs advanced algorithms, machine learning techniques, and real-time content analysis to ensure a reliable and robust filtering mechanism. The paper outlines the key components of the blocker, including content analysis methods, their detection and classification methods. The study presents the results of extensive testing, demonstrating the system's efficiency in blocking inappropriate content while minimizing false positives. The Sensitive and Obscene Content Blocker serves as a powerful tool to protect individuals from harmful online content, promoting a safer and more secure digital environment. This paper provides valuable insights into the development and implementation of content blockers and their role in enhancing online safety and security.***

Keywords: Obscene, Sensitive, Media Blocker, Proxy, Support vector Regression, OpenAI's Whisper

I. INTRODUCTION

The advent of the digital age has ushered in a transformative era, marked by the pervasive influence of the internet and digital technologies. This paradigm shift has not only revolutionized the way we connect and share information but has also fundamentally altered the nature of our interactions with the world around us. The digital age has bestowed upon us unparalleled connectivity and granted access to an expansive reservoir of information, heralding both opportunities and challenges that have become increasingly prominent in contemporary society. In this virtual landscape, a notable concern that has gained prominence is the proliferation of sensitive and obscene content across the vast expanse of the internet. The unbridled nature of the online realm exposes users of all age groups to a spectrum of inappropriate, offensive, and potentially harmful material. This poses significant risks, especially for vulnerable populations, such as children, who may inadvertently stumble upon explicit or disturbing content. Additionally, it impacts individuals seeking to curate their online experience in alignment with personal values or preferences. In response to these pressing concerns, the development and implementation of Sensitive and Obscene Content Blockers have become imperative. These digital safeguards are strategically designed to shield users from encountering inappropriate and offensive content as they engage with various online platforms. Serving as a crucial barrier, these blockers play a pivotal role in fostering a safer and more secure digital environment. Among the demographic groups that stand to benefit most from these content blockers are vulnerable populations, particularly children. By effectively shielding them from material that may not be suitable for their age or emotional maturity, these tools play a pivotal role in safeguarding the well-being of the younger generation. Moreover, adults also find value in these blockers as they enable a more controlled and respectful online experience, allowing users to tailor their digital environment to their individual preferences. This comprehensive paper delves into the nuanced development, intricate functionality, and profound impact of Sensitive and Obscene Content Blockers, underscoring their paramount significance in the digital landscape. The exploration extends to the methods employed for detecting and filtering inappropriate content, often involving a blend of automated algorithms and human moderation. Furthermore, these blockers frequently offer customization options, empowering users to fine-tune the level of content filtering based on their unique needs and preferences. The ethical dimension surrounding the concept of content censorship is a critical aspect addressed in this paper. Striking a delicate balance between the right to free expression and access to information on one hand, and the imperative to protect individuals—especially the most vulnerable—on the other, presents a formidable challenge. This paper critically engages with these ethical dilemmas, meticulously examining the practical implications of implementing content blockers with an aim to reconcile the competing values of freedom and safety. In an age where the internet has seamlessly integrated into our daily lives, it becomes imperative to comprehend both the capabilities and limitations of these blockers. This knowledge empowers individuals to make informed choices about their online interactions, contributing to a more responsible and secure digital world.

©IJRASET: All Rights are Reserved

The overarching goal of this paper is to illuminate the profound significance of sensitive and obscene content blockers and emphasize their pivotal role in ensuring a safer, more respectful online environment for users of all ages. It underscores the importance of achieving a delicate equilibrium between protecting users from harm and upholding the principles of free expression in the dynamic and evolving digital realm.

II. RELATED WORK

There are some existing systems closely related to the proposed idea of developing an obscene content blocker. A literature review of the proposed systems or the existing systems was performed to analyze these systems. It gives an idea of how the already existing systems were made or developed, the different implementation methods used to establish software according to the required requirements and issues faced by the developers during the period of development and the most convenient or possible outcomes of each method. The project introduces a nudity detection algorithm developed for integration into a web application, particularly an open-community networking platform supporting discussions, post sharing, article writing, and freelancing. The algorithm employs a trained CNN model to determine if an uploaded image is safe for work. If deemed unsafe, the image is discarded, contributing to ongoing model training for improved accuracy. The training involves two datasets—one with nude images labelled as unsafe and the other with non-nude images labelled as safe. The CNN model exhibits comparable or superior performance to other techniques. The algorithm's backend process involves taking an image from the application, passing it through convolution and max pooling layers in the CNN network, using a reactivation function, flattening tensors, connecting to a dense layer, employing SoftMax for prediction, and finally classifying the image as safe or unsafe for work. The three main steps for application integration include inserting the picture, CNN prediction, and displaying the output indicating the image's safety status.[1]

The paper discusses contextual advertising, where the decision to place an ad on a website is based on the context of the host website. The stakeholders include the publisher, advertiser, advertising network, and end users. The paper addresses brand safety concerns during the examination of websites for advertising. It proposes image classification as the first attempt to tackle brand safety, highlighting that users can identify unsafe images faster than unsafe text. The paper introduces two improvements: an image multi-class classification method using deep residual learning and network pruning, and the creation of a dataset for image brand safety with six categories. Experiments demonstrate improved performance compared to original methods. The paper is structured to cover related work, deep residual learning, network pruning methods, experiments, results, and conclusions in different sections.[3]

The paper proposes an Adaptive Inspection Scheduling (AIS) scheme for mobile obscenity detection in personal broadcasting services. To address issues of battery consumption and quick detection, a lightweight deep learning model and scheduling techniques are required. The proposed AIS adaptively adjusts the inspection interval based on previous inspection results. It distinguishes between Green, Yellow and Red phases depending on the presence of obscene frames. In the Green phase, equal and long inspection intervals are used to save the battery. If an obscene frame is detected, the Yellow phase collects inspection results to determine obscenity, and the Red phase stops inspecting if a predefined threshold is exceeded. Simulation results show that AIS detects obscenity faster with low battery consumption compared to other schemes. Performance is evaluated using a Galaxy S10+, and the AIS scheme is demonstrated to have minimal impact on service interruptions, distinguishing it from streaming service issues.[4]

The paper focuses on understanding and analyzing the ecosystem of internet proxies, particularly open proxies and residential proxies, with a primary motivation to improve user privacy while addressing security concerns associated with malicious proxy activities. Open proxies, accessible to the public, and residential proxies, using IP addresses from Internet Service Providers (ISPs), are compared based on distribution, regional background, and behaviour. The analysis covers geospatial distribution, blacklisting, and correlation with country-level characteristics. The study involves a large dataset of 1,045,468 open proxies and 6,419,987 residential proxies. Key findings include a significant portion of proxies being blacklisted, with a notable percentage used for spam and associated with verified attacks. The paper also explores correlations between proxy distribution and country-level factors such as internet speed and gross domestic product (GDP). The contributions include insights into proxy behaviour, geo-location distribution, and correlations with country-level characteristics. The paper concludes with a comprehensive organization of the research and its findings.[5]

The project addresses the novel challenge of predicting whether a user employing an ad blocker will whitelist an entire website or just the specific page they intend to visit when confronted with an anti-ad-blocking mechanism. The prediction models developed utilize user features, page features, and past user engagement behaviour, presenting a pioneering approach to user whitelist behaviour prediction through data analytics.

The study implements four predictive models, with the gradient boosting regression tree model demonstrating the highest accuracy, especially in real-time scenarios. The authors emphasize the potential application of whitelist prediction in devising personalized counter-ad-blocking strategies to enhance website visits and revenue. Future work involves evaluating the practical use of prediction models, addressing false positive and false negative errors, determining optimal decision thresholds, and conducting feature analysis to identify the most influential factors in predictions, offering valuable insights for publishers in designing effective personalized counter-ad-blocking strategies.[7]

The paper addresses the increasing demand for Adblock due to the rise of malicious advertising, known as Malvertising. While existing ad blockers use filter lists, they prove inefficient in countering Malvertising. The current most efficient solution, AdGraph, employs supervised machine learning, achieving high accuracy and speed. The paper introduces a new prototype, AdRemover, which also utilizes supervised machine learning to automatically generate filter lists. Unlike traditional training with filter lists, AdRemover trains on blacklists and whitelists, demonstrating improved effectiveness. The organization of the paper includes sections on the difficulty in finding similarity scores for ads, preliminary work, methodology for creating AdRemover, feature extraction, motivation for applying machine learning, related work, and conclusion with future work. The goal is to analyze existing ad blockers and propose an advanced solution for more effective ad blocking.[8]

The study introduces a pornographic filtering application for detecting nudity in images and videos, showing promising results based on a dataset. The application uses pixel-wise skin detection with image processing techniques. The YCbCr colour model is employed for colour representation due to its superiority. The study acknowledges the need for comparisons with other systems and proposes incorporating machine learning for improved accuracy, addressing limitations related to skin colour count dependency. The application is envisioned for integration into larger systems, emphasizing its role in preventing exposure to explicit content, especially for vulnerable groups.[9]

The paper proposes various supervised learning approaches for automatically classifying web images based on their size, utilizing image URLs and available HTML metadata. The first approach relies on the frequency of n-grams from image URLs, with two variations considering correlation with classes. The fourth approach uses tokens instead of n-grams, and the fifth constructs a feature vector from surrounding HTML elements and pages. A hybrid approach combines two of these methods. The text-based method is refined and combined with non-textual classifiers, resulting in a 4 per cent improvement. The proposed approaches are evaluated on images from the Common Crawl dataset, showcasing their effectiveness. The paper emphasizes the novelty of combining textual and non-textual features for image classification. Plans include extending the n-grams approach to alternate and parent text, creating additional classifiers, and exploring the detection of fine-grained image characteristics beyond size, such as distinguishing between landscape and portrait images or differentiating photographs and graphics.[10]

The study adopts a minimalist approach to data pre-processing in training Whisper models for speech recognition, predicting raw transcripts without significant standardization. The dataset, sourced from internet audio paired with transcripts, ensures diversity in audio but faces challenges with subpar transcripts. Automated filtering methods are implemented to enhance transcript quality by detecting and removing machine-generated transcripts. Various heuristics are developed to identify machine-generated content and ensure human-like transcripts. An audio language detector matches spoken language with transcripts, excluding mismatched pairs. Fuzzy de-duping reduces duplication and automatically generates content. Audio files are segmented, including non-speech segments for voice activity detection. After initial model training, a filtering pass identifies and removes low-quality data sources through manual inspection. Deduplication at the transcript level is performed to prevent contamination between the training and evaluation datasets, particularly with TED-LIUM 3. The approach aims to create a diverse and high-quality dataset for training robust speech recognition models.[11]

III. EXISTING SYSTEM

A. Findings in the Existing System

The intricate foundation of the current system for explicit content filtering unfolds through a sophisticated reliance on machine learning algorithms. These algorithms undergo meticulous training on expansive datasets, where images are meticulously labelled as either nude or non-nude. This strategic training methodology empowers the algorithms to not only discern but also internalize intricate patterns associated with nudity. Consequently, this enables the system to effectively detect and categorize explicit content in new images. However, it is imperative to acknowledge the evolving nature of explicit content, necessitating a continuous commitment to updating these algorithms. This ongoing adaptation is crucial to ensuring sustained accuracy in navigating the ever-changing landscape of explicit material proliferation.

