Archives available at journals.mriindia.com

International Journal on Advanced Computer Theory and Engineering

ISSN: 2319-2526
Volume 14 Issue 01, 2025

Securing children from inappropriate and harmful things on the Internet

Akanksha Deshmukh[1], Sejal Jadhav[2], Omkar Bankar[3], Prof. Santosh Kawade[4]
[1,2,3]U.G. Students, Department of Computer Engineering, Dr. D Y Patil College of Engineering and Innovation, Varale, Talegaon Dabhade, Pune, India
[4]Assistant professor of Computer Engineering, Dr. D Y Patil College of Engineering and Innovation, Varale, Talegaon Dabhade, Pune, India

Peer Review Information

Submission: 21 Feb 2025
Revision: 25 March 2025
Acceptance: 30 April 2025

Keywords

Convolutional Neural Networks
Natural Language Processing
Real-Time Content Filtering
Child Online Safety

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.

PROPOSED SYSTEM

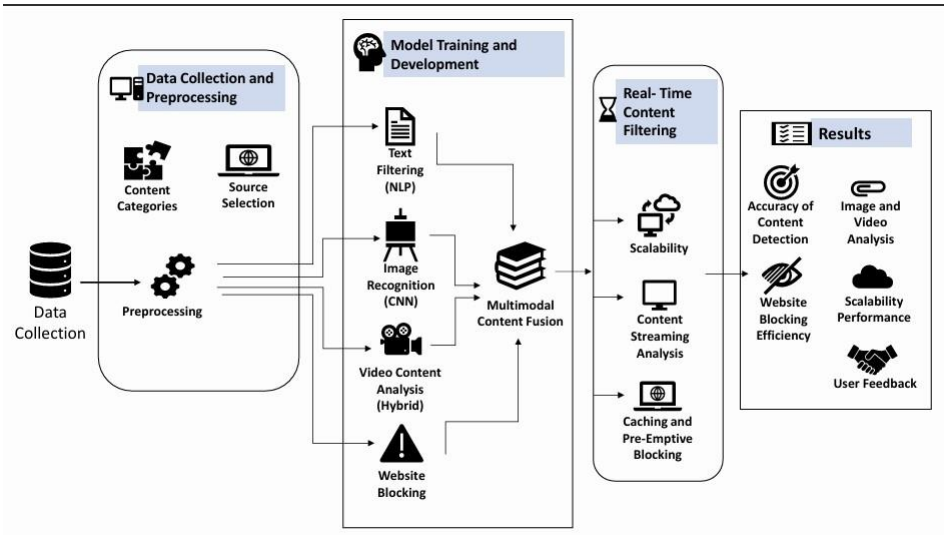
The proposed system is a robust and intelligent browser extension aimed at safeguarding children from exposure to harmful, explicit, or inappropriate online content. It leverages state-of-the-art, pre-trained machine learning models—accessed via secure APIs from platforms such as Hugging Face and AI research initiatives—to enable real-time analysis of web content. The extension dynamically extracts and evaluates textual and visual elements from websites, employing advanced natural language processing models for hate speech and offensive language detection, alongside image classification models for identifying NSFW and violent imagery. Upon detecting harmful content, the system initiates appropriate actions such as content blocking, blurring, or warning overlays, thereby creating a safer and more controlled browsing experience.

Scope of the System

Real-time content filtering: Filters text and images in real-time using pre-trained machine learning models to detect hate speech, offensive language, and NSFW or violent content.
Customizable filtering sensitivity: Allows users to adjust content filtering based on specific needs or preferences.
Real-time alerts: Notifies users when inappropriate content is detected and blocked.

Privacy compliance: Ensures adherence to data protection standards, including COPPA (Children's Online Privacy Protection Act).
Target audience: Aimed at parents, guardians, educational institutions, and any stakeholders looking to create a safer online environment for children**.**

ARCHITECTURE DIAGRAM



ALGORITHMS

1. **RoBERTa (Transformer-Based NLP Model) Purpose:** Detect hate speech in text.

Why Used: State-of-the-art for text classification.
Handles context better than keyword filters (e.g., understands sarcasm or coded language). Pre-trained on large datasets, fine-tuned for hate speech.

2. **Convolutional Neural Network (CNN) (Implied for Images) Purpose:** Detect NSFW/violent images.

