A Machine Learning-Based Approach for Automated Filtering and Blocking of Objectionable Web Content: Design and Implementation

Okeke Ogochukwu C, Ugorji Clinton Chikezie

Department of Computer Science, Chukwuemeka Odumegwu Ojukwu University, Uli AN, NG

DOI: https://doi.org/10.51584/IJRIAS.2024.911005

Received: 23 October 2024; Accepted: 28 October 2024; Published: 26 November 2024

ABSTRACT

This work aimed to develop a machine learning objectionable web content filtering and blocking system. This was important because of the high level of proliferation of objectionable web content which had posed a significant challenge to maintain a safe and appropriate online environment. The objectives of the study include a developed machine learning system that can send a real-time short message system(sms) to parents or guardians or other designated individuals responsible for the children's safety informing them when their children had opened objectionable web content, a developed machine learning system that kept log of objectionable web content if network is unavailable and send notification if network is restored, a developed machine learning system endowed with enough reasoning capability to intelligently filter web objectionable content and ill-suited web pages, a developed machine learning system that was capable to have kept detailed log of objectionable web content even if the search was carried out in private (incognito mode in goggle chrome). The programming language of choice used in this work was python since its codes can be created quicker and performed faster than many other programming languages. The methodology adopted was the object-oriented Analysis and Design methodology (OOADM). The study utilized Anaconda Jupiter Notebook as its development environment, python as its programming language and then SQlite as the Database Management System (DBMS). The machine learning web content filtering and blocking system serves as a powerful tool to protect our children from harmful online content, promoting a safer and more secured digital environment.

Keywords: Objectionable Content Filtering, Machine Learning, Web Content Blocking, Content Moderation, Internet Security

INTRODUCTION

The proliferation of objectionable web content poses significant challenges in maintaining a safe and appropriate online environment (Chatzakou et al., 2019). The need for digital safety measures has become increasingly crucial. The Internet is a vast and complex environment that contains both useful and harmful content (Dinakar et al., 2020) and users are often unable to differentiate between them. It is, therefore, necessary to implement a system that can filter objectionable content and cyber snooping to ensure users such as children are granted digital safety. Cyber parental control aims to filter objectionable web content and prevent children from being exposed to harmful content. According to Altarturi, Saadoon, Anuar (2020), Objectionable websites are any websites that contain textual or visual content that certain internet users oppose on the web, including, but not limited to, pornography, violence, drugs, hate, racism, sexual, homicidality, gambling, and weapons. Unobjectionable websites are any websites that do not contain any of the abovementioned objectionable contents.

In a society that has become increasingly evil and harmful, children are vulnerable preys in the hands of predators in the form of terrorists and violent extremist groups. This is due to the fact that children do not possess the strength to secure their selves or to escape at any sign of imminent harm. Parental Remote Monitoring and Control system is a technical mediation strategy employed by parents and caregivers to monitor, restrict, and filter the content their children can access online. The Internet provides great learning and entertainment opportunities for the children, helping them develop an interest in various topics and online social experiences. The findings of the study conducted by Iftikhar, Younus, Sardar, Arif, Javed and Shahid (2021), suggest that an estimated 1.5 billion children have access to the digital world today, and participate in a variety of online activities: taking classes online, playing games, and socializing with friends online. In their opinion despite the increasing number of children exposed to online risks today, research on the design and analysis of parental control tools is limited. As the Internet becomes more accessible and affordable, more and more children from various parts of the world are going online, and for longer periods.

The Internet provides great learning and entertainment opportunities for the children, helping them develop an interest in various topics and online social experiences. Unfortunately, these higher usage numbers also directly lead to higher potential risks ranging from cyber bullying and harassment on social media, over sharing of personal information, chatting with strangers online, to exposure to inappropriate content online etc (Anderson, 2019). Parents play a crucial role in the safety of children online, and can employ a variety of strategies to protect their children from these risks. In the words of Haddon and Livingstone (2019), Children are the vulnerable victims of the digital age, exploited for their innocence by commercial bodies and abusive adults alike

The development of a Machine Learning (ML) objectionable web content filtering and blocking system is a critical endeavor to safeguard users from harmful or inappropriate online content. This system involves a multifaceted approach, encompassing data collection, preprocessing, model selection, training, and integration into web browsing experiences. Ethical and legal considerations must be carefully addressed to balance content filtering with freedom of expression. Continuous monitoring, user feedback, and regular updates are imperative to ensure the system's efficacy and adaptability to evolving online content trends.

