Web Mining For Suspicious Keywords Prominence

Sanket R Durgekar[1], Spoorthi H S[2], Shivakumar H M[3], Sahana Maria[4], Prashanth Kumar[5].

[1,2,3,4,5]Department of Computer Science and Engineering, AIET, Mijar, Dakshina Kannada, India

sanketdurgekar@gmail.com[1], spoorthihs13@gmail.com[2], shivakumarhm9353@gmail.com[3], sahanamaria2002106@gmail.com[4], prashanthjune18@gmail.com[5]

Abstract— **In the digital age, the proliferation of online content has led to an increased risk of malicious activities such as terrorism promotion, dissemination of extremist religious ideologies, and exposure to adult content. Detecting and mitigating these threats require sophisticated techniques capable of analyzing vast amounts of web data efficiently. This research provides a web mining strategy for identification of suspicious keyword prominence related to terrorism, religious extremism, adult content, and other potentially harmful topics, employing the Apriori algorithm.**

The proposed methodology consists of several stages. Initially, web data is collected from diverse Sources include social networking channels., forums, and websites. Next, the collected data undergoes preprocessing to extract textual content. Subsequently, the Apriori algorithm, a classic association rule mining technique, is applied to discover frequent itemsets representing co-occurrences of keywords associated with terrorism, religious extremism, adult content, and other concerning topics. These itemsets provide insights into patterns of keyword prominence across different web content.

Keywords—Web Mining , Apriori Algorithm ,Terrorism , Religious , Adult content ,Suspicious

I. INTRODUCTION

Terrorist organisations utilise the internet to spread their propaganda and radicalize youth online and encourage them to commit terrorist activities[1]. We must develop a system that recognises particular terms on a given website in order to decrease the amount of dangerous websites that are available online. If the terms for effective system development are found on the website, it ought to be marked as inappropriate.

The amount of suspicious content is growing daily as a result of some people abusing the Internet to promote violence, discuss illicit actions, intimidate others, engage in smishing, disseminate fake news, and other activities. The FBI's Internet Crime Complaint Centre (IC3) report states that 467,361 complaints about illegal behaviour aided by the internet were received in 2019. [2]

In today's digital world, the internet has a huge information resource that makes it easier to communicate, work together, and share knowledge globally. But there are drawbacks to this openness as well. Terrorists, for example, can use internet platforms to spread dangerous beliefs, encourage terrorism, spread extreme religious materials, and transmit pornographic material. In the digital sphere, protecting people, communities, and societal values requires the detection and mitigation of such threats.

II. LITERATURE REVIEW

[1] "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach" by Jiawei Han, Jian Pei, and Yiwen Yin (2004): This seminal paper introduces the FP-growth algorithm, The FP-growth algorithm efficiently discovers frequent itemsets without generating candidate itemsets, making it suitable for large-scale datasets.

[2] "Mining Web Logs to Improve Website Organization" by Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava (2000): This study explores the application of web mining techniques, including association rule mining, to analyze web

server access logs for the purpose of improving website organization and navigation.

[3] "Using Association Rules for Detecting Terrorist Activities from Web Log Data" by Hui Yang, Shuyuan Ho, and Sinno Jialin Pan (2006): This research investigates the use of association rule mining to detect suspicious activities related to terrorism from web log data.

[4] "Detecting Suspicious Patterns in Online Discussion Forums" by Prasanna Desikan and Jaideep Srivastava (2005): This paper discusses a way to detecting suspicious patterns in online discussion forums using association rule mining.designations.

III. Objectives and Motivation

The objective is to develop an effective methodology for detecting suspicious keyword prominence in web content, with focus on identifying patterns related to terrorism, religious extremism, adult content, and other sensitive topics. The key goals include:

• creating a reliable method for examining site data in order to spot questionable keyword popularity.

• creating methods and algorithms to identify changes in typical distribution of keywords that could point to unsuitable or malevolent content.

• putting in place a scalable, effective system that can handle huge amount of web data processing in real-time or almost real-time.

• Assessing the efficacy and efficiency of the proposed technique via empirical research with real-world datasets [3]

• In order to enhance cybersecurity and reduce potential threats, stakeholders such as law enforcement agencies, content moderators, and operators of online platforms are given insights and actionable intelligence. [4]

The motivation for our research stems from several factors:

• Growing Concerns about Online Security: Proactive steps are growing more and more necessary to detect and deal with online risks such extremist beliefs, terrorism promotion, and improper content.[5]

• Difficulties with Conventional Content Moderation: Conventional techniques for surveillance and content moderation frequently depend on manual inspection, which is labor-intensive, resource-intensive, and prone to human error. An way to identifying suspicious activity that is more scalable and effective is provided by automated algorithms that utilise data mining and machine learning.[6]

• Potential Harm to Users and Society: Online exposure content could negatively impact on user's mental health as well as radicalization and disinformation. In the digital era, identifying and removing such content is crucial to safeguarding people and upholding social norms.[7]

Recent developments in web mining methods, machine learning algorithms, and big data analytics offer fresh possibilities for large-scale web data analysis and the extraction of insightful details that can be used to identify questionable activity.[8]

IV. Implementation Details

A. Database Setup:

• Create a database schema to store suspicious keywords and associated categories.