Why Used: Best for image classification (spatial feature extraction).
Proven in models like NudeNet/OpenNSFW.

3. **Regular Expression (Regex) Matching**

Purpose: Block exact keyword matches (e.g., "porn", "violence").
Why Used: Simple and fast for exact matches.
Low computational cost (runs locally in the browser). Complements ML models by catching obvious violations.

4. **MutationObserver + TreeWalker**

Purpose: Monitor dynamic page changes (e.g., infinite scroll).
Why Used: Simple and fast for exact matches.
Low computational cost (runs locally in the browser). Complements ML models by catching obvious violations.

5. **Whitelist Filtering (Set Difference)**

Purpose: Allow educational terms (e.g., "sex education") despite keyword matches.
Why Used: Reduces false positives (e.g., blocking anatomy lessons). Uses set theory logic: Blocked = Keywords - Whitelist.

6. **Thresholding (Score > 0.85)**

- **Purpose**: Decide when to block content based on ML confidence.

- **Why Used**:
  - Balances precision/recall (high threshold reduces false positives).
  - Tuneable based on strictness requirements.

MATHEMATICAL MODEL

i. Self-Attention:

$Attention(Q, K, V) = softmax(QKT / \sqrt{dk})V$

Q,K,VQ,K,V: Query, Key, Value matrices (learned during training).

dk: Dimension of key vectors (scaling factor).

ii. Classification Head:

$P(hate | x) = softmax(W \quad h[CLS] + b) \quad (\ ,\ ,\ ) = \{f1, f2, ..., fp\};$

$h[CLS]h[CLS]$: Hidden state of the classification token. W,b: Weights and bias.

iii. Convolution Operation:

$$S_{ij} = (I * K)_{ij} = \sum_{a=0}^{m-1} \sum_{b=0}^{n-1} I_{i+a,j+b} K_{a,b}$$

I: Input image, K: Convolution kernel.

iv. Sigmoid Output: P(NSFW|x)=1/1+e−(W·f(x)+b) $f(x)$: Feature vector from CNN layers.
W,b: Weights and bias.

v. **Thresholding (Score > 0.85)**

Decision Boundary: Block={True if $P(\text{hatel}\, x) \geq 0.85$, False otherwise.