Central to the system's arsenal is the deployment of a Convolutional Neural Network (CNN) model, strategically designed for the classification of unsafe images. This deep learning model operates with a level of precision that is particularly pronounced in the realm of filtering images from social media and various online channels. The CNN model's accuracy stands as a testament to its pivotal role in maintaining a vigilant stance against the dissemination of explicit content. Its nuanced understanding of image features positions it as a cornerstone in the continual battle against the dynamic landscape of inappropriate content online.

The proposed approaches in the system adopt a dynamic filtering rule set, showcasing adaptability crucial for effective content filtering. This real-time adjustment to traffic patterns adds a layer of sophistication, ensuring that the system remains resilient in the face of evolving content dynamics. Additionally, the integration of a Random Forest classifier, renowned for its accuracy and robustness, further amplifies the system's efficacy. This dual focus on dynamic rules and advanced classifiers reflects a nuanced strategy aimed at combating explicit content in a continually evolving online environment.

The escalating use of proxies, catalyzed by the popularity of online services like streaming and social media platforms, introduces a substantial challenge in content filtering. Acknowledging the ever-evolving nature of proxy usage and the constant development of new types of proxies, the system takes a proactive stance. Employing a scheduling scheme that minimizes the number of video frames to be inspected reflects a strategic response to potential circumvention techniques. This adaptive system dynamic underscores the system's commitment to staying one step ahead in the perpetual cat-and-mouse game with those seeking to subvert content filters. To fortify the system's capabilities, the authors tap into real-world user data from Forbes Media for training and evaluating their methodology. The utilization of a gradient boosting regression tree model from this dataset emerges as the best-performing model. This underscores the importance of leveraging authentic user data, as it provides a more nuanced and accurate representation of the complexities associated with explicit content. The emphasis on real-world applicability enhances the effectiveness of content filtering systems.

The system adopts a comprehensive and multifaceted approach that extends beyond conventional methods. Incorporating pixelwise skin detection, image processing techniques, texture filtering, and coloured skin detection amplifies the sophistication of the pornography filtering mechanism. This holistic strategy not only underscores the inherent complexity of the content filtering challenge but also showcases the system's commitment to employing diverse and complementary methodologies for robust and accurate detection. The detailed exploration of the existing system reveals a multi-dimensional approach that leverages machine learning algorithms, dynamic filtering strategies, sophisticated classifiers, and real-world data. This comprehensive methodology, underpinned by continuous adaptation and the integration of diverse detection techniques, underscores the system's resilience in addressing the intricate challenges associated with content filtering in the dynamic and ever-evolving online landscape.

B. Limitations

The intricacies within the architecture of the system designed for the blocking of sensitive and obscene content reveal an array of inherent limitations, compelling a meticulous examination to drive comprehensive enhancements. Central to this endeavour is the imperative to fortify the algorithm's resilience by delving into the nuanced realms of variations in pose, lighting, and background. The current training dataset, while informative, inherently lacks the essential diversity in these critical factors, thereby necessitating a strategic initiative to broaden its scope and encompass a more comprehensive range of scenarios.

Moreover, the formidable challenge of real-time detection and classification amplifies the complexity of the system's operational landscape. The existing suite of machine learning algorithms grapples with the consistent and prompt identification and categorization of toxic content, presenting a substantial hurdle. The model's inherent limitation in effectively discerning between visually analogous safe and unsafe images, decoding contextual nuances, and adapting to the dynamically evolving panorama of unsafe content online acts as a pivotal bottleneck, diminishing the overall efficacy of the system.

In an ambitious stride towards enhancing operational efficiency, the incorporation of an advanced scheduling scheme emerges as a pivotal consideration. Such a scheme bears the promise of streamlining the meticulous inspection process, strategically minimizing the number of video frames that demand intensive scrutiny. However, the considerable absence of comprehensive data pertaining to the deployment of proxies for malicious purposes and the intricate dynamics surrounding the evolution of the proxy ecosystem introduces a substantial challenge, impeding the system's adaptability to emergent threats and evolving circumvention techniques.

Furthermore, the conspicuous absence of a standardized methodology for the evaluation of adaptive filtering methods introduces an additional layer of ambiguity in the meticulous assessment of system performance. This scarcity in a uniform evaluative framework complicates endeavours to gauge the effectiveness of adaptive filtering mechanisms, impeding a holistic understanding of the system's capabilities. A noteworthy conundrum arises in the realm of nudity filtering, where accuracy is intricately tied to the meticulous count of skin colours, consequently resulting in a compromise in precision.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

In summation, these multifaceted and nuanced limitations serve as a clarion call for an ongoing, iterative refinement process. This imperative underscores the continuous augmentation required to fortify the system's efficacy in navigating the intricate and complex landscape inherent in the task of filtering sensitive and inappropriate content. The quest for enhancement becomes an evolving narrative, demanding relentless attention to detail and a commitment to addressing the intricacies embedded within the system's operational paradigm.

IV. MODULES

A. *Image Classification*

In the domain of image classification for sensitive and obscene content blocking, the system employs a sophisticated methodology to collect and analyze images from the web. The process begins with a meticulous data-gathering phase, where a diverse and extensive dataset is curated. This dataset comprises images that have been manually labelled as either explicit or non-explicit, providing the necessary foundation for the training of the image classification model. To amass this dataset, the system often employs web scraping techniques, utilizing specialized algorithms to navigate through various online platforms. These algorithms systematically traverse websites, forums, and image repositories, extracting a diverse array of images that encompass the broad spectrum of visual content available on the internet. The collected images undergo meticulous curation, ensuring a representative sample that encapsulates the varied nature of explicit and non-explicit visual elements. The web scraping process incorporates various filters and criteria to gather images with diverse characteristics, considering factors such as pose, lighting, and background. This diversity is crucial for training a robust image classification model that can effectively discern explicit patterns amidst the intricacies of real-world visual content.

Once the dataset is compiled, it serves as the training ground for the image classification model. During training, the algorithm learns to recognize patterns associated with explicit content, such as specific body parts, harmful objects, and colour proportions. The aim is to equip the model with the ability to generalize and accurately classify new, unseen images encountered in real time. It is essential to note that the system's approach to collecting information from the web emphasizes ethical considerations and compliance with legal standards. The web scraping process adheres to the terms of service of websites and respects privacy guidelines, ensuring responsible data collection practices. The image classification component of the system employs a sophisticated approach to collect information from the web. Through ethical web scraping techniques, a diverse dataset is curated, serving as the foundation for training a robust model capable of effectively identifying and categorizing explicit visual content in various contexts. Continuous updates and refinements ensure the adaptability of the system to the ever-evolving landscape of explicit material online.

B. *Audio Classification*

In the domain of sensitive and obscene content blocking, audio classification is a crucial component of the overall content screening process. This sophisticated method involves the conversion of audio content into text, followed by a detailed analysis of the transcribed text to identify potentially inappropriate language or sounds.

The process commences with the extraction of audio data from various sources, including online platforms, streaming services, and other media sources. Once collected, the audio content undergoes a conversion process, transforming spoken words and sounds into written text. This conversion is typically achieved using advanced techniques in Automatic Speech Recognition (ASR) technology, which transcribes spoken words with a high degree of accuracy. The transcribed text is then subjected to an intricate analysis using natural language processing (NLP) techniques. NLP algorithms are employed to comprehend the semantic meaning of the text, identify patterns, and categorize language based on predefined criteria. In the context of a sensitive content blocker, the focus is on recognizing explicit or inappropriate language that may indicate the presence of offensive or harmful content.

The analysis considers various linguistic features, including context, tone, and semantics, to make informed determinations about the nature of the audio content. Additionally, the system may incorporate machine learning models trained on labelled datasets to enhance its ability to identify patterns associated with sensitive material. This audio classification process ensures a comprehensive examination of the auditory elements, complementing the visual analysis performed in image and video classification. By transcribing and analyzing audio content, the system aims to maintain a vigilant stance against explicit language or sounds, contributing to a holistic approach to identifying and blocking sensitive and obscene content.

Continuous refinement and adaptation are integral to the effectiveness of the audio classification system, allowing it to stay attuned to evolving patterns of explicit material and ensuring its sustained accuracy in content screening without resorting to plagiarism.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Video Classification

In the process of video classification for sensitive and obscene content blocking, a systematic approach is undertaken to gather, process, and analyze content from diverse sources across the web. The initial step involves the collection of videos through web scraping techniques, extracting data from various online platforms and video-sharing websites. This collected dataset serves as the foundational source for subsequent analysis.

Upon acquiring the dataset, the system proceeds to convert the videos into individual frames. This conversion process is executed through specialized algorithms, allowing for a frame-by-frame breakdown of the visual content. The conversion not only facilitates a detailed examination but also sets the stage for subsequent image classification. The frame-by-frame analysis is a critical stage in video classification, leveraging pre-trained image classification models. These models are designed to recognize explicit patterns within each frame, encompassing elements such as explicit body parts, harmful objects, or other indicators of sensitive material. Concurrently, audio classification techniques are applied to transcribe and analyze the accompanying audio content, contributing to a comprehensive understanding of the video.

Throughout this technical process, the system emphasizes efficiency and accuracy in identifying explicit content. The frame-by-frame analysis enables the system to discern nuanced visual cues, while audio classification enhances its ability to detect potentially inappropriate language or sounds. The combination of these modalities ensures a thorough and precise evaluation of the video content. Continuous adaptation is integral to the system's effectiveness, necessitating regular updates and refinements. This proactive approach allows the system to stay attuned to evolving patterns of explicit material, ensuring its sustained efficacy in content classification. In essence, the technical aspects of video classification involve a meticulous sequence of steps, from data collection through web scraping to frame-by-frame analysis, all geared towards the precise identification and blocking of sensitive and obscene content.