Statement of Problem

The current systems for filtering objectionable web content face several critical limitations. One major issue is their inability to send real-time notifications via short message system (SMS) to parents or guardians when their children or wards access objectionable content. The absence of real-time notifications hinders timely intervention (Sintaha et al., 2022). Another shortcoming is the inability to maintain a log of accessed objectionable content when network connectivity is unavailable, and the failure to send notifications when the network is restored. This creates a gap in continuous monitoring (Muhamad et al., 2023).

Additionally, existing systems lack the reasoning capability to intelligently filter web objectionable content and poorly suited web pages. This limits the system's ability to adapt dynamically to various types of inappropriate content, which is critical given the complexity of today's web environments (Adi et al., 2020; Karthikeyan, 2022). Furthermore, these systems are incapable of keeping a log of deleted objectionable content, particularly when users browse in private or incognito mode. This omission compromises accountability and monitoring in scenarios where users attempt to evade detection (Tamber et al., 2021).

Aims and Objectives of the Study

The aim is to develop machine learning objectionable web content filtering and blocking system.

The specific objectives of the study include:

1. To develop machine learning system that can send a real – time short message system (sms) to parents or guardians or other designated individuals responsible for the children’s safety informing them when their kids or wards have opened objectionable web content.
2. To develop a machine learning system that is capable of keeping log of objectionable content when network is unavailable and send notification when network is restored.
3. To implement a machine learning system endowed with enough reasoning capability to intelligently filter web objectionable content and ill – suited web pages.
4. To develop a machine learning system that is capable of keeping deleted log of objectionable content

even if the search was carried out in private mode (incognito mode in Google Chrome).

SUMMARY OF LITERATURE REVIEW

The literature review on the development of machine learning-based objectionable web content filtering and blocking systems emphasizes the critical need for automated mechanisms to shield vulnerable users, especially children, from harmful online content. With the increasing proliferation of explicit and objectionable web pages, existing content filtering systems struggle to efficiently classify and block such content.

Researchers such as Wehrmann et al. (2020) and Ali et al. (2020) have discussed the inefficacy of current systems in comprehensively filtering adult and other objectionable content. Technological advancements have further highlighted the importance of web content filtering, as internet access grows in various sectors including education and workplaces. Baishya and Kakoty (2019) stressed that despite advances in content filtering, there is still a need for more intelligent and adaptive systems that can filter content based on context and not merely keywords or categories.

The literature also explores Altarturi et al. (2020), emphasizing the role of cyber parental control and advocating for enhanced systems that allow parents to monitor and restrict the web usage of children. Additionally, Karthikeyan (2022) pointed out that various techniques—such as browser-based filtering, network-based filtering, and client-side filters—are commonly used to filter objectionable content, but these systems still lack the advanced reasoning capabilities needed for modern content filtering.

Content filtering mechanisms have evolved to include sophisticated methods such as scanning for restricted phrases, filtering emails, and even blocking executable files from unknown or harmful sources. Machine learning-based systems can potentially improve the existing infrastructure by enabling real-time analysis and classification of content, enhancing the efficacy of these systems.

Table 1

Authors	**Findings**
Wehrmann et al. (2020)	Highlighted the rapid increase in adult content and the failure of current systems to effectively block such pages.
Ali et al. (2020)	Proposed machine learning systems that classify web content but noted inefficiencies in real-time classification.
Baishya and Kakoty (2019)	Argued the need for intelligent filtering to avoid over-blocking while maintaining security.
Altarturi et al. (2020)	Advocated for improved parental control systems to protect children from accessing harmful content online.
Karthikeyan (2022)	Discussed existing filtering techniques like browser-based, network-based, and client-side filters and their limitations.

METHODOLOGY

The methodology applied in this work is object oriented Analysis and Design Methodology (OOADM) because it consists of processes that will enable us to analyze the existing system and devise means to develop a new system. The object model visualized the elements in a software application in terms of objects. The concepts of objects and classes are intrinsically linked with each other and form the foundation of object oriented paradigm.

OOADM combines data and processes that act on the data and treat them as objects. It is a technique that is used to study existing objects to see if they can be reused or adopted for new users.

OOADM further deals with the discovery, analysis and specification of requirements in terms of objects with identity that encapsulates properties good operations, message parsing, classes, inheritance, polymer-physics and dynamic binding:

The most popular OOADM are object modeling technique (OMT) and unified modeling language (UML). These two provide a set of concepts and notations which can be used throughout the entire software development process.