RESULT

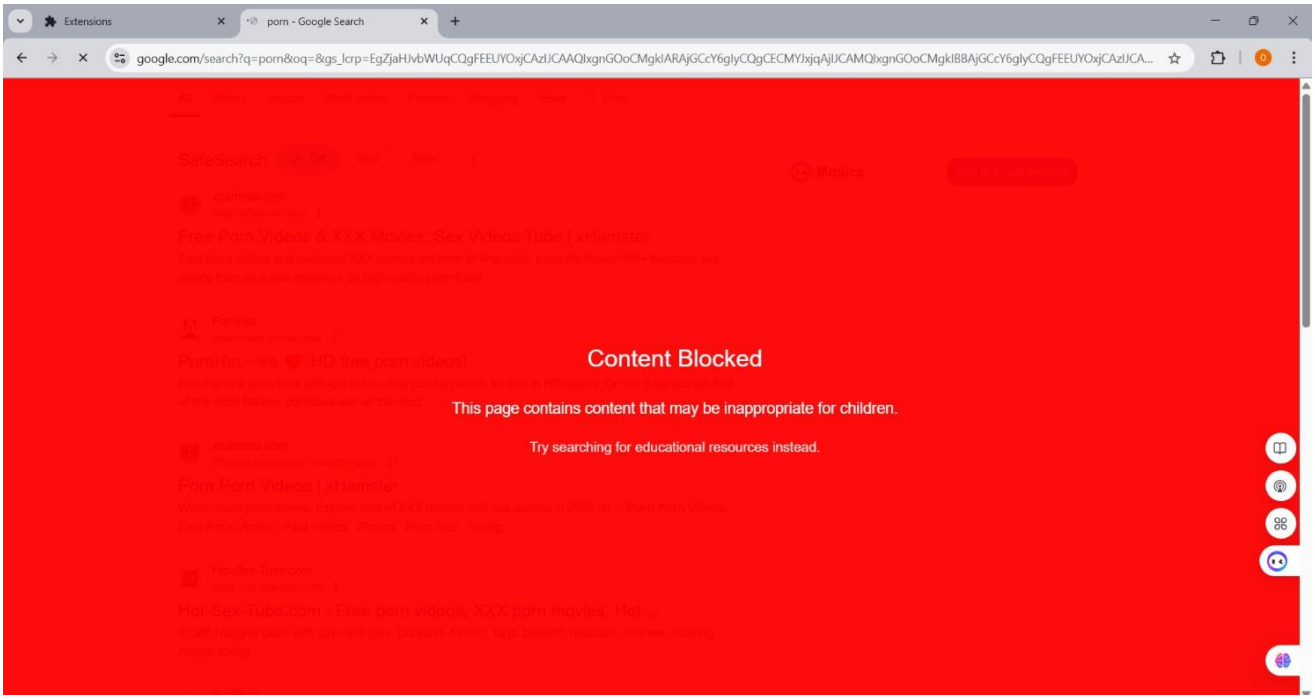


Fig.1 Real-Time Content Blocking of Explicit Websites

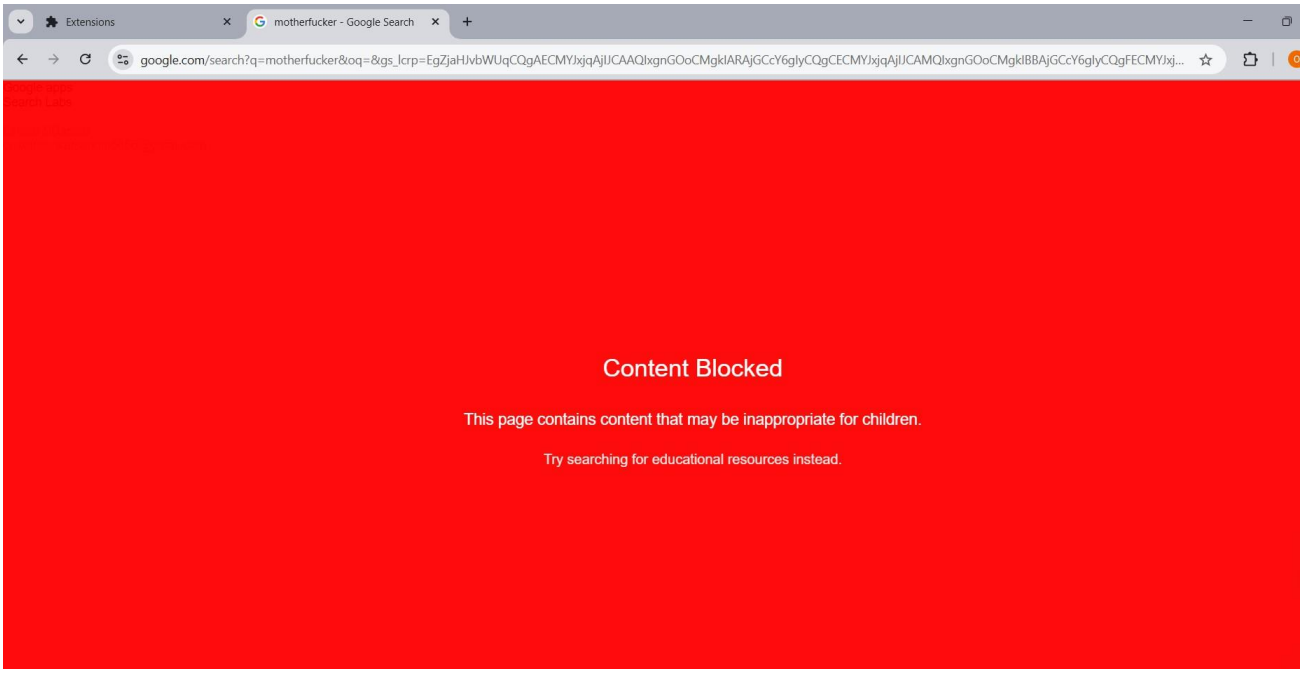


Fig.2 Blocking of Offensive Language-Based Search Queries

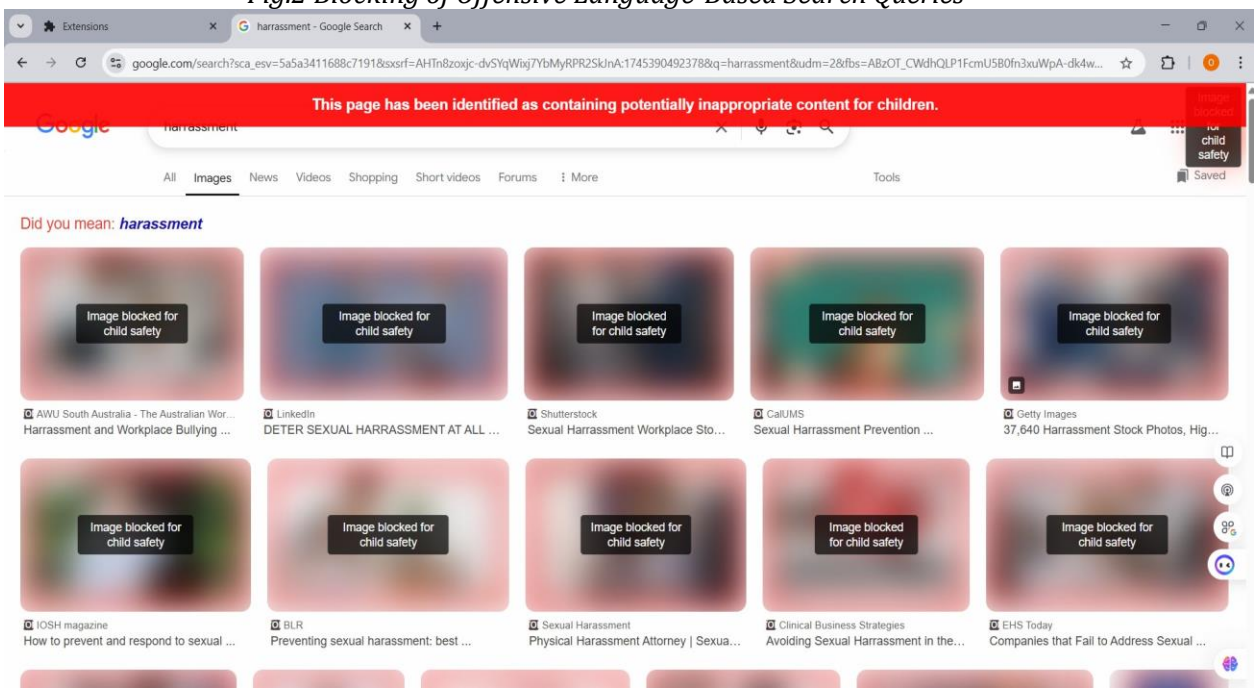


Fig.3 Image Content Filtering Based on Contextual Sensitivity

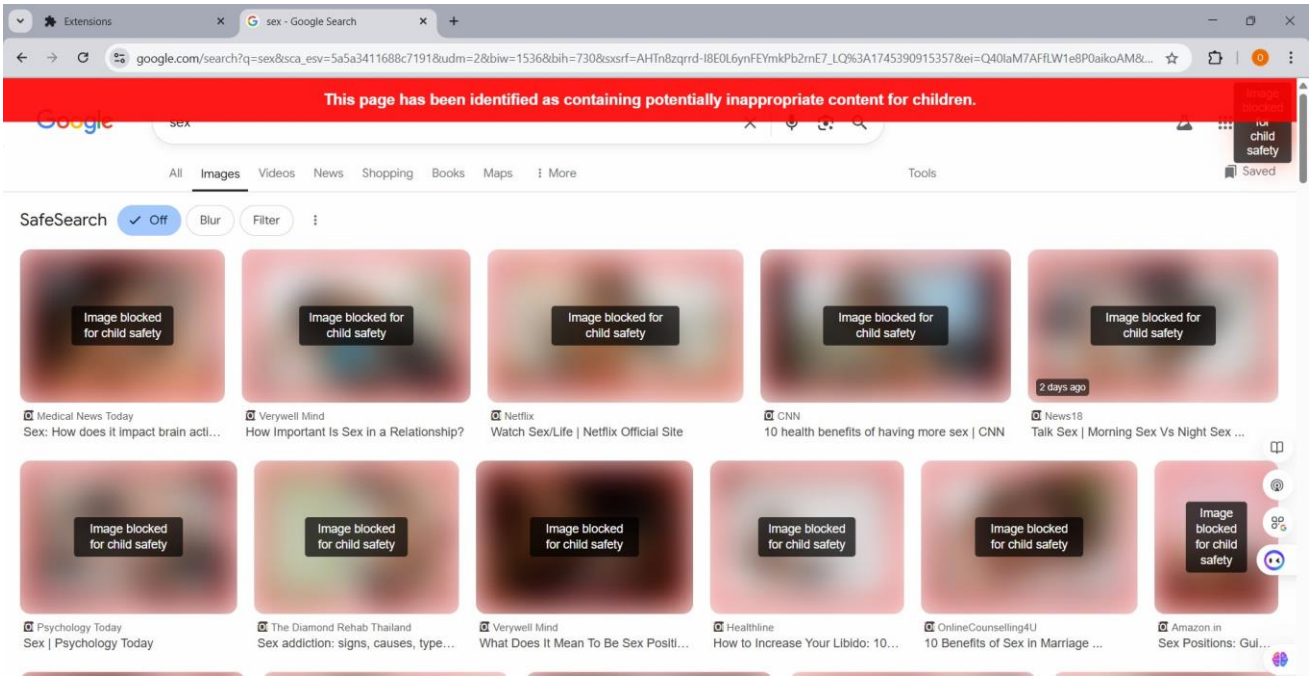


Fig.4 Automated Filtering of Explicit Content for Child Safety

Purpose of Study

This study develops an AI-enhanced browser extension to protect children from harmful online content by integrating machine learning models (RoBERTa for text analysis and CNNs for image recognition) with rule-based keyword filtering, creating a hybrid system that automatically detects and blocks inappropriate material like hate speech, explicit content, and violence in real-time while minimizing false positives; it evaluates the solution's effectiveness in real-world home and educational environments, addresses technical challenges including API dependency and privacy preservation, and demonstrates how intelligent content filtering can enhance online safety for young users without disrupting their browsing experience or requiring constant supervision.

APPLICATION

- **Child Online Safety**: Automatically detects and blocks exposure to hate speech, explicit images, offensive language, and violent content while browsing the internet.
- **Educational Institutions**: Can be deployed in schools and learning environments to create a safe digital space for students using shared computers or online educational platforms.