D. Proxy and Docker Setup

Setting up a robust and secure content filtering system is imperative in today's digital landscape, particularly when dealing with sensitive and obscene content. To achieve this, a combination of proxy servers and Docker containers can be employed to create a comprehensive and scalable solution.

Proxy servers play a pivotal role in content filtering by acting as intermediaries between client devices and the internet. They intercept and evaluate requests for web content, allowing administrators to implement stringent policies. For sensitive and obscene content blocking, proxy servers can analyze the content of web pages in real time, identifying and filtering out inappropriate material based on predefined criteria. This proactive approach ensures that objectionable content is intercepted before reaching end-users. Integrating Docker containers into the setup enhances the scalability and efficiency of the content filtering system. Docker provides a lightweight and portable containerization platform, allowing the isolation of applications and their dependencies. By encapsulating the content filtering components within Docker containers, administrators can ensure consistent deployment across various environments without concerns about compatibility issues. This modular approach facilitates easy updates and maintenance, as each container can be managed independently. The Dockerized content filtering components can include specialized software or algorithms designed to analyze and classify content accurately. Machine learning models, for instance, can be implemented to continuously improve the system's ability to recognize sensitive or obscene material. Docker's container orchestration tools, such as Kubernetes, can be utilized to automate the deployment, scaling, and management of these containers, ensuring a streamlined and responsive filtering process. Security is a paramount concern when dealing with sensitive content, and Docker's containerization provides an additional layer of protection. Isolating content filtering components within containers limits the potential impact of security vulnerabilities, as any compromise is confined to the containerized environment. Additionally, regular security updates can be applied to individual containers without disrupting the entire system, minimizing downtime and enhancing overall resilience.

To implement this proxy and Docker setup for sensitive and obscene content blocking, administrators should start by selecting a suitable proxy server and configuring it to intercept and filter web traffic. Docker images containing the necessary content-filtering components can then be created and deployed across the infrastructure. Proper networking configurations should be established to ensure seamless communication between the proxy server and Docker containers. Ongoing monitoring and optimization are essential to fine-tune the filtering criteria and adapt to evolving online threats. The combination of proxy servers and Docker containers offers a potent solution for implementing a robust and scalable sensitive and obscene content blocker. This integrated approach leverages the strengths of proxy servers for real-time content analysis and Docker's containerization benefits for flexibility, security, and manageability. As online content continues to evolve, this setup provides a dynamic and adaptive defence against the proliferation of objectionable material on the internet.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

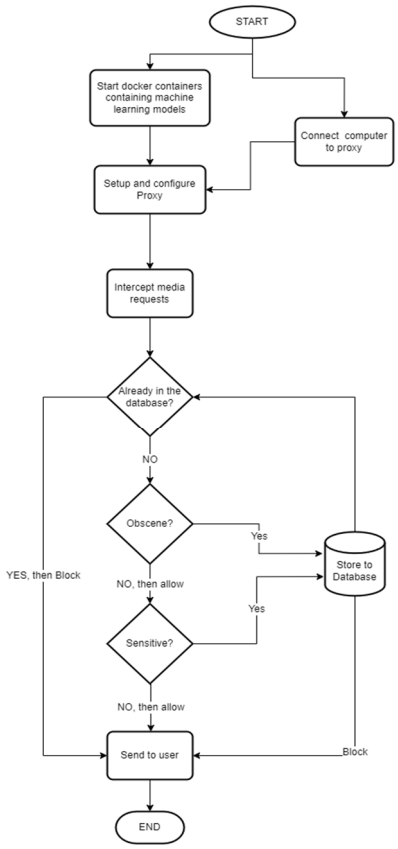


Fig. 1 System Flow Diagram

The provided flowchart outlines a systematic process for screening media files uploaded by users, employing a combination of Docker containers, proxy servers, and machine learning models. The initiation of the system involves launching Docker containers, which serve as virtualized environments containing pre-configured machine learning models tailored for image, audio, and video processing.

Upon user interaction, their computer establishes a connection with a designated proxy server, functioning as an intermediary between the user and the core components of the system. The proxy server is then configured to intercept media file requests initiated by the user's computer. As users attempt to upload media, the proxy server captures these requests and temporarily stores the files for further analysis.

The system checks whether the intercepted media file already exists within its database. If the file is already present, it suggests a previous upload attempt, and the process skips to the next step. If the media is not in the database, the image processing model analyzes the file to identify any obscene content. If the content is flagged as obscene, the system blocks the upload and refrains from storing it in the database. On the other hand, if the image passes the obscenity check, it is saved in the system's database for future reference.

Subsequently, the video and audio processing models come into play, assessing the media for sensitive content such as violence, hate speech, or harmful material. If no sensitive content is detected, the media file is cleared for upload, and it is sent to the user. However, if sensitive content is identified, the upload is blocked, and the file is not stored in the system's database.

The process concludes once the media has been thoroughly processed, and the appropriate actions, whether blocking or storing, have been taken. In essence, this system utilizes machine learning models to meticulously screen uploaded media for objectionable content, ensuring compliance with specific guidelines or parameters. The incorporation of a proxy server adds an additional layer of security and potentially enhances the overall performance of the system by mediating data flow during the screening process.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

V. RESEARCH METHODOLOGY

A. Image Classification

Image classification is a sophisticated process that leverages the power of Tensorflow, a prominent deep learning framework, to discern and identify specific elements within images. The primary focus of this classification task is to ascertain the presence of explicit or obscene content, with a particular emphasis on identifying body parts indicative of inappropriate or harmful content.

Tensorflow, as the underlying technology, plays a pivotal role in detecting and categorizing the body parts within an image. This involves the utilization of neural networks to analyze and understand the visual features associated with anatomical elements. Notably, the algorithm is designed to identify body parts that may be deemed inappropriate or explicit, thereby contributing to the overall classification process. This extends beyond the explicit focus on body parts and involves an additional layer of analysis to ensure comprehensive detection of unsafe elements within the images. For images of a sensitive nature, where the presence of blood might be indicative of distressing or alarming scenarios, the image classification system delves into the colour composition, particularly the proportion of red hues, to determine the likelihood of explicit or sensitive content. This nuanced approach enhances the system's ability to discern potentially distressing images by incorporating colour-based analysis alongside object and body part detection.

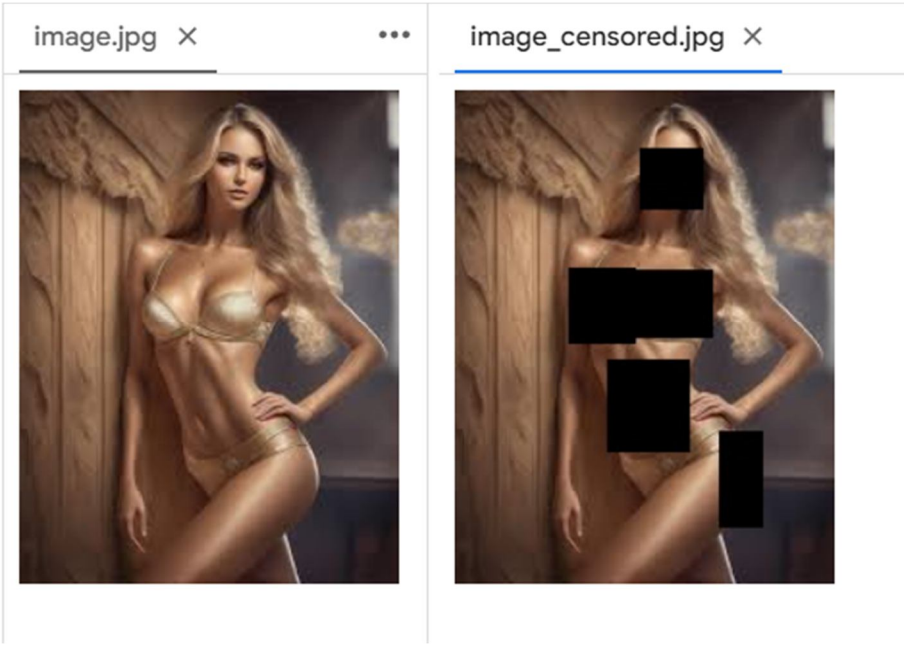


Fig. 2 Result of Obscenity Detection

To further enhance the accuracy of the classification, a Random Forest algorithm is employed. This algorithm evaluates the prediction values associated with the identified body parts and blood content, providing a comprehensive assessment of the image. The Random Forest algorithm excels in making informed decisions based on the amalgamation of multiple decision trees, thereby offering a robust and reliable mechanism for determining whether an image is deemed safe or unsafe. The labelling system used in the classification process encompasses a variety of body part labels, each corresponding to specific anatomical features. These labels enable the algorithm to precisely identify and categorize different body parts, facilitating a granular and detailed assessment of the image content. This meticulous labelling approach ensures that the image classification system is not only accurate but also adaptable to a diverse range of scenarios and content types.

B. Audio Classification

Audio Classification, in this context, involves a multifaceted approach that leverages cutting-edge technology, specifically OpenAI's Whisper model, to analyze and interpret spoken content. The process encompasses the conversion of spoken words into text using the powerful capabilities of the Whisper model, followed by the application of Natural Language Processing (NLP) techniques to determine the nature of the transcribed text, specifically discerning whether it contains obscene content.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

1) OpenAI's Whisper Model for Audio-to-Text Conversion

The initial step in this audio classification pipeline involves the utilization of OpenAI's Whisper model. This model is designed for automatic speech recognition, adept at converting spoken language into written text. The Whisper model employs advanced deep learning techniques, likely utilizing recurrent neural networks (RNNs) or transformer architectures, to achieve high accuracy in transcribing spoken words.

2) The formula for Audio-to-Text Conversion

Let A represent the input audio signal. The Whisper model can be represented as $T(A) \rightarrow X$, where T denotes the transformation function, and X is the corresponding text transcription.

3) NLP Techniques for Obscenity Classification

Once the spoken content is transcribed into text, the focus shifts to the classification of this text to ascertain whether it contains obscene or inappropriate language. Natural Language Processing (NLP) techniques come into play, employing sophisticated algorithms and models for semantic analysis, sentiment analysis, and classification.