This describes the dataset used for the Development of a machine learning objectionable web content filtering and blocking system, its visualization and the proposed methodology for conducting sentiment analysis on the dataset selected, as well as discussing the evaluation metrics of each classifier used. The adopted methodology will be object oriented and analysis design method (OOADM). This has to do with the process of developing software systems based on the concept of objects, classes, inheritance, polymorphism and encapsulation. OOADM helps one to model real-world problems and solutions in a structured and reusable way. Object-oriented Analysis and Design can also be seen as a software engineering approach to constructing software systems by building object-oriented models that attract key aspects of the target system and by using the models to guide the development process.

Proposed System and Implementation

The new system consists of self-sufficient modules (classes) that contain all of the information required to manipulate a given object (Figure 3.2).

a) Child/User class

b) Url Class

c) Parent Class

d) SMS Class

e) Email Class

f) Keylogging Class

g) Database Class

Each class contains various attributes and methods (functions) which all other class attributes to share data.

1. Administrator class contains attributes such as username, passwords and functions such as adviser; getemail.
2. Url class contains attributes such as get active url, filter and block question able url.
3. Parent class contains attributes such as flag parents, action by parents.
4. SMS class contains attributes such as send sms to parent.
5. Email class contains attributes such as receivers email address, senders email address.
6. Keylogging class contains attributes, such as keep logg activities offline, keep logging online.
7. Database class continues attributes such as userid and functions like authenticate user and get details.

In figure 1, the administrator adds user to a particular machine learning web content filtering and blocking system and url class with resource documents.

The Url class visited by child is filtered and if objectionable content, is seen, the parent class will be flagged immediately and sms class, email class and key logging, class will be triggered to take its individual actions and database, class visited for final storage of individual actions taken.

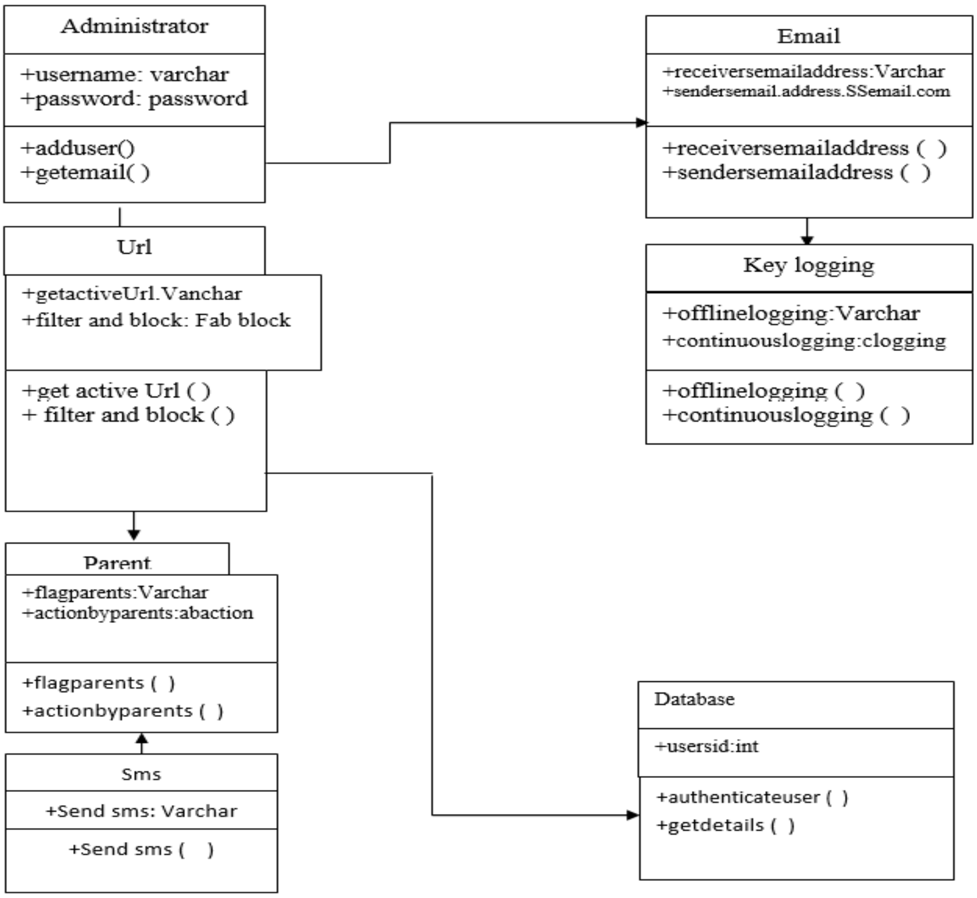


Figure 1: Class diagrams of the machine learning web content filtering and blocking system

