e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

SECURING CHILDREN FROM INAPPROPRIATE AND HARMFUL THINGS ON THE INTERNET

Santosh Kawade[*1], Akanksha Deshmukh[*2], Sejal Jadhav[*3], Omkar Bankar[*4]

[*1]Asst. Prof., Department Of Computer Engineering Dr. D.Y. Patil College Of Engineering And Innovation, Pune, India.

[*2,3,4]Student, Department Of Computer Engineering Dr. D.Y. Patil College Of Engineering And Innovation, Pune, India.

DOI: https://www.doi.org/10.56726/IRJMETS63320

ABSTRACT

This project proposes an intelligent, automated system to safeguard children from harmful online content by using advanced machine learning. It dynamically monitors and filters inappropriate material in real time, minimizing exposure to risks like explicit images and violent media. The system combines image recognition, natural language processing, and contextual analysis for comprehensive content detection, allow- ing for a safer digital experience without constant parental intervention. It offers a user-friendly, efficient tool for parents to protect their children, promoting responsible internet use in an increasingly connected world.

Keywords: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), Natural Language Processing (NLP), Clustering Algorithms.

I. INTRODUCTION

The internet has become a key part of children's lives, offering access to educational resources, entertainment, and social interaction. While it provides many benefits, it also exposes children to harmful content like explicit images, violent videos, and unauthorized websites, which can neg- atively impact their mental well-being and safety. Parents often struggle to monitor their children's online activities be- cause traditional parental control tools are time-consuming, difficult to manage, and often ineffective against new types of harmful content.

These tools usually rely on preset filters that don't adapt to the constantly changing online environment, requiring parents to supervise their children's activities constantly. However, many parents may lack the time or technical know-how to manage these controls effectively.

To address these challenges, this project aims to create an intelligent, automated system that uses machine learning to monitor and block inappropriate content in real-time. This system will include advanced features like image recognition, natural language processing, and contextual analysis, allowing it to continuously adapt and protect children from new threats across various platforms. Unlike traditional tools that focus on restriction, this system will create a safer, more enriching online experience by balancing protection with positive online exploration. The goal is to provide parents with an easy-to-use, highly effective tool that helps create a safer digital environment for children.

II. MOTIVATION

- The internet is a great tool for learning and con- necting with others, but it can also expose children to harmful and inappropriate content, like violence or explicit material, often by accident. This can lead to negative experiences and harmful ideas. Unfortunately, most parental control systems rely on parents constantly watching what their kids do online, which is time- consuming, difficult to manage, and doesn't keep up with the fast-changing online threats.
- Our project aims to solve this problem by making the internet safer for children. Using machine learning, we are developing a system that automatically filters and blocks harmful websites and content, so parents don't have to monitor their kids' activities all the time. The goal is to provide parents with a smart tool that not only blocks dangerous material but also adapts to new threats as they appear, offering ongoing protection. Ultimately, we want to fill the gaps in current safety methods and ensure that children can explore, learn,and enjoy the internet without the risk of encountering harmful content.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

III. OBJECTIVES

- Website Blocking: Create a system that automatically blocks access to websites that are flagged as inappropriate or dangerous, based on a trained machine learningmodel.
- Low False Positive Rates: We'll work on improving the accuracy of these tools to avoid blocking safe content by mistake, while still being highly effective at catching harmful material.

IV. LITERATURE SURVEY

- Neural Network-Based Cyber-Bullying-Aggression De- tection Using (Twitter) Michael Agbaje et al. (2024) This paper explores the detection of cyberbullying and cyber aggression on social media platforms, with a particular focus on Twitter. The authors employed neural network-based models to analyze and detect harmful behaviours in real time. Using deep learning techniques such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). The ex- periments were conducted on a Windows 11 machine, and their results demonstrated that RNNs were particu- larly effective in detecting cyberbullying, achieving an impressive accuracy. This study underscores the utility of advanced neural networks in accurately identifying harmful online behaviours, contribut ing significantly to improving online safety through dynamic monitoringand filtering.[1]

S/N	MODEL	ACCURACY	F-MEASURE
1	Logistic Regression	0.671	0.637
2	Stochastic Gradient Descent	0.655	0.601
3	Bernoulli Naïve Bayes	0.640	0.605
4	Random Forest	0.558	0.512
5	K-Nearest Neighbour	0.543	0.476
6	Linear Support Vector machine	0.669	0.648
7	Recurrent Neural Network	0.951	0.910
8	Convolutional Neural Network	0.911	0.890

