Web Mining for Suspicious Keyword Prominence

Mayank Tiwari[1], Rakshit Bhatt[2], Mayank Goel[3], Ayush Garg[4], Siddeshwari Dutt Mishra[5]

CSE, FET, MRIIRS, Faridabad, Haryana 121003

Abstract
The exponential growth of online data and content in recent times has precipitated a surge in malicious activities, including the incitement of youth to engage in illegal behaviour ,exposure to unsuitable material, propagation of false information, hate speech, and other related concerns. Therefore, there is a demand for efficient web content screening mechanisms. The detection and mitigation of these threats necessitate the deployment of sophisticated and complex techniques capable of efficiently processing large volumes of web data. This research project proposes a web mining strategy for identifying suspicious keyword prominence related to harmful activities and topics, with the aim of developing a proactive approach to mitigating online threats.

Keywords: **Web mining, Data extraction, Tags removal, Dictionary approach**

INTRODUCTION

Web mining is the technique to discover and extract valuable information from the vast amount of data available on the world wide web. This field combines techniques from data mining, machine learning, artificial intelligence, statistics and information retrieval to identify patterns , trends and relationships within the web data.

Web mining includes utilizing information mining strategies on web information to introduction unacceptable fabric, proliferation of untrue data, despise discourse, and other related concerns. When it comes to detecting suspicious keyword prominence, web mining techniques can be used to analyze the density and frequency of selected keywords on the web pages. Websites may employ tactics to artificially increase the visibility of specific terms by overusing or manipulating keywords to enhance their prominence, which could be sign of malicious content.
By using web mining to monitor keyword prominence, it's possible to proactively identify and mitigate suspicious online activities before they can cause any harm.

LITERATURE REVIEW

1. **Aakash NE Gandhi et al.** utilized multiple machine learning algorithms in the paper Identify the Online Spread of Terrorism Through Data Mining to extract textual information from web pages and determine their relevance to terrorism. Their work demonstrates the potential of machine learning in identifying and filtering web content related to terrorist activities.

2. **Chen, H. et al.** used sentiment analysis features to segregate and classify words on a web page, assigning a score to each word in "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion This study emphasizes the importance of sentiment analysis in classifying web content according to its emotional tone, which is vital for identifying harmful or suspicious material.

3. **Fawad Ali et al.** explored various methods to fetch, scan, and analyze textual data for countering terrorism on online social networks using web mining techniques. Their research emphasizes the importance of web mining in Observing and Addressing the Spread of Extremist Content.

4. **Naseema Begum et al.** focused on classifying web pages into various categories and sorting them using data mining and web mining techniques. This study illustrates how categorization and sorting can help in managing and filtering large volumes of web data for specific purposes.

5. **T. Anand et al.** created a system that combines data mining with web mining techniques for enhanced efficiency in development. They employ a method to monitor and analyze web pages, demonstrating the synergy between these two techniques in creating effective web monitoring systems.

6. **Raymond Kosala & Hendrik Blockeel**provides a comprehensive overview of the emerging field of Web mining, which integrates methods from diverse research domains such as databases, information retrieval (IR), and Artificial Intelligence (AI), particularly machine learning and natural language processing, utilizes synonyms and rephrasing techniques to enhance understanding and communication.

7. **Chakrabarti's** offers a thorough survey of the Web mining domain, exploring its core methodologies and applications. The paper categorises Web mining into three key areas: content mining, structure mining, and usage mining.

8. **Liu** offers an in-depth examination of the core principles of web data mining. This extensive review examines the three primary domains of Web mining: links, content, and usage data, providing insights into both traditional techniques and new developments.

9. **Agerri et al.** present a comprehensive survey of Web mining research, focusing on the methodologies and applications that have evolved up to that point. The paper categorizes web mining into three primary types: content mining, structure mining, and usage mining, providing a comprehensive overview of the research landscape in each area.

10. **Huan and Huang** provide an in-depth exploration of Web mining techniques, focusing on the extraction of knowledge from hyperlinks, content, and usage data. Their paper offers a detailed examination of the methodologies and challenges associated with each aspect of Web mining, presenting a holistic view of the field.

11. **Muhammad Jawad Hamid Mughal** provides a comprehensive examination of the landscape of web data mining, an area of increasing relevance due to the exponential growth of online data. The authors systematically evaluate different methods, tools, and algorithms related to web data mining, providing a range of valuable resources for researchers and practitioners in the field for their work.