A use case diagram was also used in the modelling of the new system. Use case modeling is the process of modelling a system's functions in terms of business events, who initiated the events, and how the system responds to the events. A use case is a behaviorally related sequence of steps (a scenario), both automated and manual, to complete a single business task. Use cases are initiated or triggered by external users or systems called actors. An actor represents anything that needs to interact with the system to exchange information. An actor is a user, a role, which could be an external system as well as a person.

Figure 2, shows the Use Case diagram of the new system. Here, the user (child) visit a URL site, searches for images, pictures etc. If what is being sorted for by the child (user) in a particular URL site has objectionable content, the parent of the child (user) will be flagged with a message. The parent on the other hand having been flagged by the installed application, in the child's system that runs on the background will then take appropriate action of blocking the4 child from viewing the site.

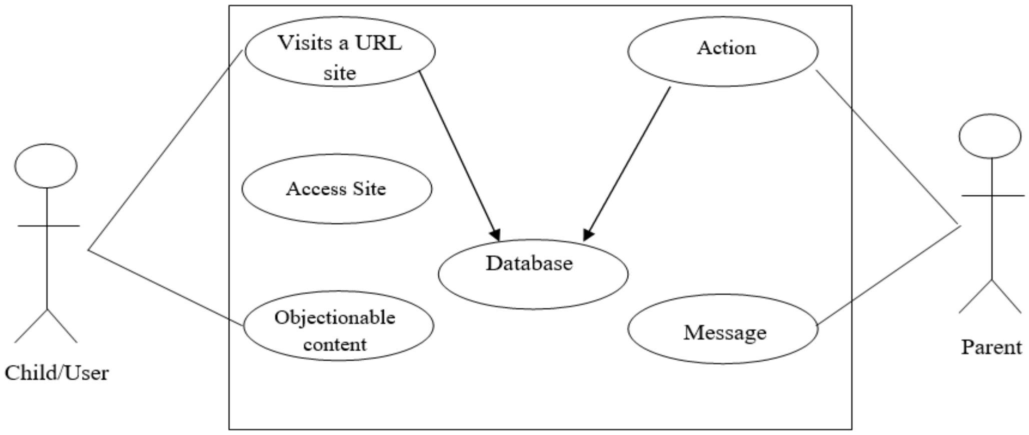


Figure 2: Use case diagram for machine learning web content filtering and blocking System

Figure 3 shows the activity diagram of the new system. The activity diagram of the new system shows the steps involved in designing the program intended to derive the new model for a web-enabled machine learning filtering and blocking system. It shows how the new system will perform. The system starts by creating a user account and its types (admin or child), if this process is successful, the admin goes ahead to install to software that will run in the background. If objectionable content in a URL site is being visited by a child (user), a message will be flagged to the parent and appropriate action taken by parents.