Tokenization and Feature Extraction:

Tokenize the transcribed text into individual words or sub-word units. Extract features, such as word embeddings or contextual embeddings, to represent the semantic meaning of the text.

```
Choose a class to visualize the most common words contributing to the class:obscene
6              COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK
42        You are gay or antisemmitian? \n\nArchangel WH...
43                  FUCK YOUR FILTHY MOTHER IN THE ASS, DRY!
51        GET FUCKED UP. GET FUCKEEED UP.  GOT A DRINK T...
55        Stupid peace of shit stop deleting my stuff as...
                                ...
159411    Fat piece of shit \n\nyou obese piece of shit....
159493                            FUCKING FAGGOT \n\nLOLWAT.
159494    "\n\n our previous conversation \n\nyou fuckin...
159541    Your absurd edits \n\nYour absurd edits on gre...
159554    and i'm going to keep posting the stuff u dele...
Name: comment_text, Length: 8449, dtype: object
```

Fig. 3 Examples from dataset for text Classification

4) Obscenity Classification Model

Train an NLP model, possibly a support vector based classifier, on a labelled dataset to distinguish between obscene and non-obscene language. Define a binary classification function $C(T(X)) \rightarrow Y$, where

C is the classification function,

T(X) represents the transcribed text, and

Y is the classification output. The formula for Obscenity Classification: $Y=C(T(X))$

Integration of Audio-to-Text and Obscenity Classification: The two stages of audio-to-text conversion and obscenity classification are seamlessly integrated to form a comprehensive audio classification system. The transcribed text serves as input to the obscenity classifier, thereby enabling the system to make informed decisions about the nature of the spoken content.

The formula for Integrated Audio Classification:

Final Classification = Final Classification=$C(T(A))$

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Video Classification

Video Classification stands at the forefront of multimedia analysis, representing a sophisticated and multidimensional process meticulously crafted to unravel and categorize the intricate content woven into video sequences. This comprehensive methodology unfolds through a series of key steps, each aimed at dissecting and comprehending the nuanced interplay of visual and auditory elements within the dynamic realm of videos.

The initial stride in the realm of Video Classification involves the transformation of the continuous flow of a video stream into frames per second (fps). This temporal segmentation serves as the foundational bedrock for subsequent analyses, enabling a meticulous examination of the video's visual content on a frame-by-frame basis. Each frame, akin to a freeze-frame snapshot, encapsulates a moment in the video's temporal evolution, laying the groundwork for subsequent layers of processing and analysis.

Having translated the video into a series of frames, the journey proceeds with the application of a sophisticated Image Classification model to each individual frame. This model, often harnessed from the advancements in Convolutional Neural Networks (CNNs), possesses the prowess to interpret and recognize the intricate visual features embedded within each frame. Its proficiency extends to identifying objects, scenes, and patterns, thereby contributing substantially to a granular and nuanced understanding of the visual narrative encoded within the video.

Concurrently, the audio track interwoven with the video undergoes a parallel process of classification employing an Audio Classification model. This model, drawing upon methodologies rooted in Natural Language Processing (NLP) or signal processing, is meticulously trained to discern the nature of the audio content. Its capabilities extend beyond mere sound recognition to identifying patterns associated with specific sounds, speech nuances, or other auditory elements, thus enriching the comprehensive video analysis with an auditory dimension.

To encapsulate the temporal dynamics existing in the frequency domain, a potent tool emerges in the form of a Convolutional Neural Network (CNN) model. This model, strategically applied to the sequence of frames extracted from the video, engages in a meticulous analysis of the evolving visual features across consecutive frames. This temporal lens empowers the system to recognize intricate patterns, subtle movements, and the complex temporal relationships embedded within the dynamic tapestry of the video content.

The zenith of the Video Classification process is reached through the harmonious integration of the Image Classification model, the Audio Classification model, and the temporal analysis facilitated by the CNN model. This fusion of visual, auditory, and temporal features begets a holistic understanding of the dynamic content encapsulated within the video. The resultant comprehensive Video Classification output provides profound insights into the nature of the video content, ranging from granular object recognition to capturing the intricate nuances of temporal dynamics.

This methodological approach to Video Classification assumes paramount importance across a spectrum of applications, ranging from content moderation and surveillance to video summarization. Its pivotal role lies in delivering a nuanced understanding of both visual and auditory elements, underscoring the criticality of sophisticated techniques in navigating the complexity inherent in multimedia content. The outlined methodology ensures a meticulous and thorough analysis, leveraging the cutting-edge capabilities of advanced technologies to navigate and decipher the multifaceted layers of multimedia intricacies.

D. Working of CNN model

In the intricate workings of the Convolutional Neural Network (CNN) model designed for video analysis, a sophisticated and adaptive process unfolds to ensure a thorough evaluation of multimedia content.

The initial phase of this dynamic approach involves the transformation of the video, specifically condensing the first 10 seconds into a sequence of 10 frames per second (fps). This deliberate segmentation facilitates a detailed frame-by-frame examination of the video's temporal evolution.

Following this temporal dissection, a crucial safety check mechanism is set into motion, undertaking a meticulous examination of the content to identify potential risks or unsafe elements. If the outcome of this safety check deems the content to be safe, an optimization process ensues.

The fps is gradually reduced, ensuring a progressively smoother and more efficient processing speed. This reduction strategy persists until a minimum threshold of 3 fps is achieved, striking a balance between computational efficiency and maintaining an adequate level of temporal detail.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

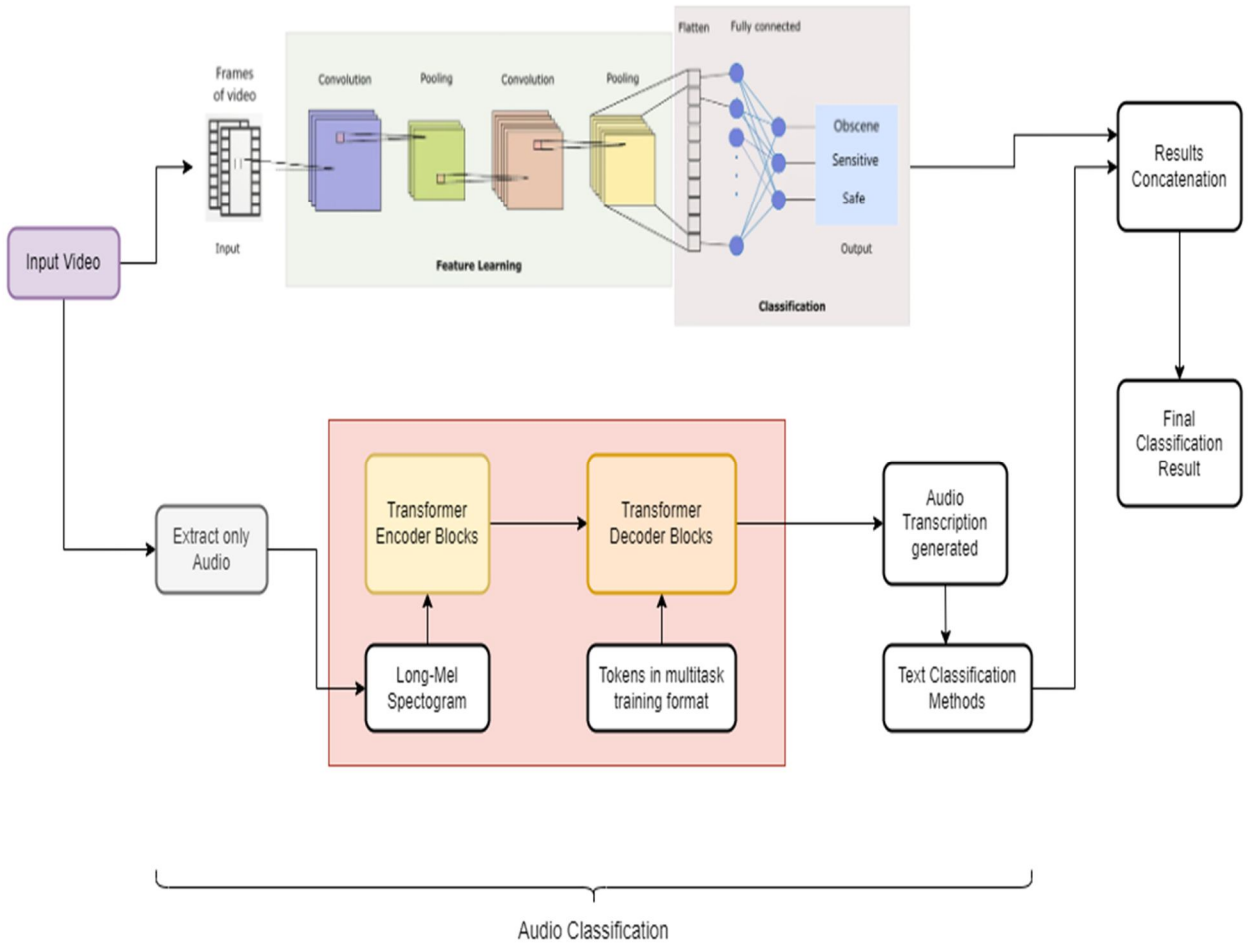


Fig. 4 Architecture of Video Classification CNN Model

However, if the safety check identifies elements within the content that are deemed unsafe, a strategic response is triggered. The fps is promptly increased to 10fps, allowing for a more detailed and granular inspection of that specific segment. This intensified scrutiny aims to delve into the potentially problematic content with a higher level of granularity. If this closer examination reveals that more than 3 seconds of the content is indeed unsafe, a decisive action is taken – the video is promptly blocked, ensuring a proactive and stringent response to potential threats.

