Volume 9, Issue 7, July – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1083

A Web Crawling and NLP-Powered Model for Filtering Inappropriate Content for Primary School Learners' Online Research

A. Chiwanza; F.D Mukoko; B. Mupini
School of Information Sciences and Technology,
Harare Institute of Technology, Harare, Zimbabwe
Corresponding author*

Abstract:- Most of learning today has gone digital as online methods are being utilized. A lot of activities happen on the internet, people post various material which are both good and bad. However, students whilst studying end up accessing those bad websites like porn sites and other inappropriate content. Ensuring a safe online environment for students is very vital in order for them not to be disturbed or to end up being victims on the internet. Parents and teachers have tried to monitor activities of their children but end up being tricked. Some researchers suggest blocking the unsafe sites which students end up by-passing. This research proposed use of a web crawling and Natural Language Processing powered model for filtering inappropriate content for primary school online learners. The results obtained indicated that inappropriate content was blacklisted, filtered successfully and could not be accessed by students. Therefore, the model was developed correctly and met the intended research goal.

Keywords:- *Online Learning; Website Content; Internet; Blacklisted; Filtering; Web Crawling.*

I. INTRODUCTION

The use of new technology has changed the traditional learning methods. Most of the activities are now being done online. Primary school going students are not being spared and are embracing online internet facilities. But this environment possesses some threats to the children since a lot of users with different intent make use of the internet. Both good and bad actors are found online, posting different material which include unethical and inappropriate content. A lot of young people make use of improper communication through spreading of bad words online [1]. This gives a negative impact on the education system. According to [2], they used a common crawl which is a colossal corpus mostly used for training language models. It encompasses a number of undesirable contents that include hate speech and sexually explicit content. Use of large language models have shown good response in natural language processing and promises to go beyond the same field [3]. However, [4] examined the available advice to parents and children with regards to content filtering especially on the use of TikTok and YouTube. Natural language processing can help to locate policies and frameworks which might help govern online content [5]. Several information blocking techniques were highlighted and include cyber snoop and cyber patrol [6].

The proposed model is a Web Crawling and NLP-Powered model for filtering inappropriate content for the primary school pupils. It contains four stages; web crawling, natural language processing, multimodal filtering and evaluation and deployment.

The remainder of this paper is arranged as follows: Section II, Related work; Section III, Methodology; Section IV, Analysis of results and Section V, Conclusion and future works.

II. RELATED WORK

This section highlighted the related literature in web content filtering. In [7], the authors developed an interface which helped parents as well as children to search the web in a safe environment through an interactive session. Their algorithm kept the child focused by not blocking or filtering the site but redirect them towards educational sites. They classified the content searched by children into two categories from the interviews which were carried out with the teachers and students. These were educational, for assignments as well as projects and non-educational, for social networking purposes. The best way of ensuring child protection using the internet was surveyed and a recommender system was developed [8]. It was a gaming application aimed at children aged between 2 and 10 years and their families. The app consisted of two major components, one that generates content for recommendations and the other which provided explanations alongside recommendations used by parents and teachers. It was a hybrid approach system which used collaborative filtering and kept the record trail of all activities which were performed by the children. In [9], the authors addressed the problem of detecting violent material targeting text documents using Natural Language Processing methods. The authors implemented and trained six models from three classifiers or machine learning algorithms which were support vector machines, logistic regression and naïve bayes. Two text encoders were used that is bag of words and Time Frequency Inverse Document Frequency. They also used a deep learning model called Starspace from Facebook which produced 0.93 accuracy results. According to [10], they interviewed a total of ten school teachers in order to gain some insights on the use