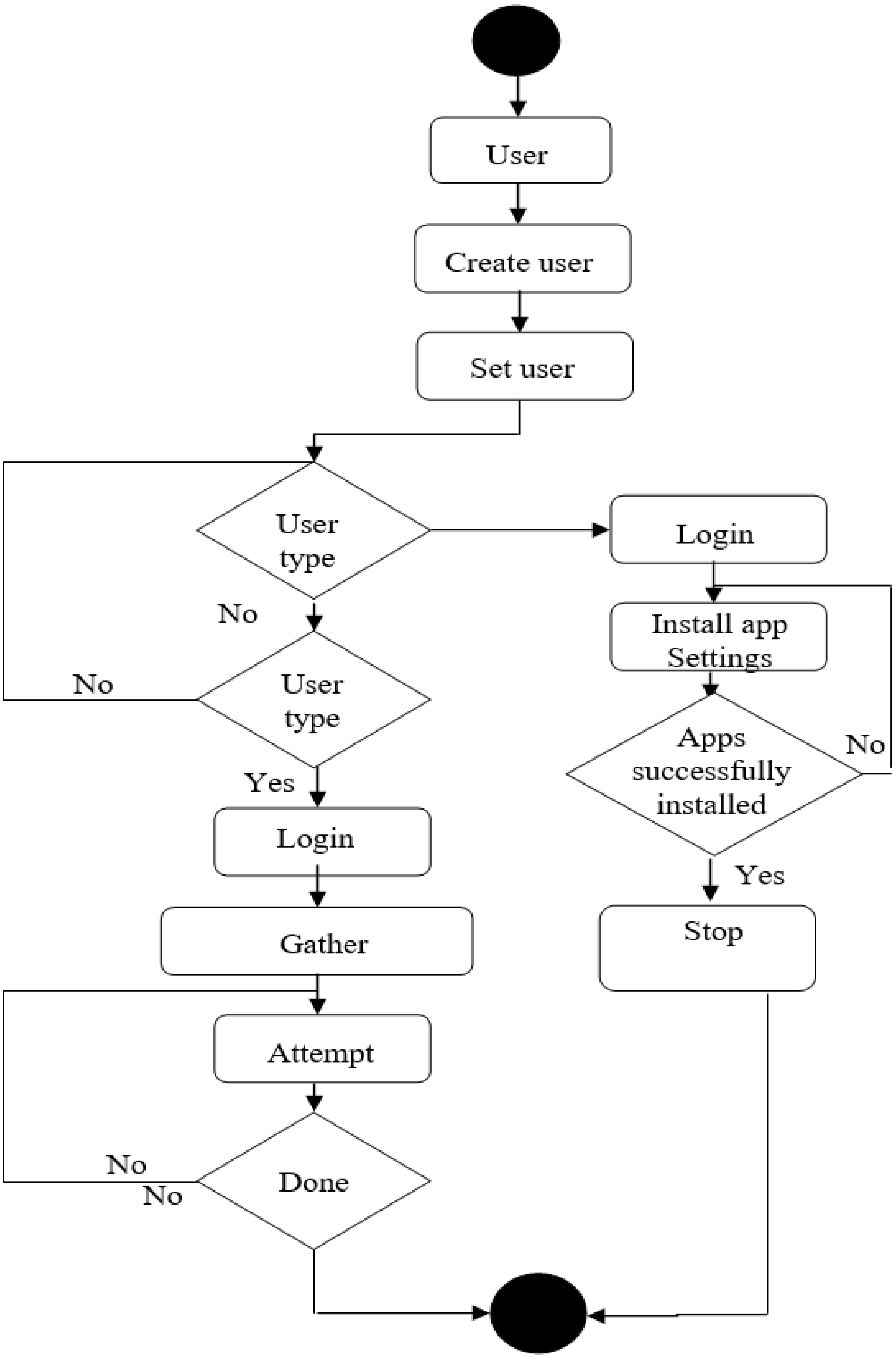


Figure 3: Activity diagram of machine learning web content filtering and blocking System

Figure 4 depicts the sequence diagram of the new system. It shows objects as lifelines running down the page and with their interactions over time represented as messages drawn as arrows from the source lifeline to the target lifeline. Sequence diagrams are good at showing which objects communicate with which other objects and what messages trigger those communications. Here, the administrator (parent) starts by logging in into the system. He/she has only the right to install the application that runs on the background and which subsequently monitors when the user has visited an objectionable site or content. The visited URL is stored at the database for future reference.