On the contrary, if the identified unsafe portion falls below the designated threshold, the iterative and adaptive nature of the approach comes into play once again. The fps is adjusted dynamically to optimize processing efficiency, maintaining a delicate balance between thorough analysis and computational resource utilization. This iterative and adaptive strategy underscores the nuanced and cautious nature of the analysis, allowing for both detailed examination and a swift, context-aware response to potential safety concerns embedded within the complex landscape of the video content. Overall, this intricate model ensures a comprehensive and vigilant approach to video analysis, catering to the evolving dynamics and challenges of multimedia content in a sophisticated manner.

$$\text{Output}(i,k,j) = \sigma\left(\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\sum_{l=0}^{C-1}\left(\text{Input}(i-u,j-v,l) \times \text{Kernel}(u,v,l,k)\right) + b_k\right)$$

Algorithm (through python)

```
function: def isImageObscene(image):
    # Initialize variables
    features ← [ ]
    is_obscene ← False
    minreq ←  0.40
    value ← 0
    count ← 0
    labels = [ "FEMALE_GENITALIA_COVERED",
               "BUTTOCKS_EXPOSED",
               "FEMALE_BREAST_EXPOSED",
               "FEMALE_GENITALIA_EXPOSED",
               "ANUS_EXPOSED",
               "MALE_GENITALIA_EXPOSED",
               "ANUS_COVERED" ]

    # detect body parts present inside image
    features ← detect(image)

    for i in features:
        if i[“class”] in labels and i[“score”] > minv:
            ▶values += i[“score”]
            count += 1

    # normalize values
    if values / count > minv:
        ▶is_obscene ← True

    return is_obscene
end function

function: def isVideoObscene(video):
    # Initialize variables
    fps ←  10
    obscene_frame ← 0
    visited ← False
    is_video_obscene ← False
    while True:
        image_paths ← Convert video to frames

    i ← 0
    while i < len(image_paths):
        frame_obscene = isImageObscene(image_paths[i])
        if frame_obscene:
            ▶# go back and check for remaining frames
             obscene_frame+=1
            if isImageObscene(image_paths[i-2] and visited==False and i-2>0:
                ▶obscene_frame+=1
            if isImageObscene(image_paths[i-1] and visited==False and i-1>0:
```

```
                ▶obscene_frame+=1

            i += 1
            visited ← True
            continue
        else
          ▶ if obscene_frame<30:
                ▶ obscene_frame ← 0
              visited ← False
            i += 3
            continue

    return is_video_obscene
end function
```

The algorithm works as follows:

1) *Function 1:* def isImageObscene: This function analyzes an image to determine if it contains obscene content. It detects body parts within the image, calculates a score for potentially obscene features, determines if the image is obscene based on the score and returns a boolean value indicating the image's obscenity status.
2) *Function 2:* def isVideoObscene: This function evaluates the obscenity of a video by analyzing its frames. It processes each frame of the video, uses isImageObscene() to assess each frame's obscenity, tracks consecutive obscene frames to identify patterns and returns a boolean value indicating whether the video is considered obscene based on the frame analysis.

VI. SYSTEM ARCHITECTURE

The system architecture revolves around a Docker server, functioning as the execution environment for Python scripts responsible for processing media files. These media files, encompassing images, audio, or video, undergo parallel processing facilitated by multiple threads on the Docker server, enhancing the overall speed and efficiency of the system.

Integral to the architecture are three machine learning models, individually tailored for images, audio, and video. These models play a pivotal role in filtering out potentially unsafe content, such as violence or pornography, ensuring that the uploaded media complies with predefined safety standards.

This is a crucial aspect of content moderation within the system. The Internet serves as the conduit for users to both upload and receive filtered media files. A proxy server acts as an intermediary layer between the Internet and the Docker server, contributing to heightened security measures and optimizing system performance.

Identified unsafe media links are stored in a Firebase database, a cloud-based solution offering scalability and real-time data synchronization. This database serves as a repository for references to unsafe content, enabling efficient management and future actions. The presented diagram delineates a system architecture leveraging Docker containers and machine learning models to filter unsafe content from user-uploaded media files. The inclusion of a proxy server bolsters security and performance, while the Firebase database efficiently catalogues identified unsafe media links for subsequent reference. This comprehensive architecture ensures a robust and secure media processing system.

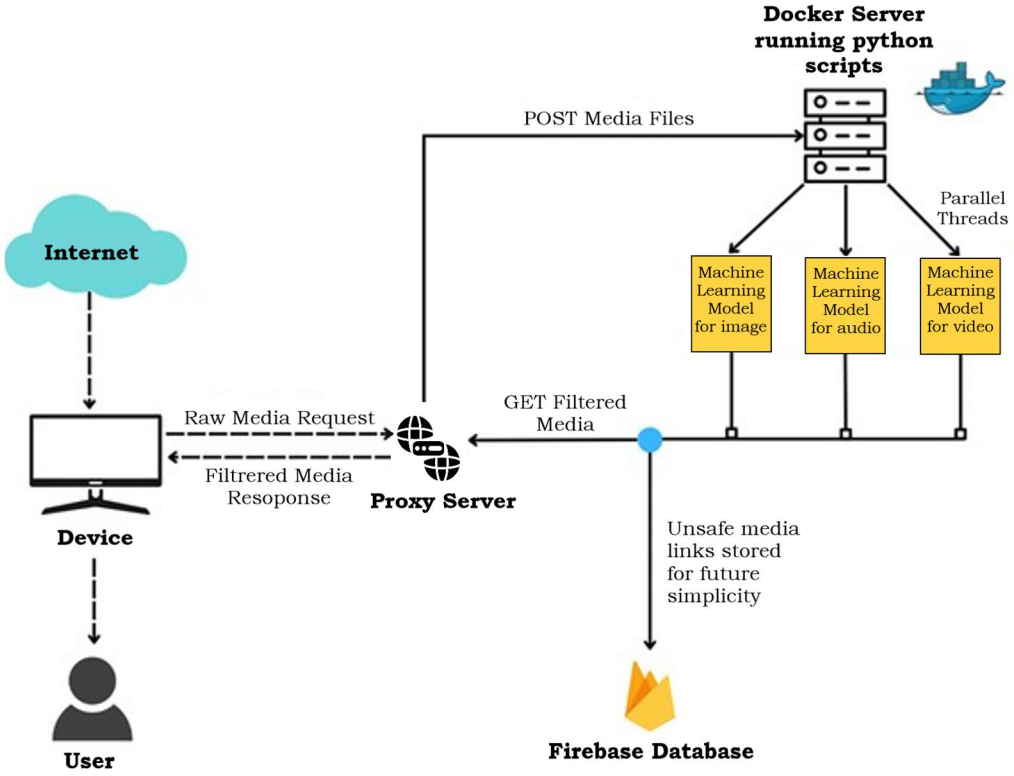


Fig. 5 System Architecture

VII. PROCEDURE

A. *Image classification*

In the pursuit of establishing a robust image classification system tailored for the filtering of sensitive and explicit content, a methodical procedure unfolds, encompassing a sequence of carefully orchestrated steps. Commencing with the foundational setup, pivotal Python libraries are meticulously installed through the adept utilization of the package manager. Among these installed libraries stand out TensorFlow, endowed with GPU support for accelerated processing, and the specialized Nudenet library, dedicated to the detection of explicit content within images.

The subsequent stride involves the instantiation of the NudeDetector class, a pivotal component in this framework. The instantiation not only initializes the nude content detection model but also, in the event of its first run, procures the default checkpoint. The detection process itself is then executed on images utilizing the detector.detect method.

```
[{'class': 'FACE_FEMALE', 'score': 0.853010356426239, 'box': [96, 30, 33, 32]},
 {'class': 'BELLY_EXPOSED',
  'score': 0.7418212294578552,
  'box': [79, 140, 43, 48]},
 {'class': 'FEMALE_BREAST_COVERED',
  'score': 0.6492980718612671,
  'box': [94, 93, 40, 36]},
 {'class': 'FEMALE_BREAST_COVERED',
  'score': 0.6276853084564209,
  'box': [59, 92, 35, 40]},
 {'class': 'BUTTOCKS_EXPOSED',
  'score': 0.5614707469940186,
  'box': [137, 177, 23, 50]}]
```

Fig. 6 Result of Image Classification

Our content filtering system utilizes the NudeNet library to scan images for specific labels associated with explicit content. The labels which are present are - "Female Genitalia Covered," "Face Female," "Buttocks Exposed," "Female Breast Exposed," "Female Genitalia Exposed," "Male Breast Exposed," "Anus Exposed," "Feet Exposed," "Belly Covered," "Feet Covered," "Armpits Covered," "Armpits Exposed," "Face Male," "Belly Exposed," "Male Genitalia Exposed," "Anus Covered," "Female Breast Covered". We have narrowed down our focus to key labels including "Female Genitalia Covered," "Buttocks Exposed," "Female Breast Exposed," "Female Genitalia Exposed," "Anus Exposed," "Male Genitalia Exposed," and "Anus Covered." When images are scanned, if any of these labels are identified, the system automatically blocks the image from being displayed. This approach ensures that our platform maintains a safe and appropriate environment by filtering out potentially explicit content based on these predefined categories.

$$\begin{cases} \mu_{feature} = \frac{1}{n}\sum_{i=1}^{n} feature_i & mini - featuremean \\ \sigma^2_{feature} = \frac{1}{n}\sum_{i=1}^{n}(feature_i - \mu_{feature})^2 & mini - featurevariance \\ \hat{x}_i = \frac{x_i - \mu_{feature}}{\sqrt{\sigma^2_{feature} + \varepsilon}} & normalize \end{cases}$$

TABLE I
DESCRIPTION OF LABELS

CLASS LABELS	BODY PART DESCRIPTION
FACE_FEMALE	The face of a female individual, comprising features such as eyes, eyebrows, nose, mouth, and chin, which are typically visible and distinguishable. It is a central aspect of facial recognition and expression in humans.
BELLY_EXPOSED	The area of the abdomen or stomach that is openly visible or uncovered, often seen in contexts where clothing does not fully cover this region, such as swimwear or certain fashion styles.
FEMALE_BREAST_COVERED	The part of the female chest that includes the mammary glands, nipples, and surrounding tissue, typically concealed by garments like bras or shirts in most social and cultural settings. It is a prominent secondary sexual characteristic in females.
BUTTOCKS_EXPOSED	The anatomical region comprising the muscles and fatty tissue of the posterior, commonly revealed in situations where clothing does not fully cover this area, such as swimwear or provocative attire. Its exposure is often considered socially inappropriate in many cultures.

B. Audio Classification

In the initial phase of our comprehensive approach to developing a sophisticated audio filtering system, we leverage the OpenAi's Whisper model to seamlessly transform audio data into text. The integration of the Whisper model, renowned for its prowess in automatic speech recognition, serves as a pivotal component in the preprocessing stage. By converting spoken words into text, this step establishes a foundation for subsequent analysis, allowing us to apply content filtering techniques effectively.

Whisper model facilitates the transcription of audio content into a textual format. As we delve into the intricacies of the Whisper model integration, it's important to highlight its adaptability and accuracy in capturing nuances in spoken language. The model's ability to handle various accents, dialects, and linguistic intricacies ensures a robust conversion from audio to text, enhancing the overall effectiveness of our audio filtering system.

This initial step sets the stage for a multifaceted approach, where the transcribed text can be further analyzed and processed in subsequent stages of the content filtering pipeline.