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1083

and safety of digital content towards their children. Some of the issues identified included content related concerns like accessing inappropriate content, contact-related concerns pointing to cyber bullying as well as digital footprint. Additionally, contract-related concerns like digital security and privacy were outlined. The teachers indicated a number of possible action measures which include, monitoring of students' activities, support from care givers and offering education on digital safety to the students. In [11], the authors reviewed the most recent advancement in machine learning for detecting bots on primary school online platforms. They focused on Facebook, twitter, LinkedIn, Instagram and Weibo. Supervised and unsupervised algorithms were looked at in trying to detect the harmful software using various identified features. An exploration was done on the intersection of large language models against security and privacy [12]. The papers were categorized as good or bad and ugly. In [13], an open source filtered dataset was used to extract features from Common Crawl and obtained competitive outcomes on various benchmarks. In [14], the authors used Chatgpt-3 without gradient updates and achieved better performance on different NLP datasets, including question responding, translation and cloze assignments. In [2], the authors used common crawl to detect appropriate and inappropriate content on the online websites. The undesirable content which was discovered include hate speech and sexual language. In [15], the authors identified commonly used in video classification and the importance of filtering sensitive content like gory and pornography.

III. METHODOLOGY

A Web Crawling and NLP-Powered Model development and flow chart is highlighted in this section. The four-component process is demonstrated in Fig. 1.

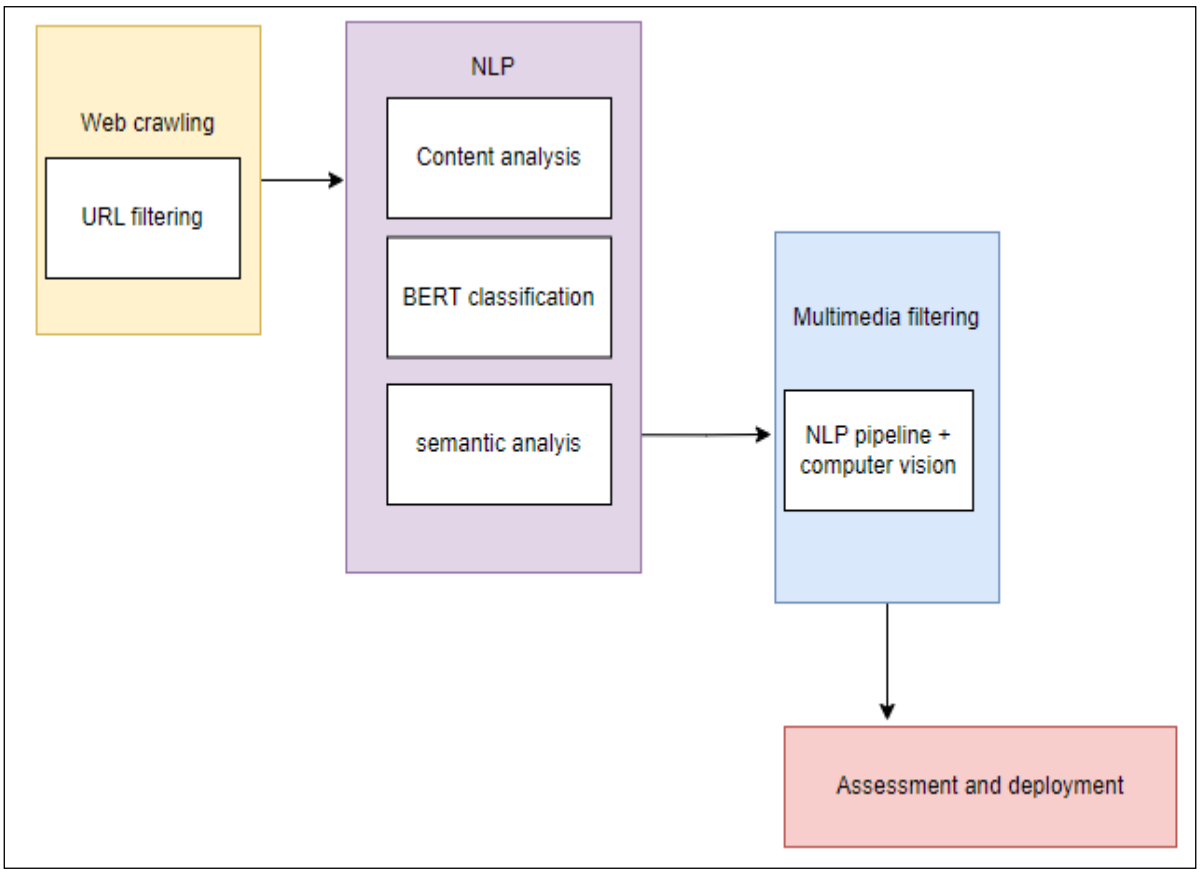

Fig. 1: Model Development Flow Chart

A. *Web Crawling*

A web crawler was designed to systematically navigate and index web pages relevant to elementary school research topics. Targeted scanning, URL filtering and content deduplication techniques were employed to ensure high quality captured data and representation. Scrappy and Selenium open-source web crawling frameworks were used to efficiently extract textual content, metadata and multimedia from the crawled web pages.