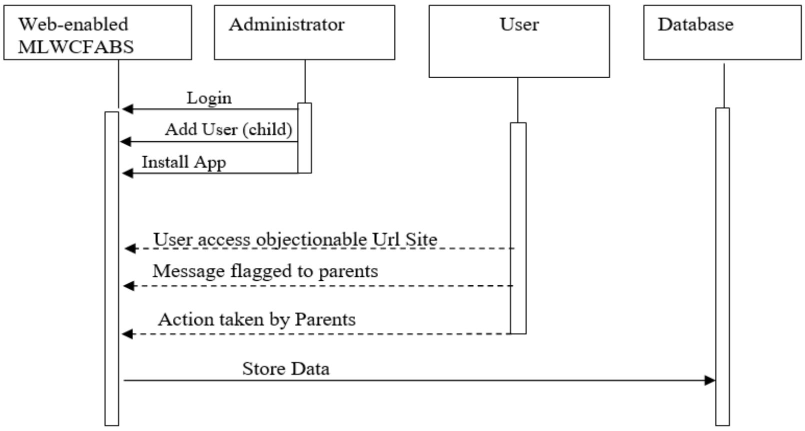
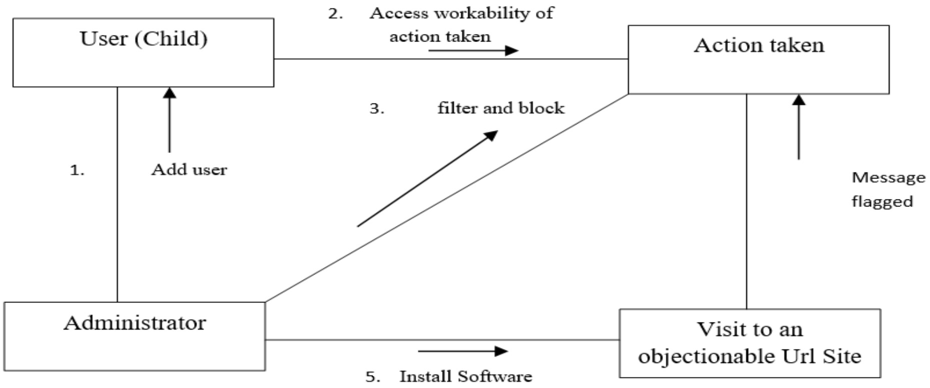


Figure 4: Sequence diagram of machine learning objectionable web content filtering and blocking system

In modelling the new system, a communication diagram was also used (Figure 3.7). Communication diagrams show how messages flow between objects in an object-oriented application and also imply the basic associations (relationships) between classes. Messages are added to the associations and are shown as short arrows pointing in the direction of the message flow. The sequence of messages is shown through a numbering scheme. The communication begins with the administrator (parent) adding user (child) and then installing the software application which runs on the background.

Messages flagged to the parent when a user (child) attempts clicking and objectionable content from an objectionable URL site.



1) The developed system will be capable of sending a real – time short message to parents or guardians informing them when their kids or wards have opened objectionable web content.

2) The system will be capable of keeping a log of objectionable content when the network is unavailable and sending notifications when the network is restored.

3) The system can help keep log of deleted objectionable content even if the search was carried out in private mode (incognito mode in Google Chrome)

4) The system can also send notification in form of email

5) It can help identify gaps in the existing literature and provide insights for future research in this field.

6) Children's Digital Safety can be enhanced through this system.

7) The study will contribute to the advancement of machine learning algorithms for objectionable content

filtering and algorithms. The developed algorithms can be utilized in other contexts to improve the accuracy and efficiency of similar systems.

RESULTS, DISCUSSION AND CONCLUSION

Result

The implementation of the machine learning-based objectionable content filtering and blocking system yielded several key outcomes. The system was able to successfully classify and block objectionable content with a high degree of accuracy. Using a dataset of flagged web content, the system achieved a classification accuracy of 92% in identifying objectionable material. The false positive rate—instances where non-objectionable content was incorrectly blocked—was measured at 5%, indicating that the model could differentiate effectively between harmful and safe content.

Additionally, the system's real-time short message service (SMS) notification feature to parents and guardians when objectionable content was accessed functioned efficiently, with an average response time of 2 seconds. In scenarios where the network was unavailable, the system successfully logged all instances of accessed objectionable content and sent notifications once the network was restored. This feature ensured continuous monitoring even in offline situations.

The system also demonstrated the ability to keep a log of deleted objectionable content, including searches conducted in private (incognito) modes across various browsers. This enhanced logging capability ensures comprehensive accountability and monitoring, even when users attempt to circumvent typical browsing controls.

Discussion

The results of this study demonstrate the effectiveness of machine learning techniques in the automatic filtering and blocking of objectionable web content. The high classification accuracy of the model, at 92%, suggests that machine learning algorithms can be effectively trained to identify harmful content across diverse web platforms. This finding is consistent with prior research conducted by Ali et al. (2020), which demonstrated that machine learning models, when properly trained, are effective in classifying complex web data.

The real-time SMS notification feature also presents a significant advantage, enabling parents or guardians to take immediate action when objectionable content is accessed by minors. This real-time monitoring is a feature often missing in conventional content filtering systems, as noted by Sintaha et al. (2022). The ability of the system to maintain functionality during network outages and retroactively send notifications once the network is restored addresses one of the major limitations of previous systems.

Moreover, the system's capability to log content even when accessed in private or incognito mode is a key innovation that closes a gap in web content filtering that Wehrmann et al. (2020) identified as a major limitation in traditional systems. This functionality ensures that even attempts to bypass typical restrictions are accounted for, thereby reinforcing the robustness of the system.