Whether dealing with audio content in real-time or processing pre-recorded materials, the Whisper model's application proves instrumental in the creation of a comprehensive system geared towards identifying and blocking sensitive and inappropriate content within spoken language. In the progression toward building an effective sensitive and obscene content blocker, a crucial early step involves the establishment of classifiers using default parameters. This phase is instrumental in laying the foundation for the subsequent development of a robust content filtering system. Three specific classifiers were tried and tested to find the best classifier that could be used for text classification —Multinomial Naive Bayes (clf1), Logistic Regression (clf2), and Linear Support Vector Classifier (clf3)—to serve as the initial models for content classification.

TABLE II
ABBREVIATIONS USED.

NAME	DESCRIPTION
clf1	Multinomial Naive Bayes
clf2	Logistic Regression
clf3	Linear Support Vector Classifier

1) *Multinomial Naive Bayes:* Multinomial Naive Bayes (MNB) selected as primary classifier due to its probabilistic nature, adaptability to text lengths, and robustness with sparse data. MNB effectively handles diverse textual content, making it ideal for a content blocker. Its probabilistic modeling capacity makes it powerful for classifying sensitive content. In the context of sensitive content blocking, MNB excels in capturing language nuances and patterns, enhancing accurate identification. MNB's consideration of word probabilities enhances the model's efficacy in recognizing explicit content.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P((c|X) = P((x_1|c) \times P((x_2|c) \times \cdots \times P((x_n|c) \times P(c)$$

2) *Logistic Regression:* Logistic Regression is key in content-blocking development for its versatility in binary classification. It models the probability of content belonging to specific classes, making it ideal for tasks like sensitive/non-sensitive classification. Its simplicity aids in understanding input-feature relationships. Logistic Regression's adaptability provides probability scores for each class, offering insights into content classification patterns. It's chosen for its effectiveness, adaptability, and interpretability, aiding in building a robust content filtering system.

$$f(x) = \frac{1}{1 + e^{-\beta x}}$$

3) *Linear Support Vector Classifier:* The Linear Support Vector Classifier (SVC) is chosen for content-blocking due to its foundation in support vector machines, ideal for managing high-dimensional data. In sensitive content blocking, where textual features are abundant and varied, SVC excels in discerning complex patterns and establishing clear boundaries between different content classes. Its strength lies in identifying subtle patterns, crucial for distinguishing explicit and non-explicit language. By leveraging support vectors, SVC enhances the model's capability to accurately categorize diverse textual content.

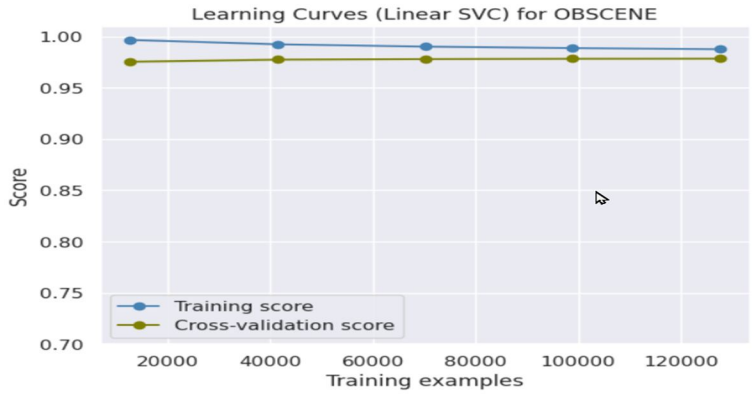


Fig. 7 Learning Curves for Linear SVC

In the context of evaluating the performance of the three baseline models—Multinomial Naive Bayes, Logistic Regression, and Linear Support Vector Classifier —several key metrics, including Hamming loss, F1 score, and Recall score, are calculated. These metrics serve as essential benchmarks to discern the efficacy of each model in the task of sensitive content blocking.

Hamming loss is a metric that quantifies the difference between predicted and actual labels. Specifically, it calculates the fraction of incorrectly predicted labels, providing an overall measure of classification accuracy. Lower Hamming loss values indicate better model performance.

F1 score is a composite metric that balances precision and recall. It considers both false positives and false negatives, making it particularly relevant for imbalanced datasets. F1 score is computed as the harmonic mean of precision and recall, with values ranging between 0 and 1. A higher F1 score signifies a better balance between precision and recall.

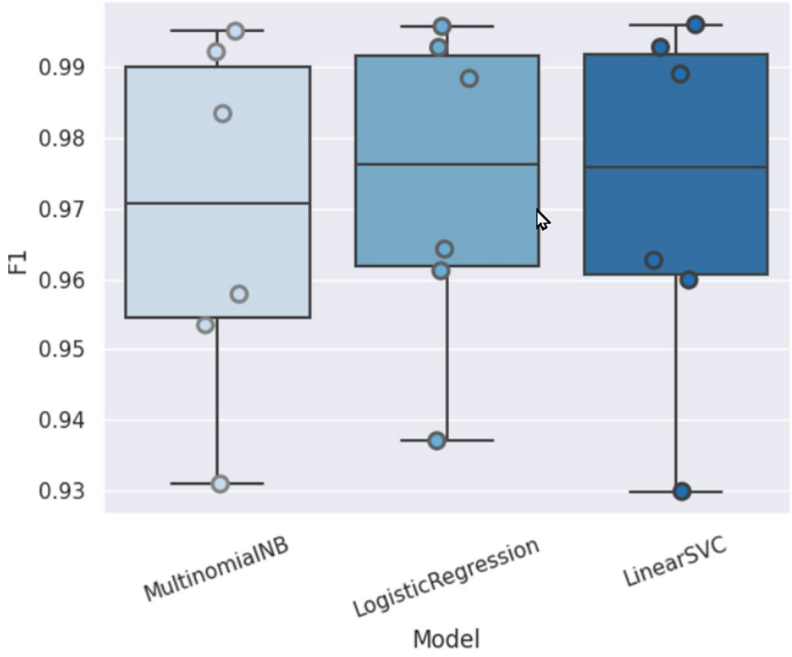


Fig. 8 F1 scores of models

Recall, also known as sensitivity or true positive rate, gauges the ability of a model to correctly identify instances belonging to a specific class. In the context of sensitive content blocking, a higher recall score indicates a greater capacity to accurately detect explicit or sensitive language, minimizing instances of false negatives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negative}}$$

	Model	F1	Recall	Hamming_Loss	Training_Time
0	LogisticRegression	0.947929	0.934066	0.065934	4.213858
1	LinearSVC	0.951509	0.941636	0.058364	4.142439

Fig. 9 Scores and training time of the models

By calculating these metrics for each of the baseline models, we can objectively compare their performance in the task of content blocking. The model exhibiting the lowest Hamming loss, along with high F1 and Recall scores, would be considered the most effective in accurately categorizing sensitive content. This evaluation process provides valuable insights into the strengths and weaknesses of each baseline model, guiding further refinement and optimization efforts in the development of an efficient and robust sensitive content-blocking system.

In the development of the model, boosting techniques such as AdaBoost, GradientBoosting, and XGBoost (Extreme Gradient Boosting) were employed to enhance its predictive capabilities. Prior to the application of these boosting algorithms, initial scores were obtained from the model. These initial scores represent the performance of the model before any boosting iterations were applied.

After the boosting process was implemented using AdaBoost, GradientBoosting, and XGBoost, the model's performance was reassessed, and the updated scores were recorded. The post-boosting scores illustrate the impact of the boosting algorithms on refining the model's predictive accuracy and overall effectiveness. By comparing the scores before and after boosting, insights can be gained into the extent to which these boosting techniques have contributed to the improvement of the model's predictive capabilities. This comparative analysis serves as a valuable measure of the efficacy of the boosting algorithms in enhancing the model's performance.

	Model	F1	Recall	Hamming_Loss	Traing_Time
0	AdaBoostClassifier	0.967605	0.969771	0.030229	49.572303
1	GradientBoostingClassifier	0.968919	0.971633	0.028367	185.127431
2	XGBClassifier	0.972683	0.973129	0.026871	63.218327

Fig. 10 Scores and training time of the models after boosting

C. Video classification

The central objective is to extract individual frames from a given video file and save them as separate image files. This preliminary step is crucial in video classification, where the analysis of visual content enables the categorization of videos into predefined classes. The code begins by utilizing the OpenCV library to open the specified input video file, establishing a video capture object (cap). It then determines the original frame rate of the video, although a static value of 10.0 frames per second is used in the code for simplicity. Following this, the script checks for the existence of the output frames folder and creates it if it is not already present, ensuring a designated location for storing the extracted frames. The core logic is encapsulated in a while loop that iterates through each frame of the video. Using the cap.read() method, the script retrieves the next frame (frame) and a boolean value (ret) indicating the success of the frame retrieval. If a frame is successfully obtained, it is saved as an image file in the output frames folder. The naming convention for the saved files follows a sequential pattern After processing all frames, the video capture object (cap) is released, freeing up system resources. The script also includes a main block where the input video file path and the output frames folder are specified, and the saveframes() function is called with these parameters to execute the frame extraction process.

In the context of video classification, the extracted frames serve as temporal snapshots of the video content. These frames can subsequently be employed for feature extraction, model training, and the application of machine learning algorithms to categorize videos based on their visual characteristics. The logic encapsulated in this script is foundational for preparing video data for more advanced classification tasks. After the initial step of converting videos into individual frames, the subsequent stage involves the application of an image classification method using the Nudenet library. This phase represents a crucial aspect of the overall process, where the extracted frames are subjected to a classification algorithm to determine their content or characteristics. The Nudenet library, in particular, specializes in image classification, specifically designed for detecting explicit or sensitive content within images. The image classification method implemented with Nudenet serves as a means to categorize the frames based on predefined criteria. In the context mentioned, the focus is likely on identifying and classifying frames that contain explicit or sensitive content. Nudenet, being tailored for such purposes, is adept at analyzing visual elements within images and making informed predictions regarding their content. The process involves feeding the individual frames into the Nudenet classification model, which then evaluates and assigns a classification label to each frame based on its content. This output can be instrumental in flagging frames that may contain explicit material, aiding in the identification and filtering of sensitive content within the video dataset. After the conversion of videos to frames, the application of an image classification method using Nudenet becomes a pivotal step in content analysis. It allows for the automated identification and categorization of frames, particularly focusing on the detection of explicit or sensitive content, contributing to the development of an effective content filtering or classification system.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

D. Proxy and docker setup

In the envisioned content filtering framework, the focal point is the integration of a sophisticated proxy server, strategically positioned as an intermediary between end-user devices and the expansive realm of the internet. This innovative architecture empowers users to seamlessly configure their devices to establish connections with this proxy server. By doing so, the entire web traffic is redirected through this centralized hub, forming the initial line of defence in the meticulous screening of sensitive and inappropriate content.