- **Home Use**: Empowers parents and guardians to protect children from inappropriate online material without needing constant supervision.
- **Content Moderation Support**: Assists in pre-screening web content by flagging harmful elements in real- time, reducing reliance on manual monitoring.
- **Digital Literacy Programs**: Supports initiatives that promote responsible internet usage by demonstrating how AI can assist in creating safer digital environments.

ADVANTAGES

1. **Real-Time Protection:** Instantly analyzes and filters harmful content, ensuring continuous safety while browsing.
2. **AI-Powered Accuracy:** Utilizes advanced pre-trained machine learning models for high-accuracy detection of hate speech, offensive language, and explicit images.
3. **No Manual Monitoring Required**: Automatically filters content without the need for constant supervision by parents or educators.
4. **Cross-Platform Compatibility:** Works across modern web browsers, making it

accessible and easy to deploy.

5. **Non-Intrusive Operation:** Runs seamlessly in the background without disrupting the user experience.
6. **Privacy-Focused:** Designed to operate without collecting or storing personal data, ensuring user confidentiality and compliance with privacy regulations**.**

DISADVANTAGES

1. **Dependency on Internet Connectivity**: The extension relies on external APIs for content analysis; it may not function effectively without a stable internet connection.
2. **API Limitations and Latency**: Using third-party APIs can introduce response delays and may be subject to usage limits or service downtime, which can affect real-time filtering performance.
3. **Privacy Concerns with External APIs**: While the system does not store personal data, the use of external services for content analysis may raise privacy concerns among users regarding third-party data handling.
4. **Browser-Specific Behavior**: The extension's behaviour may vary slightly across different browsers due to inconsistencies in browser extension APIs and permission handling.

CONCLUSION

This project presents a practical and effective solution for enhancing child online safety through a browser extension that leverages pre-trained machine learning models. By integrating real-time text and image analysis using powerful APIs, the system can accurately detect and block harmful, explicit, or offensive content. The extension is designed to operate seamlessly within modern browsers, offering a non-intrusive, privacy-conscious approach to content filtering.

While it depends on internet connectivity and third-party APIs, the system delivers significant value in environments where safeguarding children from inappropriate content is a priority. Overall, the extension demonstrates the potential of AI-driven tools in creating a safer and more responsible digital experience for young users.