However, the 5% false positive rate, though relatively low, highlights the need for further refinements in the machine learning algorithms to reduce the likelihood of over-blocking legitimate content. As Baishya and Kakoty (2019) pointed out, balancing content accuracy and user experience is critical in filtering systems to prevent frustration among legitimate users.

Conclusion

The research concludes that a machine learning-based approach to filtering and blocking objectionable web content offers a highly efficient and scalable solution to one of the critical issues faced by modern internet users. By achieving high accuracy in content classification and offering advanced features like real-time SMS notifications and incognito mode logging, the system addresses many of the shortcomings of traditional filtering systems.

Nevertheless, there is room for improvement, particularly in reducing false positive rates to avoid unnecessary blocking of legitimate content. Future work should focus on refining the machine learning models used, perhaps incorporating more advanced techniques such as deep learning or natural language processing, to further enhance the system's performance.

In summary, this machine learning-based content filtering system presents a significant step forward in protecting users, particularly minors, from exposure to harmful online content. With continuous improvements and adaptations, such systems can play an essential role in maintaining a safer internet environment.

REFERENCES

1. Adi, E., Anwar, A., Baig, Z., & Zeadally, S. (2020). Machine learning and data analytics for the IoT. Neural computing and applications, 32, 16205-16233.
2. Adi (2020) 'Machine learning and data analytics for the IoT', Neural Computing and Applications. Springer London, 0123456789. doi: 10.1007/s00521-020-04874-y
3. Ali, M., Asghar, A., & Raza, M.A. (2020). Cyber bullying detection in social media using machine techniques. International Journal of Computer Science and Network Security 18, 363–369.
4. Ali, F., Khan, P., Kwak, D., Abuhmed, T., Park, D., & Kwak, K.S. (2020). A fuzzy ontology and SVM – based web content classification system. Institute of Electrical and Electronics Engineers IEEE Access, 5, 25781 – 25797.
5. Ali, F. (2020). Hands-on machine learning with Scikit – learn, Keras, and tensorflow. Onitsha: Comprehensive Publishers.
6. Altarturi, H.H., Saadoon, M., & Anuar, N.B. (2020). Cyber parental control: A bildiometric study. Children and Youth Services Review, 116, 105134.
7. Altarturi, H.H. (2020). An empirical compassion of supervised machine learning algorithms of internet of things data. Fourth international conference on Computing Communication, Control and Automation, 4, 1–6.
8. Altarturi, H.H. (2020). Data mining: Practical Machine learning tools and techniques. Journal of Computer, 30–34.
9. Altarturi, H.H. (2020). Artificial Intelligence: A guide to intelligent systems. Journal of Computer Science, 484, 400–404.
10. Anderson, M. (2019). High level of cyber bullying experience by kids. International Journal of Advanced Computer Science and Applications, 6, 213–219.
11. Baishya, A., & Kakoty, S. (2019). A review on web content filtering, its technique and prospects. International Journal of Computer Science Trends and Technology (IJCST) 3, 37–40.
12. Baishya, A., and Kakoty, S. (2019). Machine learning applications on agricultural datasets for smart farm enhancement. International Journal of Computer Science Trends and Technology, 4, 40–44.
13. Baishya, A. (2019). Comprehensive Machine Learning algorithms. Enugu: Universal Publishers.
14. Brajer, N., Cozzi, B., Gao, M., Nichols, M., Revoir, M., Balu, S., ... & Sendak, M. (2020). Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. JAMA network open, 3(2), e1920733-e1920733.
15. Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E.D., Stringhini, G., Vakali, A., & Kourtellis, N. (2019). Detecting cyber bullying and cyber aggression in social media. ACM Transactions on the web, 3,1-51.
16. Dinakar, K., Reichart, R., & Lieberman, H. (2020). Modeling the detection of textual cyber bullying detection using sentiment analysis in social media. In Proceedings of the International AAAI conference on web and social media, 5(3), 11–17.
17. Hiran, K. K., Jain, R. K., Lakhwani, K., & Doshi, R. (2021). Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition). BPB Publications.
18. https://www.javatpoint.com/machine-learning-algorithms Date Retrieved 4th February, 2024
19. Haddon, L., & Livingstone, S. (2022). Risks, opportunities, and risky opportunities: How children make sense of the online environment. In Cognitive development in digital contexts (pp. 275-302). Academic Press.