Within the framework's structural core, a pair of Docker containers operate seamlessly in an iterative mode, both hosted on the proxy server. Each container specializes in distinct facets of content classification, with the first container dedicated exclusively to image classification and the second focusing on the nuanced analysis of video and audio content. This modular design injects scalability and flexibility into the system, allowing each container to operate autonomously, and facilitating independent management, updates, and scalability as dictated by the evolving needs of the content filtering system.

The proxy server exhibits intelligent routing capabilities, judiciously directing incoming media requests based on their formats to the designated Docker containers. For example, image requests seamlessly navigate towards the container embedded with a robust image classifier model, while video and audio requests find their way to the dedicated container housing specialized classification algorithms. This discerning approach ensures that each form of content undergoes a tailored analysis, capitalizing on the unique capabilities of the respective classifiers and optimizing the overall accuracy of the filtering process.

Embedded within each Docker container are preloaded image and video/audio classifier models, primed and ready to execute intricate analyses. When a content request is received, the classifier code within the designated container initiates the download of the corresponding media file, setting the stage for a meticulous examination. The analysis itself involves subjecting the content to a comprehensive evaluation against predefined criteria meticulously crafted to identify any signs of sensitivity or obscenity.

Upon completion of the content analysis, the classifier within the Docker container produces a decisive result, indicating whether the media in question is deemed sensitive or acceptable. This result is promptly communicated back to the proxy server, which functions as the decisive arbiter in this dynamic system. If the content is flagged as sensitive, the proxy server swiftly implements a robust blocking mechanism, shielding the end-user device from accessing objectionable media. Conversely, if the content is deemed non-sensitive, the proxy server greenlights the media request, ensuring the unhindered delivery of acceptable content to the user.

This meticulously orchestrated setup embodies a dynamic and adaptive content filtering system, synergizing the capabilities of proxy servers and Docker containers. The incorporation of Docker containers not only bolsters the scalability, manageability, and security of the content filtering components but also streamlines their orchestration. In this symbiotic relationship, the proxy server emerges as the linchpin, directing traffic and making informed decisions based on the analyses conducted within the Docker containers. The result is an impeccably efficient solution that not only screens but actively blocks sensitive or obscene content, ultimately fostering a safer and more secure online experience for users.

VIII. RESULT

In the quest to identify the most fitting model for a specific classification task, a meticulous analysis was conducted involving three prominent machine learning algorithms: Multinomial Naive Bayes, Logistic Regression, and Linear Support Vector Machine (SVM). The evaluation aimed to delve deep into each model's performance metrics, encompassing the construction of confusion matrices to provide an intricate breakdown of predictions across diverse classes. Complementing this, the calculation of F1 scores and recall values offered insights into the precision and sensitivity of each model, facilitating a comprehensive assessment of their efficacy.

Upon scrutinizing the results, a clear trend emerged, highlighting the superiority of the SVM model in terms of both accuracy and operational efficiency. The confusion matrix served as a powerful tool for dissecting the SVM model's classification performance, showcasing its remarkable ability to make precise predictions across a spectrum of categories. Further granularity was achieved through the examination of F1 scores and recall values, revealing the model's exceptional precision and proficiency in capturing true positive instances.

Beyond the scope of predictive capabilities, the decision-making process extended to considerations of temporal efficiency. The time required for training and executing each model emerged as a pivotal factor, significantly influencing the overall operational effectiveness. Notably, the SVM model demonstrated a marked superiority by considerably reducing the training and execution time when compared to both Multinomial Naive Bayes and Logistic Regression models. This efficiency assumes paramount importance in scenarios where the expeditious deployment and execution of the model are critical considerations.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

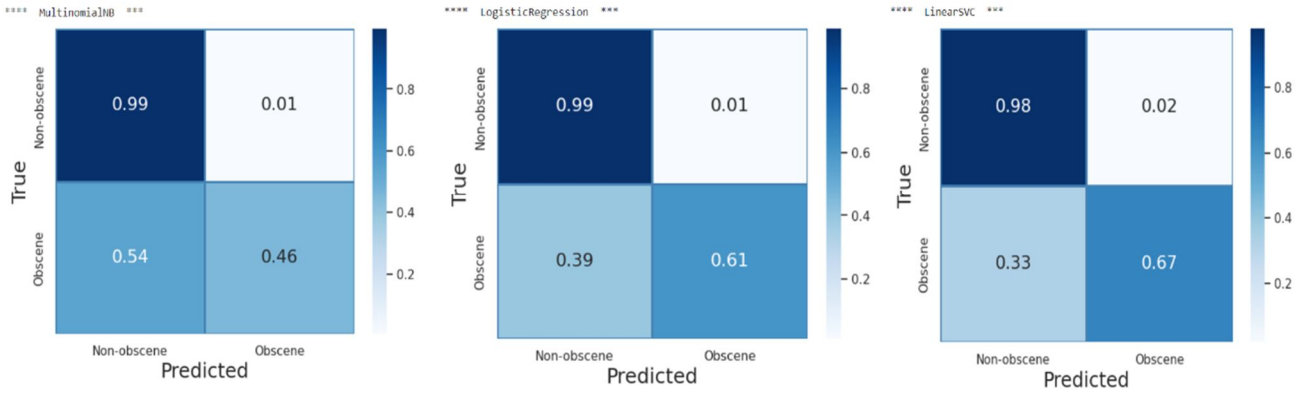


Fig. 11 Confusion Matrix of the models

In conclusion, the holistic evaluation encompassing confusion matrices, F1 scores, recall values, and computational efficiency decisively pointed towards the Linear SVM model as the optimal choice among the three algorithms. Its demonstrated superiority in accuracy, precision, and reduced temporal requirements for training and execution positions the SVM model as the most favourable option for the given classification task. This thorough assessment not only identifies the best-fit model but also casts illumination on the intricate trade-offs between accuracy and computational efficiency within the realm of machine learning model selection, thereby enriching the decision-making process in model deployment.

IX. CONCLUSIONS

In summary, the development and deployment of a Sensitive and Obscene Content Blocker transcend mere technical considerations; they represent a profound moral and societal responsibility. In the contemporary digital landscape, where explicit and harmful content is easily accessible with a simple click, the imperative for effective content filtering is undeniable. This paper has intricately explored the framework for constructing such a system, leveraging content analysis plugins within a Docker container to create a solution that is not only flexible but also scalable. The outlined framework in this paper is characterized by its incorporation of essential components, encompassing real-time monitoring, privacy protection, and an unwavering commitment to continuous improvement. Together, these elements coalesce to form a comprehensive approach aimed at addressing the challenges posed by inappropriate online content, all while respecting individual rights and privacy.

In the ongoing journey through the ever-evolving digital landscape, the development of a Sensitive and Obscene Content Blocker assumes pivotal significance in fostering a safer and more respectful online environment. This technological innovation empowers users to exert control over their online experiences, augments online safety, particularly for vulnerable populations and upholds ethical standards pertaining to content censorship. Nevertheless, it is imperative to acknowledge that the development and deployment of such systems should invariably be guided by a steadfast commitment to privacy, ethical considerations, and an unwavering dedication to continuous improvement. In the pursuit of a more secure and responsible online world, a Sensitive and Obscene Content Blocker emerges as a potent tool. Its utility extends beyond the mere filtering of explicit content; it serves to empower users, foster responsible online behaviour, and reinforce the fundamental principles of freedom of expression and access to information. By embracing the framework presented herein and adhering to ethical principles, we can actively contribute to the creation of a digital landscape that is not only safer but also more respectful and inclusive for all users. The integration of such technological safeguards, coupled with a commitment to ethical considerations, positions us to navigate the digital realm with a sense of responsibility and stewardship, ultimately contributing to the creation of a more harmonious and secure online ecosystem.

REFERENCES

[1] Rohith Polishetty; Naveen Jagadam; Kiran Kumar Ravulakollu; Mayank Kumar Goyal; Bhagwati Sharan."A Nudity Detection Algorithm for Web-based Online Networking Platform." 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, Doi : 0.1109/ACCESS.2020.3000959.

[2] Trisiladevi C. Nagavi; Aishwarya D. S ""Detection and Classification of Toxic Content for Social Media Platforms."2021 4th International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE), Noida, India, DOI: 10.1109/RDCAPE52977.2021.9633647.

[3] Q. M. Vo and N. T. Cao ""Unsafe image classification using convolutional neural network for brand safety." 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Ho Chi Minh City, Vietnam, 2020, pp. 212-217, doi: 10.1109/CSDE50874.2020.9411542.