- Cyber Bullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continu- ous Bag of Words Feature Extraction Suliman Mo- hemed Fati et al. (2023) This research provides a comparative analysis of various deep learning mod- els for cyberbullying detection on Twitter, focusing on attention mechanisms and continuous bag of-words for feature extraction. After performing natural lan guage processing (NLP) techniques to preprocess the dataset, the study applied several deep learning algorithms to analyze problematic comment patterns. The study evaluated mod els like Long Short-Term Memory (LSTM), Conv1DLSTM, CNN, and BiLSTM, with Conv1DLSTM showing the best overall performance. The analysis demonstrates that attention-based mecha- nisms can significantly enhance the detection of cyberbullying when compared to standard machine learning techniques, positioning deep learning models as pow- erful tools for real-time online content monitoring and prevention of abusive behaviour. [9]

S/N	MODEL	ACCURACY	PRECISION	RECALL	F1SCORE
1	LTSM	0.8011	0.8142	0.7281	0.7687
2	Conv1DLTSM	0.8649	0.8146	0.8919	0.8515
3	CNN	0.8496	0.8836	0.7908	0.8346
4	BiLSTM	0.7795	0.8373	0.8130	0.8250
5	BiLSTM_Pooling	0.7982	0.8167	0.7862	0.8012
6	GRU	0.7093	0.7089	0.7561	0.7317

- Deep Learning-Based Cyber Bullying Early Detec- tion Using Distributed Denial of Service(DDoS)Flow Muhammad Hassan Zaib et al. (2020) This paper proposes a novel method for the early detection of DDoS attacks by comparing flow-based and non-flow- based datasets. The authors employed a combination of Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for classification, optimizing feature selection with statistical tests and correlation methods. The research achieved high accuracy rates. The results highlight the robustness of deep learning models, espe- cially SVM, in the early detection of DDoS attacks. Themethodology used in this paper provides crucial insightsfor developing efficient models capable of predicting and mitigating network-based cyber threats, further strengthening the application of machine learning indynamic threat prevention.[12]

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

S/N	MODEL	ACCURACY	PRECISION	RECALL	F1SCORE
CSE-CIC IDS 2018 DATASET 1					
1	SVM	96.26	91	88	88
2	ANN	94.87	95	94	95
NSL KDD DATASET 2					
1	SVM	98.73	97	97	97
2	ANN	89.98	90	89	90

- Cyberbullying Detection: Hybrid Models Based on Ma-chine Processing Techniques Chahat Raj et al. In their 2023 study, Nagy et al. explore the effectiveness of various machine learning (ML) and deep learning (DL) tech niques for detecting phishing URLs, addressingthe growing threat of online scams. Utilizing a dataset of 54,000 records for training and 12,000 for testing, the authors implement models such as Random Forest (RF), Na¨ ıve Bayes (NB), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN)to classify websites as either legitimate or phishing.The study highlights the use of multiprocessing andmultithreading techniques in Python, enabling efficient parallel training of the models. By following key steps of preprocess ing, classification, and evaluation, the research demonstrates the impact of these techniques ondetection accuracy and training efficiency. The compar-ative analysis offers valuable insights into the strengths and weaknesses of the various ap proaches, contributing to advancements in phishing detection methodologies and enhancing online user safety. [7]

DATASET	MODEL	ACCURACY	PRECISION	RECALL	F1SCORE
Wikipedia attack Dataset	Logistic Regression	80.90	79.36	80.97	97.74
	Bayes Exp. Max	82.70	81.33	82.83	81.36
	Bayes	83.11	81.78	83.14	81.58
	CNN	92.91	92.09	83.78	88.63
	BERT	95.31	92.61	93.57	95.70
	Bi-GRU with GloVe	96.98	99.22	96.74	98.56
Wikipedia Web Toxicity Dataset	Logistic Regression	80.42	78.91	80.46	79.23
	Bayes Exp. Max	82.10	80.60	81.87	80.57
	Bayes	82.19	80.63	82.01	80.60
	CNN	93.52	92.79	88.67	91.56
	BERT	95.69	92.71	95.11	96.82
	Bi-GRU with GloVe	96.01	99.45	96.8	98.63