RESOURCES USED

A. Visual Studio

Visual Studio, a widely-used integrated development environment (IDE) created by Microsoft, offers an extensive suite of tools for software development, including a code editor, debugger, and project management features. Visual Studio supports various programming languages, including C#, VB.NET.

B. Python Libraries

- **Beautiful Soup:** Beautiful Soup is a Python library for web scraping and data extraction. It provides a simple and intuitive way to parse HTML and XML documents, enabling developers to extract data from web pages with ease. Beautiful Soup is often used in conjunction with other libraries, such as Requests, to build powerful web scraping applications.

- **Requests:** Requests is a library used for sending HTTP requests. It provides a straightforward and user-friendly interface for making HTTP requests and receiving responses. Requests supports numerous HTTP methods such as GET, POST, and PUT, allowing developers to seamlessly interact with web services and APIs.
- **OS:** The OS module is a standard Python library designed for interfacing with the operating system. It provides a set of functions for performing tasks such as file manipulation, directory creation, and system information retrieval. The OS module allows developers to create applications that are independent of the platform and can operate across various operating systems.

C.HTML

HTML (Hypertext Markup Language) is a coding language utilized for building web pages. It offers a collection of tags and attributes to outline the structure, content, and design of web pages. HTML is frequently paired with CSS (Cascading Style Sheets) and JavaScript to develop dynamic and interactive web applications. HTML is an essential technology for web development and is compatible with all contemporary web browsers.

METHODOLOGY

Introduction

The approach that we have taken involves multiple steps to identify suspicious web pages. Firstly, we use a web scraping technique to extract all data from the web page. After the complete extraction of data from the web page we use our custom built “clean data” function to remove HTML tags and extract only the text content.

Suspicious Keyword Detection

For identification of suspicious keywords, we employ a dictionary-based approach. We have compiled a comprehensive dictionary of predefined words that are indicative of malicious content. Our program compares the extracted text content with this dictionary to detect any matching words. Upon detecting a matching word , our program flags the word and enables a frequency analysis to determine the number of occurrences of the word within the extracted content.

Frequency Analysis and Threshold Evaluation

After the frequency analysis, our program evaluates the number of occurrences of the flagged word against a predetermined limit. If the frequency of the word exceeds the predefined limit for the occurrence of that word, then that web page is categorized as suspicious.

Output and Reporting

After the web page has been classified as suspicious , our program generates a report that includes the extracted content where the suspicious word is discovered. The said report provides valuable insights into the nature of the suspicious content , enabling further analysis and mitigation strategies to be employed.

Technical Approach

The approach that we have used leverages a combination of natural language processing(NLP) and ,machine learning(ML) techniques to analyse web page content and identify the suspicious keywords. The dictionary-based approach which we have used enables efficient and effective detection of malicious content, while providing a robust mechanism for minimizing false positives.

PROPOSED SYSTEM

We have implemented a dictionary-based approach for the initial detection of the suspicious words from the website. The total percentage of the suspicious words is evaluated relative to the total website content. If the count percentage exceed the 30% criteria the website will be marked as suspicious and further the machine learning algorithms are applied on the content for the analysis of the sentiment for those suspicious words. We have applied a range of machine learning techniques, such as Vader for the sentiment analysis of the textual content and the nltk for the text processing and tokenization.