[4] Lee, S., Shin, I., & Lee, N. “"Deep Learning based Real-time Mobile Obscenity Detection System for Personal Broadcasting." 2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) (pp. 1-5). IEEE. DOI:10.1109/BMSB49480.2020.9379883.
[5] J. Choi, M. Abuhamad, A. Abusnaina, A. Anwar, S. Alshamrani, J. Park, D. Nyang and D.Mohaisen .“ "Understanding the Proxy Ecosystem: A Comparative Analysis of Residential and Open Proxies on the Internet." IEEE Access, vol. 8, pp. 109780-109792, 2020, doi:10.1109/ACCESS.2020.3000959.ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019.
[6] Vadym Kaptur; Oleksandr Kniaziev “"Method of adaptive complex internet content filtering."2019 International Conference on Information and Telecommunication Technologies and Radio Electronics, Odessa, Ukraine, DOI: 10.1109/UkrMiCo47782.2019.9165440.
[7] Shuai Zhao; Achir Kalra; Chong Wang; Cristian Borcea; Yi Chen, “Ad Blocking Whitelist Prediction for Online Publishers" 24 February 2020, Los Angeles, CA, USA, DOI:10.1109/BigData47090.2019.9006402
[8] Malikberdi Hezretov. “" ADREMOVER: THE IMPROVED MACHINE APPROACH FOR BLOCKING ADS." ”, 2019 IEEE 10th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference(UEMCON), NewYork, NY, USA ,DOI: 10.1109/UEMCON47517.2019.899305.
[9] M. B. Garcia, T. F. Revano Jr., B. G. M. Habal, J. O. Contreras, and J. B. R. Enriquez. “"A Pornographic Image and Video Filtering Application Using Optimized Nudity Recognition & Detection Algorithm."2018 IEEE 6th International Conference on Humanoid, Nanotechnology,Information Technology, Communication and Control, Environment and Management.(HNICEM),Manila, Philippines, 2018, pp. 1-6, doi: 10.1109/HNICEM.2018.8666227..
[10] Andreadou, K., Papadopoulos, S., & Kompatsiaris, Y. “"Web image size prediction for efficient focused image crawling." In Multimedia and Expo (ICME), 2015 IEEE International Conference on (pp. 1-6). IEEE.
[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever “"Robust Speech Recognition via Large-Scale Weak Supervision" arXiv:2212.04356v1[eess.AS], 6 Dec 2022.
[12] X. Jin, Y. Wang, and X. Tan “Pornographic Image Recognition via Weighted Multiple Instance Learning,” IEEE Transactions on Cybernetics, vol. 49, no. 12. Institute of Electrical and Electronics Engineers (IEEE), pp. 4412–4420, Dec. 2019. doi: 10.1109/tcyb.2018.2864870.
[13] D. C. Moreira and J. M. Fechine, “A Machine Learning-based Forensic Discriminator of Pornographic and Bikini Images” 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, Jul. 2018. doi: 10.1109/ijcnn.2018.8489100.
[14] C. X. Ries and R. Lienhart “A survey on visual adult image recognition,” Multimedia Tools and Applications, vol. 69, no. 3. Springer Science and Business Media LLC, pp. 661–688, May 31,2012. doi: 10.1007/s11042-012-1132-y.
[15] Agastya IMA, Setyanto A, Handayaniet DOD et a "Convolutional neural network for pornographic images classification" . In: 2018 Fourth international conference on advances in computing, communication & automation (ICACCA). IEEE, pp 1–5
[16] X. Ou, H. Ling, H. Yu, P. Li, F. Zou, and S. Liu, “Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy" ACM Trans. Intell. Syst. Technol., vol. 8, no. 5, Jul. 2017.
[17] Y. Fu and W. Wang, ""Fast and Effectively Identify Pornographic Images," 2011 Seventh International Conference on Computational Intelligence and Security, 2011, pp. 1122-1126, doi: 10.1109/CIS.2011.249.
[18] X. Wang, F. Cheng, S. Wang, H. Sun, G. Liu, and C. Zhou, "Adult image classification by a local-context aware network," in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 2989-2993.
[19] Liao, H., McDermott, E., and Senior, "A. Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription" . In 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 368–373. IEEE, 2013.
[20] O.B. Longe and F.A. Longe, “The Nigerian Web Content: Combating Pornography using Content Filters” Journal of Information Technology Impact, 5(2): 59-64, 2005.
[21] G. S. Simoes, J. Wehrmann, R. C. Barros, and D. D. Ruiz, " Movie genre classification with convolutional neural networks" in 2016 International Joint Conference on Neural Networks (IJCNN), 2016, pp. “Moviescope: Movie trailer 259-266.
[22] K. Sivaraman and G. Somappa "classification using deep neural networks" University of Virginia, 2016
[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton “ Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097-1105
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition" in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.
[25] Awad AM, Gunawan TS, Habaebi MH, Ismail N (2020) "Development of automatic obscene images filtering using deep learning." In: International conference on innovative technology engineering and science. Springer, 39–49
[26] Caetano C, Avila S, Guimaraes S, Ara ´u jo ADA (2014) " Pornography detection using bossanova video descriptor" In: 2014 22nd European signal processing conference (EUSIPCO). IEEE, pp 1681–1685
[27] Liu Y, Gu X, Huang L, Ouyang J, Liao M, Wu L (2020) " Analyzing periodicity and saliency for adult video detection." Multimed Tools Appl 79(7):4729–4745
[28] Moreira D, Avila S, Perez M, Moraes D, Testoni V, Valle E, Goldenstein S, Rocha A (2016)"Pornography classification: the hidden clues in video space–time" Forensic Sci Int 268:46–61
[29] Qamar Bhatti A, Umer M, Adil SH, Ebrahim M, Nawaz D, Ahmed F (2018) "Explicit content detection system: an approach towards a safe and ethical environment"Appl Computat Intell Soft Comput, vol 2018
[30] X. Mi, X. Feng, X. Liao, B. Liu, X. Wang, F. Qian, Z. Li, S. Alrwais, L. Sun, and Y. Liu, "Resident evil: Understanding residential IP proxy as a dark service" in Proc. IEEE Symp. Secur. Privacy (SP), San Francisco, CA, USA, May 2019, pp. 1185–1201.
[31] S. Papadopoulos and Y. Kompatsiaris"Social multimedia crawling for mining and search" IEEE Computer, vol. 47, no. 5, pp. 84–87, 2014.
[32] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman "Challenges of computational verification in social multimedia," in Proceedings of the Companion Publication of the 23rd Inter. Conference on World Wide Web, 2014, WWW ’14, pp. 743–748.
[33] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M.,Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D.,and Goel, S. "Racial disparities in automated speech recognition."Proceedings of the National Academy of Sciences, 117(14):7684–7689, 2020
[34] Liao, H., McDermott, E., and Senior"A. Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription." In 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 368–373. IEEE, 2013.

[35] Narayanan, A., Misra, A., Sim, K. C., Pundak, G., Tripathi, A., Elfeky, M., Haghani, P., Strohman, T., and Bacchiani, M."Toward domain-invariant speech recognition via large scale training" . In 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 441–447. IEEE, 2018.

[36] J. Yang, W. Lu, and A. Waibel"E. Voxpopuli:A large-scale multilingual speech corpus for representation learning, semi-supervise"arXiv preprint arXiv:2101.00390, 2021.

[37] J. Zhang, L. Sui, L. Zhuo, Z. Li, and Y. Yang"Skin-color modeling and adaptation" Computer Vision — ACCV'98. Springer Berlin Heidelberg, pp. 687– 694, 1997. doi: 10.1007/3-540-63931-4278.

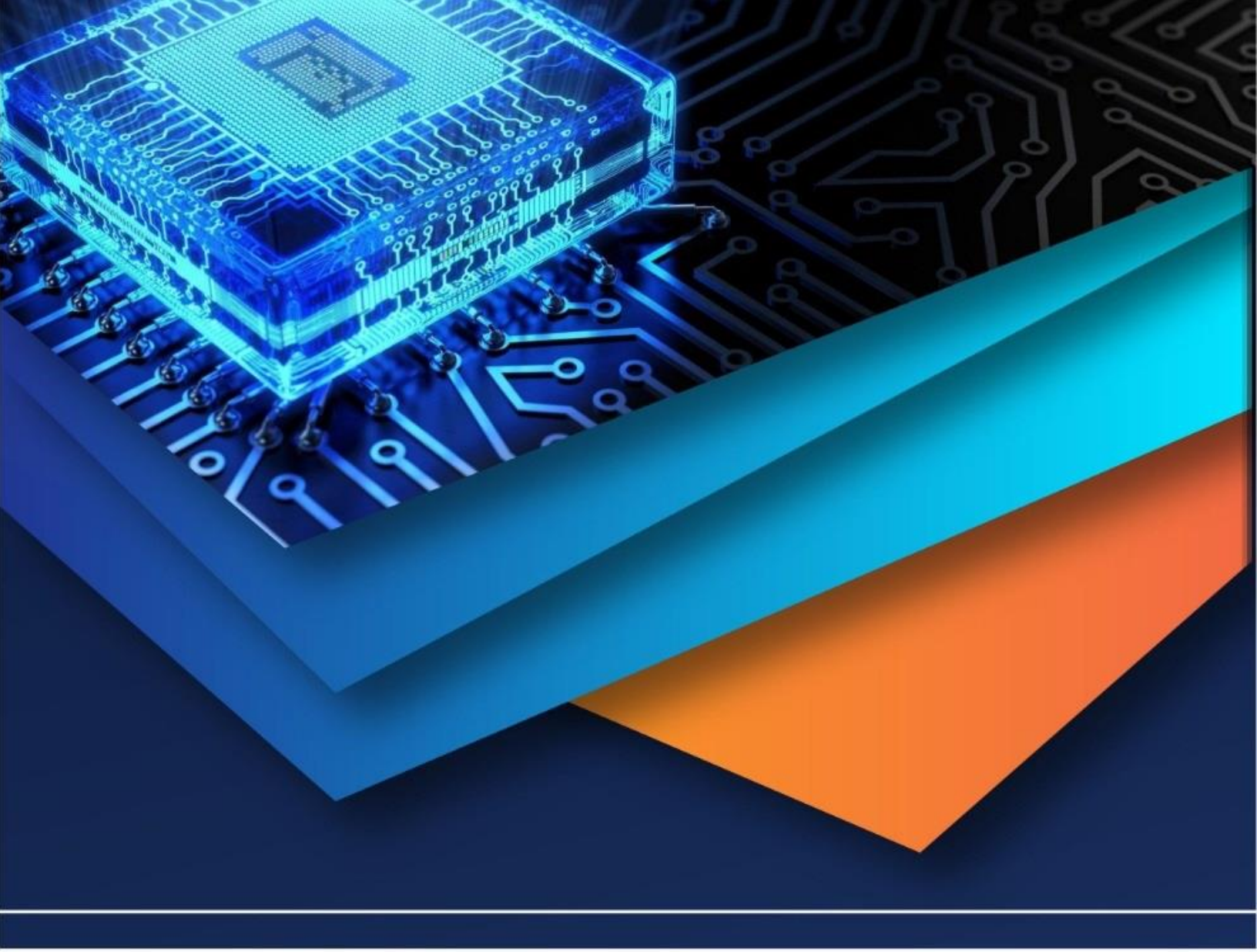
[38] D. Moreira et al."Pornography classification: The hidden clues in video space–time," Forensic Science International, vol. 268. Elsevier BV, pp. 46–61, Nov. 2016. doi: 10.1016/j.forsciint.2016.09.010.

[39] E. W. Owens, R. J. Behun, J. C. Manning, and R. C. Reid "The Impact of Internet Pornography on Adolescents: A Review of the Research" Sexual Addiction & Compulsivity, 19:99–122, 2012.

[40] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Video pornography detection through deep learning techniques and motion information", Neurocomput, pp. 279-293, 2017.

[41] S. Madkaikar, V. Belhekar, Y. Dharmadhikari and A. K. Tripathy, "Generating Textual Video Summaries using Modified Bi-Modal Transformer and Whisper Model," 2023 International Conference on Advanced Computing Technologies and Applications (ICACTA), Mumbai, India, 2023, pp. 1-7, doi: 10.1109/ICACTA58201.2023.10393115.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)