- Phishing URLs Detection Using Sequential and Par-allel ML Techniques Naya Nagy et al.. In their 2023 study, Nagy et al. explore the effectiveness of various machine learning (ML) and deep learning (DL) tech niques for detecting phishing URLs, addressing the growing threat of online scams. Utilizing a dataset of 54,000 records for training and 12,000 for testing, the authors implement models such as Random Forest (RF), Naıve Bayes (NB), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN) to classify websites as either legitimate or phishing. The study highlights the use of multiprocessing and multithreadingtechniques in Python, enabling efficient parallel trainingof the models. By following key steps of preprocessing,classification, and evaluation, the research demonstratesthe impact of these techniques on detection accuracy and training efficiency. The comparative analysis offers valuable insights into the strengths and weaknesses of the various approaches, contributing to advancements in phishing detection methodologies and enhancing onlineuser safety.[8]

S/N	MODEL	ACCURACY	PRECISION	RECALL	F-MEASURE
1	Random Forest	95.14	87.28	100	93.21
2	Naïve Bayes	96.01	95.65	92.25	93.92
3	CNN	95.13	87.24	100	93.19
4	LTSM	95.14	87.28	100	93.21

- The proposed method combines the squirrel search algorithm, an optimization technique inspired by nature, with BiLSTM for anomaly recognition. The framework uses the knowledge gained from a sequence of frames to categorize the video as either typicalor abnormal. The proposed method was exhaustively tested in several benchmark datasets for anomaly detection to confirm its functionality in challenging surveillance circumstances. The results show that the proposed framework outperforms existing methods in terms of area under curve (AUC) values, with a testset AUC score of 93.1%.The paper also discusses

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

the importance of feature selection and the benefits of using BiLSTM over traditional unidirectional long short-term memory (LSTM)models for anomaly detection in videos. Overall, the proposed framework provides a highly precise computerization of the system, making it an effective tool for identifying abnormal human behavior in surveillance footage.[13]

V. CONTENT TYPES

- Explicit Images: Filtering out adult content, pornog-raphy, and other inappropriate visuals.
- Violent Videos: Blocking videos with violent orgraphic content.
- Unsafe Websites: Restricting access to websites thatpromote harmful behaviour or contain inappropriate material.

Technological Framework

- Multi-Modal Content Analysis: Using natural lan-guage processing (NLP) for text analysis, image recognition for visual content, and context-aware algorithms to assessthe relevance and appropriateness of online materials.
- Real-Time Filtering: Implementing dynamic content monitoring that instantly blocks inappropriate content duringonline activities.

User Interface and Experience

- Child-Friendly Interface: An intuitive, engaging design for children, seamlessly integrating safety features for a positive user experience.

Implementation and Testing

- Prototype Development: Creating an initial prototype for real-world testing.
- Pilot Programs: Partnering with schools, community organizations, and families to evaluate the system in diverse settings.
- Feedback Mechanisms: Gathering user feedback to improve filtering algorithms and interface usability.

Ethical Considerations

- Data Protection: Ensuring secure handling of any datacollected, with transparent consent protocols.
- Transparent Filtering: Providing users with insights into filtering decisions, fostering trust and accountability.

Target Audience

S/N	MODEL	ACCURACY	PRECISION	RECALL	F1
1	CNN	89.7	84.9	81.6	82.5
2	F-CNN	92.6	87.2	84.6	87.4
3	LSTM	96.8	97.3	85	96.5
4	Bi-LSTM	98.2	97.3	95	96.5

VI. SCOPE

ARCHITECTURE

1. Data Collection and Preprocessing: To build robustmachine learning models, the system requires extensive datasets of both appropriate and inappropriate content. Data collection will be conducted through the following steps:

Content Categories: The dataset will be categorized into

1.1. The system will primarily cater to children aged 3 to 14, addressing their need for protection from harmful contentwhile enabling safe online exploration.

1.2. Parents and guardians will also be a focus, providing them with tools to effectively prevent their children's online activities without constant supervision.

images, videos, and textual content, focusing on explicit materials (e.g., pornography, violent media), safe content, and borderline cases.