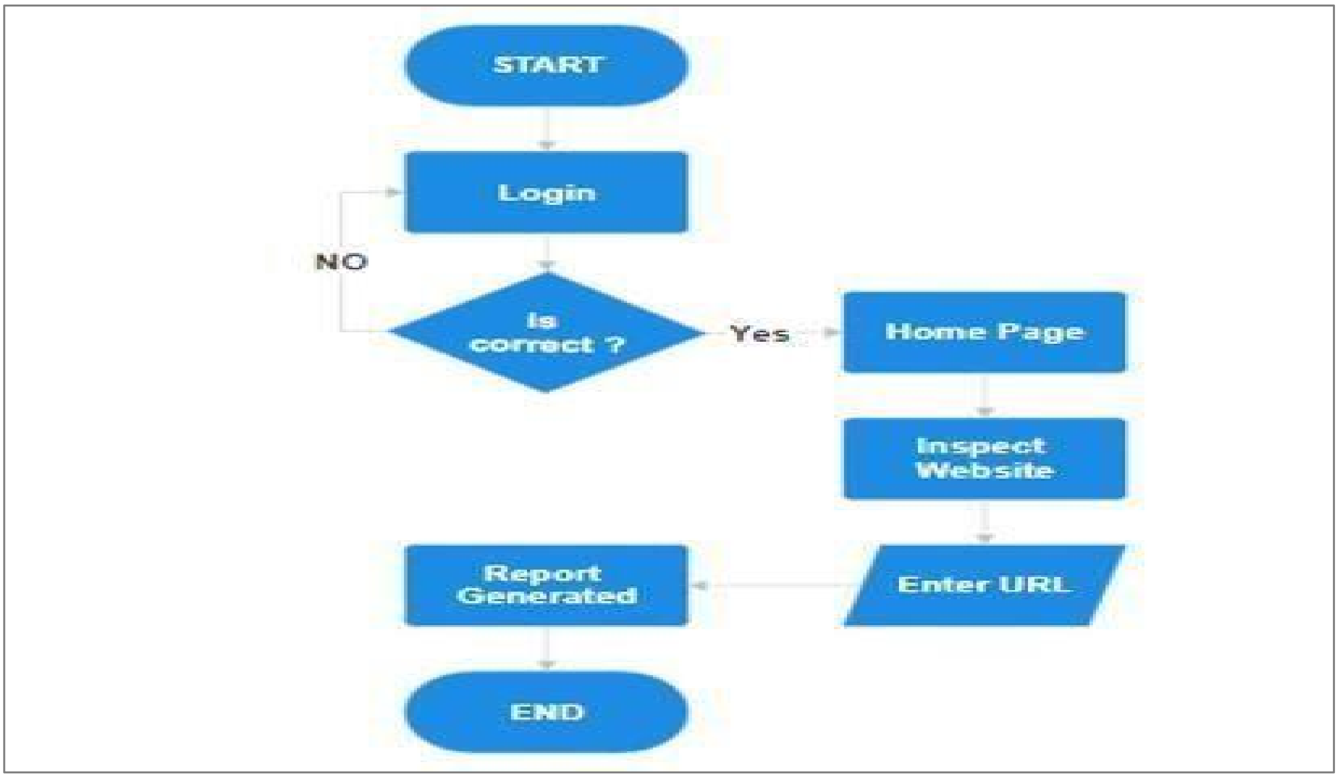

FIGURE 1 : Flowchart Depicting steps for inspecting the content of website

All above features are encapsulated in a website. The user needs to sign in on the website to use the functionalities of the website. Once logged in, the screen to enter the URL of the website opens up. The user provides the website URL which he/she wants to analyse and then the suspicious content percentage is counted and sentiment analysis is performed on the content, followed by a screen displaying the results, shown below