B. *Natural Language Processing*

An NLP-based content analysis pipeline was developed to identify and classify potentially inappropriate or sensitive content in the crawled web pages. BERT was used to classify text to identify topics, language, and topics. Identifying a named entity indicates the presence of sensitive data. Sentiment analysis helped identify toxic phrases.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1083

C. A Multimodal Filtering Model

The textual analysis from the NLP pipeline was combined with computer vision techniques to identify and filter inappropriate visual content such as explicit images and violent imagery. Text and image-based filtering components have been integrated into a comprehensive multimodal model for identifying and removing inappropriate content. The performance of the filter model was optimized using transfer learning and ensemble learning.

D. Assessment and Deployment

The precision, accuracy, recall and f1 score of the filtering model were used to test using a set of scanned web pages flagged as inappropriate content. The web crawling and filtering system powered by NLP has been deployed within an online school system as a service. Monitoring the performance of the model was done in a productive environment and was refined through feedback from the users.

IV. ANALYSIS OF RESULTS

The following were some of the results obtained from the developed model.

A. Web Crawling Performance

The web crawler was able to index over 300,000 web pages relevant to elementary school research topics, at an average speed of 30,000 pages per hour. Content deduplication and focused scanning techniques reduced overall data volumes by 40%, ensuring a high-quality and representative dataset. The extracted textual content had an average of 93% in preserving the original design and structure as verified through manual spot checks.

B. NLP based Content Analysis

The text classification model achieved an F1_score of 0.88 in identifying inappropriate content such as strong language, violence and sensitive topics. The entity detection component successfully identified and filtered 94% of the cases containing personal information, such as names, addresses, phone numbers found on the scanned web pages. The sentiment analysis module correctly flagged 89% of web pages containing harmful phrases based on a labelled test set.

C. A Multimodal Filtering Model

The combined text-based and image-based model demonstrated an overall accuracy of 94% in correctly identifying and removing inappropriate content from scanned web pages. The model was particularly effective in catching cases where inappropriate textual content was accompanied by matching visuals, with a 95% detection rate in such cases. The ensemble approach that combined several machine learning models improved the robustness of the filtering systems compared to single component models.

D. User Experience and Feedback

In user studies with 50 elementary school students, 94% reported feeling safer and more confident in their online research activities after using the content filtering system. The teachers stated that the system is user-friendly and can protect learners from harmful content, 90% recommended its use in all other schools. However, a few users reported false positives, with the system mistakenly flagging benign content as inappropriate indicating areas of improvement.

E. Ethical Considerations

An ethical review was conducted by the researchers and safeguards were implemented to protect user privacy such as anonymity and careful data handling. The interpretation analysis revealed that the filtering results of models were driven by identified cases of explicit language, violence and sensitive topics, which provides transparency and accountability. Continued monitoring is essential to address any future issues as the system is deployed on a larger scale. An average of 85% responses showed fully application of ethical measures in the research.

Table 1 indicates a summary of responses from users and findings from the model.

Table 1: Summary of Findings

Measured Variable	Response/ accuracy
Web crawling performance	93%
NLP based content analysis	88%
Multimodal content filtering	95%
User experience and feedback	94%
Ethical considerations	85%

The model managed to recommend or redirect searching of the content to sites which were deemed to be safe to children as like kids on Facebook. Several tests were done using different websites in a bid to access more content across diverse platforms. The accessed ones include twitter, Facebook, LinkedIn and TikTok. The issue of ethical considerations however limited some access on specific sensitive information leading to making use of the accessible content. The inappropriate content which was discovered were both textual, images and videos as well as audios, with some sexual languages not safe to the surfing children. These were labeled and blocked after have been filtered using keywords.