Source Selection: Publicly available datasets, web scraping (with ethical and legal considerations), and partnerships withcontent moderation agencies will be used to gather relevant content.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

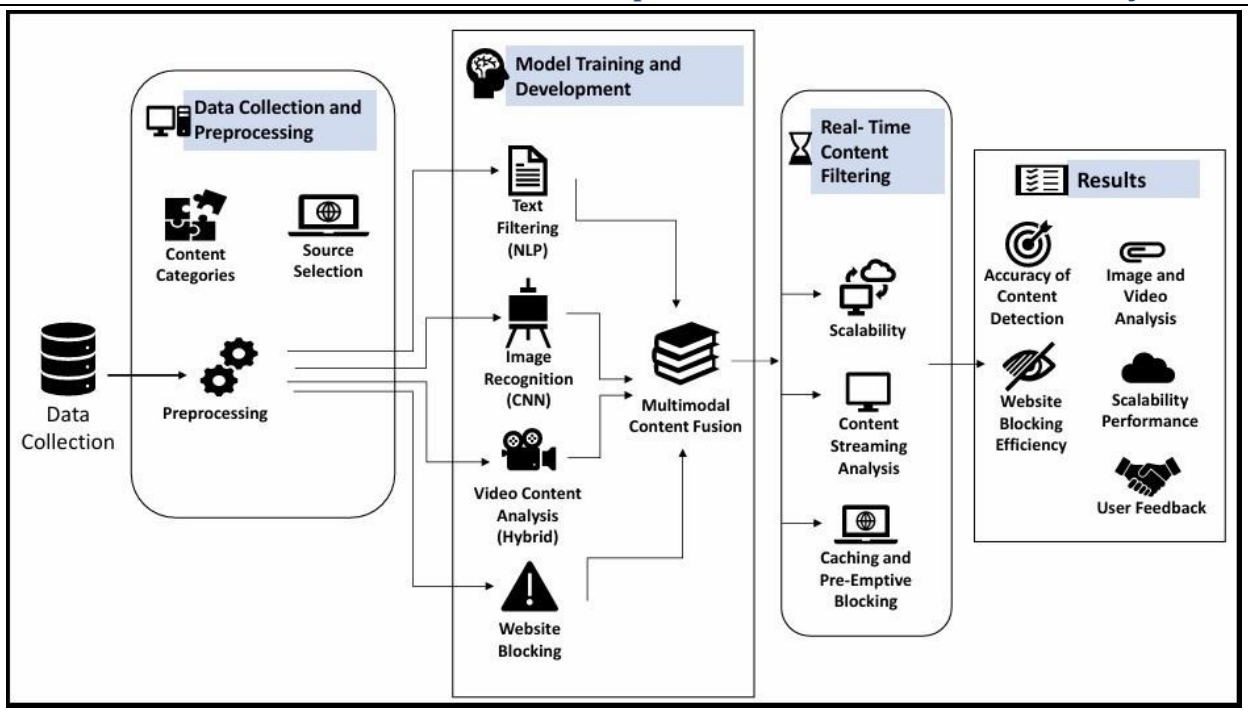

Fig. 1. Project Architecture

Preprocessing: The collected data will be preprocessed to remove noise and ensure high-quality inputs. This involves normalizing images, cleaning text data, removing duplicates, and labelling content based on appropriateness levels.

1. Machine Learning Model Development The core func- tionality of the system relies on real-time content filtering driven by machine learning models.

Natural Language Processing (NLP) for Text Filtering: NLP models will be trained to analyze text content on web pages, social media, and other sources. Techniques such as sentiment analysis, keyword detection, and context-aware processing (using Transformerbased models like BERT) will be employed to detect harmful or inappropriate language, slurs, or unsafe content.

Image Recognition for Visual Content: Convolutional Neural Networks (CNNs) will be employed to analyze images, identifying explicit images (e.g., nudity, violence) usingmodels trained on large-scale image datasets Video Content Analysis: A hybrid approach combining object detection (using YOLO or SSD) and frame-based analysis will beemployed to detect violent or explicit scenes within video content. Recurrent Neural Networks (RNNs) may also be used to understand sequences in videos to ensure contextual accuracy in detecting harmful scenes.

2. Real-Time Content Filtering: To ensure a seamless and efficient filtering process, the system will be designed to operate in real-time. This will require:

Content Streaming Analysis: The system will monitor and analyze web pages, images, and videos in real-time, detectingand blocking harmful content before it is presented to the child.

Caching and Pre-emptive Blocking: By using previouslygathered data and pre-learned models, the system will pre- emptively block websites and content known to be harmful. Scalability: The real-time filtering mechanism will be implemented in a scalable cloud environment using mi- croservices to handle large volumes of data and multiple simultaneous users.