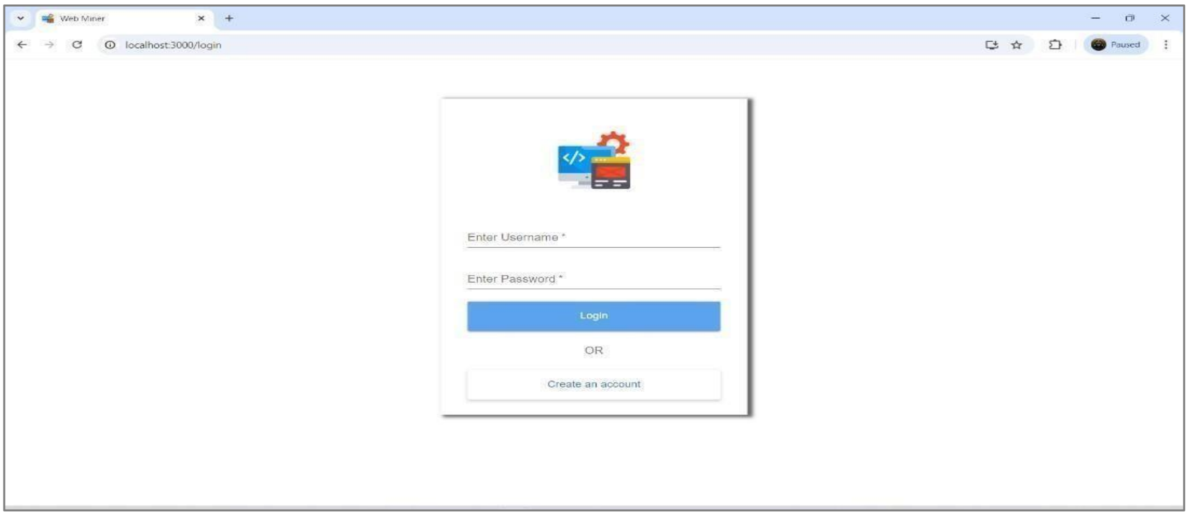

FIGURE 2 : Login Page

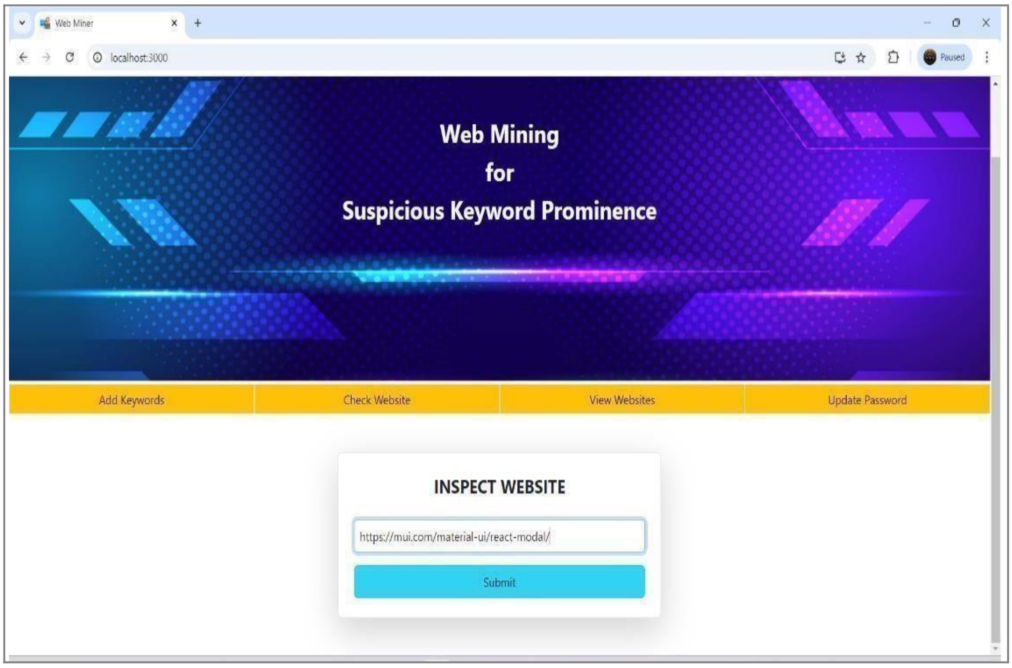

FIGURE 3 : Home Page

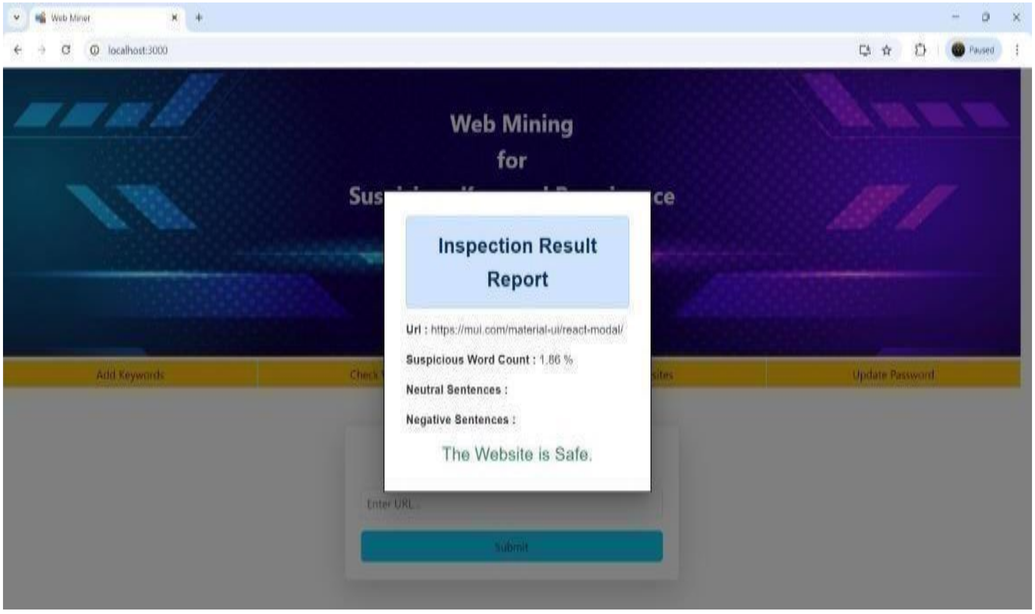

FIGURE 4: Inspection Report displayed

CONCLUSION

In conclusion, the proposed web mining strategy for identifying suspicious keyword prominence has demonstrated its potential in mitigating online threats.

By leveraging web mining techniques, this approach can proactively identify and report suspicious online activities, enhancing online safety and security. The methodology employed, which involves web scraping, data extraction, suspicious keyword detection, frequency analysis, and threshold evaluation, provides a robust and efficient mechanism for identifying malicious content. The use of Visual Studio, Python libraries such as Beautiful Soup, Requests, and OS, and HTML, has enabled the development of a sophisticated and complex system capable of processing large volumes of web data. The output and reporting features of the system provide valuable insights into the nature of suspicious content, enabling stakeholders to take informed decisions. This research project has contributed to the development of a proactive approach to mitigating online threats, and its findings have implications for the development of more effective web content filtering mechanisms. Ultimately, this project has the potential to make a significant impact in enhancing online safety and security, and its outcomes can inform the development of more effective strategies for mitigating online threats.

FUTURE SCOPE

The project can be used in any number of ways such as to protect from cybercrimes as well as to investigate the suspicious things happening on the internet. The future of the project can be fruitful to everyone as to be protected and secure, web dump and unnecessary activities may limit down and strict laws can be implemented if anyone is caught by the algorithm of the project. Our project can useful to the law department and the security of any company as if they can take a precaution and can detect the addresses or risks of the suspicious things happening in the internet in the further models of the project of web mining for suspicious keyword prominence