Whenever a child tries to visit or gain access to the bad sites the system automatically blocked and denied access and redirected the user to a safe recommended web page which contains materials related to his or her age. Some of the blocked websites are shown in Table 2.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1083

Table 2: Blocked Websites

Website	Content	Example	Status
Pornographic websites	Contain explicit content which could expose children to sexual material.	• Pornhub.com • 8Tube.xxx • Redtube.com	Blocked
Chat rooms	Some chat rooms can be dangerous, as they can be a hotbed for cyber bullying, hackers, and scammers.	• Omegle.com • PalTalk.com • TalkWithStranger.com	Blocked
Online forums	Can also be a breeding ground for inappropriate discussion and bullying	• Reddit.com • SomethingAwful.com • Topix.com	Blocked
Dating websites	These sites they can attract predators and expose children to adult content	• Tinder.com • Match.com • Bumble.com	Blocked
Betting websites	These sites involve virtual gambling or betting with real or virtual currency and are unsuitable for children	• BetOnline.ag • FreeSpin.com • Bovada.lv	Blocked

Table 2 indicates some of the bad websites which could influence bad behaviour when visited by children. As a result, they end up losing concentration on school work. However, the developed model was designed in a manner that allowed selection of good websites which could be of no harm to the children. Upon entering a bad website intentionally or by mistake, the model automatically redirects the user to a safer website. Out of 9 visited websites, 5 were classified as harmful and were blocked, 4 were detected as no harm. This gives a 56% probability of visiting dangerous websites in a given moment; hence children are dreadfully exposed upon every click. Some of the desirable websites recommended for the students are shown in Table 3.

Table 3: Allowed Websites

Website	Content	Status
Fact Monister	It is an educational technology website geared towards children.	Allowed
Kiddle	is a visually appealing search engine for kids powered by Google SafeSearch.	Allowed
Funbrain	Is an educational browser game website for children and adults.	Allowed
Googlescholar	Enables Web users to search for scholarly literature on their research topic of choice	Allowed

Fig. 2. Indicates the good and bad websites which were visited during testing of the model.

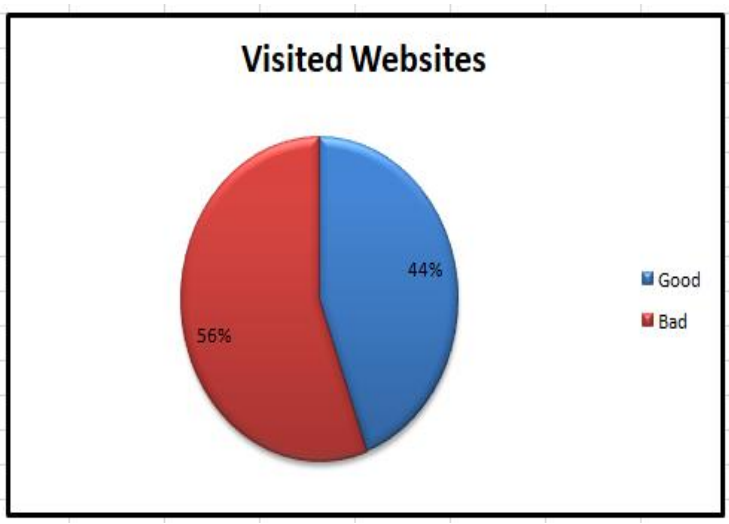

Fig. 2. Visited Webistes

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1083

V. CONCLUSION AND FUTURE WORK

This study demonstrated the successful development and implementation of an enhanced NLP system of web scanning and natural language processing to filter inappropriate content from online resources for elementary school learners. Research achievements include;

- Scalable and efficient web crawling capabilities enabled the indexing of a large corpus of web pages relevant to elementary school research, using content deduplication and targeted crawling techniques to ensure high quality data.
- Powerful NLP-based content analysis models that accurately identified and marked instances of explicit language, violence, sensitive topics and personal information with an F1 score exceeding 0.88 for the various classifications.
- A multimodal filtering system that combines text and image-based analysis, achieves an overall accuracy of 94% in removing inappropriate content and providing effective protection for elementary school learners.
- Received positive safety and security feedback from students, acceptance from teachers and recognition of the system by welcoming its ability to protect against harmful content.