3. Testing and Validation The system's effectiveness willbe evaluated through several stages of testing: Offline Testingwith Datasets: Initial validation of the machine learningmodels will be performed using test datasets separate fromthe training data, focusing on accuracy, precision, recall, andF1-score to ensure reliable detection of inappropriate content.Pilot Testing: The system will undergo pilot testing in controlled environments such as schools or family homes, where parents will monitor its performance and provide feedback on its functionality.

Real-World Deployment: After refining the system based on pilot tests, a broader deployment will take place. Dur-ing this phase, performance metrics such as latency, false- positive rates, and false-negative rates will be monitored.

VII. CONCLUSION

Securing children from inappropriate and harmful things on the internet requires a multi-faceted approach that involves parents, educators, and technology providers working together. It's essential to strike a balance between ensuring children's safety online and safeguarding their freedom of expression and autonomy. Protecting children from harmful content online requires a team effort from parents, teachers, and tech companies. It's important to finda balance between keeping kids safe on the internet while still allowing them to express themselves and explore freely.

VIII. REFERENCES

[1] Agbaje, M., and Afolabi, O. (2024). Neural Network-Based Cyber- Bullying and Cyber-Aggression Detection Using Twitter(X) Text. Revue d'Intelligence Artificielle, 38(3), 837.

[2] Apandi, S. H., Sallim, J., and Mohamed, R. (2024). Use Word Cloud Image Of Web Page Text Content On Convolutional Neural Network (CNN) For Classification Of Web Pages. International Journal of Computing and Digital Systems, 15(1), 347.

[3] Wen, L., Zhang, M., Wang, C., Guo, B., Ma, H., Xue, P., Ding, W., and Zheng, J. (2024). MEDAL: A Multimodality-Based Effective Data Augmentation Framework for Illegal Website Identification. Electronics, 13(11), 2199.

[4] Kocyigit, E., Korkmaz, M., Sahingoz, O. K., and Diri, B. (2024). Enhanced Feature Selection Using Genetic Algorithm for Machine- Learning-Based Phishing URL Detection. Applied Sciences, 14(14), 6081.

[5] Kapan, S., and Gunal, E. S. (2023). Improved Phishing Attack Detection with Machine Learning: A Comprehensive Evaluation of Classifiers and Features. Applied Sciences, 13(24), 13269.

[6] Abdul Samad, S. R., Balasubaramanian, S., Al-Kaabi, A. S., Sharma, B., Chowdhury, S., Mehbodniya, A., Webber, J. L., and Bostani, A. (2023). Analysis of the Performance Impact of Fine-Tuned Machine Learning Model for Phishing URL Detection. Electronics, 12(7), 1642

[7] Allouch, M., Mansbach, N., Azaria, A., and Azoulay, R. (2023). Utilizing Machine Learning for Detecting Harmful Situations by Audio and Text. Applied Sciences, 13(6), 3927.

[8] Nagy, N., Aljabri, M., Shaahid, A., Ahmed, A. A., Alnasser, F., Al- makramy, L., Alhadab, M., and Alfaddagh, S. (2023). Phishing URLs Detection Using Sequential and Parallel ML Techniques: Comparative Analysis. Sensors, 23(7), 3467.

[9] Fati, S. M., Muneer, A., Alwadain, A., and Balogun, A. O. (2023). Cyberbullying detection on Twitter using deep learning-based attention mechanisms and continuous bag of words feature extraction. Mathe-matics, 11(16), 3567.

[10] Wu, T., Xi, Y., Wang, M., and Zhao, Z. (2022). Classification of Malicious URLs by CNN Model Based on Genetic Algorithm. AppliedSciences, 12(23), 12030.

[11] Raj, C., Agarwal, A., Bharathy, G., Narayan, B., and Prasad, M. (2021). Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques. Electronics, 10(22), 2810.

[12] Vashistha, N., and Zubiaga, A. (2020). Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media. Information, 12(1), 5.

[13] Malphedwar, L., and Rajesh Kumar, T. (2023).Squirrel search methodfor deep learning-based anomaly identification in videos.

[14] Sagar Dhanake, International Journal of Computer Applications, Volume 174 – No. 28, April 2021, Video Calling through Augmented Reality.