REFERENCES

[1] A. Negandhi, S. Gawas, P. Bhatt, P. Porwal. "Identifying the Online Dissemination of Terrorism via Data Mining." *IOSR Journal of Engineering*, vol. 13, April 17, 2019.

[2] 2. H. Chen. "Opinion Mining Across Multiple Languages: Feature Selection for Classification in Web Forums." *ACM Transactions on Information Systems*, upcoming, June 2008.

[3] 3. F. Ali, F.H. Khan, S. Bashir, U. Ahmad. "Counter-Terrorism on Online Social Platforms Using Web Mining Strategies." Department of Computer Science, Federal Urdu University of Arts, Science and Technology (FUUAST), Islamabad, Pakistan.

[4] 4. N. Begum. "Detection of the Online Spread of Terrorism Using Web Data Mining." *International Journal of Advanced Research, Ideas, and Innovations in Technology*, vol. 5, issue 1, Institute of Engineering and Technology, Coimbatore, Tamil Nadu.

[5] 5. T. Anand, S. Padmapriya, Kirubakaran. "Tracking Terrorism Using Advanced Web Mining." *2009 International Conference on Intelligent Agent & Multi-Agent Systems*.

[6] Kosala, Raymond, and Hendrik Blockeel. "Web Mining Research: A Survey." ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, 2000, pp. 1-15.

[7] Chakrabarti, S. (2003). Web mining: Information and pattern discovery on the World Wide Web. ACM Computing Surveys (CSUR), 31(4es), 1-40.

[8] Liu, B. (2006). Web data mining: Exploring hyperlinks, contents, and usage data. Data Mining and Knowledge Discovery, 12(1), 1-15.

[9] Agerri, R., Wang, D. A. S. W., & Faulds, L. P. G. P. C. (2004). Web mining: A survey of current research. Data Mining and Knowledge Discovery, 8(2), 77-98.

[10] Huang, C. L., & Huang, J. K. (2009). Mining the Web: Discovering knowledge from hyperlinks, contents, and usage data. IEEE Intelligent Systems, 24(2), 15-25.

[11] Muhammad Jawad Hamid Mughal. Data Mining: Web Data Mining Techniques, tools and Algorithms: An Overview June 2018 International Journal of Advanced Computer Science and Applications DOI:10.14569/IJACSA.2018.090630LicenseCC BY 4.0