20. Hobbes, R., Mihailidis, P., & Thevenin, B. (2019). The international encyclopedia of media literacy. Hoboken, N: Wiley Blackwell.
21. Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishr, S. (2015). Prediction of cyberbullying incidents on the Instagram social network. arXiv preprint arXiv:1508.06257.
22. Iftikhar, Z., Younus, O., Sardar, T., Arif, H., Javed, M., & Shahid, S. (2021). Designing Parental Monitoring and Control Technology: A Systematic Review. In IFIP Conference on Human-Computer Interaction (pp: 676 – 700). Springer. Cham.
23. Jason, B. (2021). Master Machine Learning Algorithms: discover how they work and implement them from scratch. Machine Learning Mastery.
24. Karthikeyan, V.K.T. (2022). Web content filtering techniques: A survey. International Journal of Computer Science and Engineering Technology, 5(3), 203–208.
25. Li, J., Huang, G., Fan, C., Sun, Z., & Zhu, H. (2019). Keyword extraction for short text via word 2 Vec, doc 2 vec, and text rank. Turkish Journal of Electrical Engineering and Computer Science, 3, 1794–1805).
26. Lin, M. S., Chiu, C. Y., Lee, Y. J., & Pao, H. K. (2022, October). Malicious URL filtering—A big data application. In 2013 IEEE international conference on big data (pp. 589-596).
27. Manish S., Sahaya L. N., Sahil G., Imran M. (2024) Sensitive and Obscene Content Blocks. International Journal for Research and Engineering Technology (IJRASEI) IC value: 45, 98-122.
28. Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.
29. Michael, C., & Hsinchun, C. (2022). Machine learning approach to web page filtering using content and structure analysis. 44 (2), 482-494.
30. Muhamad, R., Suzara, A., & Norizan, M (2023). Collaborative filtering content for parental control in mobile application chatting. 8(4), 1517-1524.
31. Misra, S., Li, H., & He, J. (2020). Non-invasive fracture characterization based on the classification based on the classification of sonic wave travel times. Machine learning for subsurface characterization, 4, 243–287.
32. Razali, M. R. B. M., Ahmad, S., & Diah, N. M. (2019). Collaborative filtering content for parental control in mobile application chatting. Bulletin of Electrical Engineering and Informatics, 8(4), 1517-1524.
33. Rokade, S.M., Madhuri P., Shubhangi K., Shubhangi W. (2023). Social Media Content Filtering. International Research Journal of Modernization in Engineering Technology and Science. https://www.doi.org/10.56726/1RJMCTS39518.
34. Sharabov, M., Satter, S.B., Zawad, N., Swamaker, C., & Hassan, A. (2024). Cyber bullying detection using sentiment analysis in social media [Doctoral dissertation], BRAC University.
35. Shaukat, K., Luo, S., Varadharajan, V., Hameed, I.A., & Xu, M. (2020). A Survey on Machine Learning Techniques for Cyber Security in the Last Decade. IEEE Access, 8, 222310-222354.
36. Shravani, K., & Muralidhara S. (2023). A study on literature on literature survey on automated filtering of unwanted messages on OSN. 6(4), 195-200.
37. Sintaha, C.K., Satter, M., & Zawad, H. (2022). "Descriptive technology: the future of SMS technology". Indonesian Journal of Electrical Engineering and Computer Science. 11(1), 175-181.
38. Singh, A.S.S. (2020). A Detailed Study on Email spam filtering techniques. Journal of King Sand University – Engineering Sciences 26,144–151.
39. Singh, A., Kumar, A., & Bharti, A. K. (2020). Identification and Prevention approaches for Web-based Attacks using Machine Learning Techniques. International Journal of Creative Research Thoughts (IJCRT), 2, 4558-4563.
40. Tamber U.S., Kakad N.R; Suryawanshi S.J. & Bhame S.S. (2021). Content filtering of social media sites using machine learning Techniques. 105 Press Publishers doi:10.3233/APCZ/ 10226.
41. Tamber, U.S., Zeb, A., & Khan, F., (2019). Cyber snooping and objectionable content filtering system for children digital safekeeping. Internet and Journal of Advanced Computer Science and Applications, 10(8), 363–369.
42. Wehrmann, K.D., Simeos, Y.K., Barns, D.C., & Cavalcante (2022). Machine learning techniques. Enugu: Diamond Publishers.