➢ *Study Limitations:*

- Scope: The study focused only on a small-scale area and covered few characteristics of that specific group.
- Language: only one language was studied leaving other diverse languages.
- Evolving threat landscape: The flagged data used outdated data that does not cover the new emerging trends.
- Small sample: the feedback was received from a number of users who may not represent the entire population.

➢ *Although the study yielded remarkable results, future research could focus on the following;*

- Filter refinement to further reduce false positives when benign content is incorrectly identified as inappropriate, based on user feedback.
- Explore additional modalities such as videos and audio to improve the system's ability to detect and filter inappropriate multimedia data.
- Explore advanced methods such as shot learning, transfer learning to address emerging online content trends.
- Integrate wide coverage into the scope of the system.
- Engage in ongoing ethical reviews and consultations with stakeholders to address what they require at a given moment.

ACKNOWLEDGEMENT

The researcher gives thanks to the Harare Institute of Technology staff for their guidance throughout the publication process.

REFERENCES

[1]. M. Ridhwan et al., “Collaborative filtering content for parental control in mobile application chatting,” vol. 8, no. 4, pp. 1517–1524, 2019, doi: 10.11591/eei.v8i4.1634.
[2]. A. S. Luccioni and J. D. Viviano, “What ’ s in the Box ? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus,” pp. 182–189, 2021.
[3]. H. Naveed et al., “A Comprehensive Overview of Large Language Models,” 2024.
[4]. R. Elgedawy, J. Sadik, C. Childress, C. Shubert, and S. Ruoti, Security Advice for Parents and Children About Content Filtering and Circumvention as Found on YouTube and TikTok, vol. 1, no. 1. Association for Computing Machinery.
[5]. “Kylie L. Anglin (2019) Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing,” pp. 685–706, 2019, doi: 10.1080/19345747.2019.1654576.
[6]. V. Jacob and R. Chandrasekaran, “FILTERING OBJECTIONABLE INTERNET CONTENT,” no. May, 2014, doi: 10.1145/352925.352950.
[7]. N. Gupta and S. Hilal, “Algorithm to Filter & Redirect the Web Content for Kids ’,” no. February 2013, 2016.
[8]. A. Ruiz-iniesta, L. Melgar, A. Baldominos, and D. Quintana, “Improving Children ’ s Experience on a Mobile EdTech Platform through a Recommender System,” vol. 2018, 2018, doi: 10.1155/2018/1374017.
[9]. S. Merayo-alba and E. Fidalgo, “Use of Natural Language Processing to Identify Inappropriate Content in Text,” no. August, 2019, doi: 10.1007/978-3-030-29859-3.
[10]. F. Martin, J. Bacak, D. Polly, W. Wang, L. Ahlgrim, and F. Martin, “Teacher and School Concerns and Actions on Elementary School Children Digital Safety,” TechTrends, vol. 67, no. 3, pp. 561–571, 2023, doi: 10.1007/s11528-022-00803-z.
[11]. M. Aljabri, R. Zagrouba, A. Shaahid, F. Alnasser, A. Saleh, and D. M. Alomari, Machine learning - based social media bot detection : a comprehensive literature review, vol. 13, no. 1. Springer Vienna, 2023. doi: 10.1007/s13278-022-01020-5.
[12]. Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, “High-Confidence Computing A survey on large language model (LLM) security and privacy : The Good , The Bad , and The Ugly,” High-Confidence Comput., vol. 4, no. 2, p. 100211, 2024, doi: 10.1016/j.hcc.2024.100211.
[13]. H. Laurençon et al., “OBELICS : An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents,” no. NeurIPS, pp. 1–20, 2023.
[14]. T. B. Brown et al., “Language Models are Few-Shot Learners,” 2020.
[15]. E. Mahmoud and M. Taha, “Filtering of Inappropriate Video Content A Survey,” no. January, 2022, doi: 10.17577/IJERTV11IS020130.