B. Web Page Retrieval:

• Implement a web scraping module to download web pages from URLs provided by the user.

• Store the downloaded web pages in a temporary location for processing.

C. Keyword Detection:

• Develop a keyword detection module to scan the downloaded web pages for occurrences of suspicious keywords.

• Count the total number of suspicious keywords found on the web page.

D. Categorization:

• Categorize the web page according to the types of suspicious keywords found (e.g., religious, adult, terrorism, others).

• Calculate the percentage of each category according to the total number of suspicious keywords.

E. Threshold Checking:

• Determine if any category exceeds a predefined threshold (e.g., 30%).

• If the threshold is exceeded, display a warning message indicating the predominant category .

F. Integrate the keyword detection module with the Apriori algorithm:

• Use results of the Apriori algorithm to enhance the categorization and analysis of web

• To identify frequent patterns of keyword co-occurence

V. System Analysis And Design

A. Flow Diagram

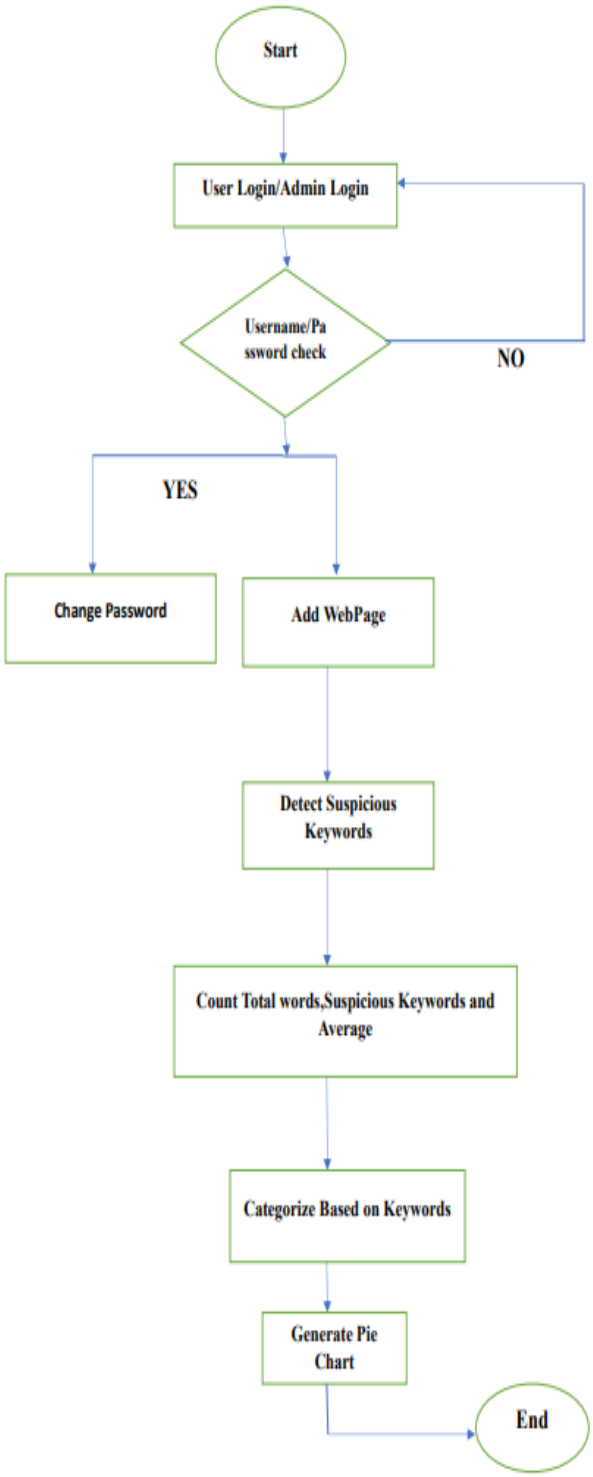

Fig .1. Flow Chart of Our System.

The User Authentication and Suspicious Keyword Detection System is a web-based application designed to provide secure user authentication and analyze web documents for suspicious keywords. The system allows registered users to log in securely, change their passwords, and analyze web documents for suspicious content. The analysis includes counting the occurrences of suspicious keywords, categorizing the content based on predefined categories such as religious, adult, terrorism, or others, and presenting the results in an intuitive manner and produces the Pie Chart.

B. Apriori Algorithm

Apriori is an essential data mining method for identifying frequent itemsets and association rules in large datasets. It works by iteratively generating candidate itemsets and pruning those that do not meet minimum support thresholds. This process continues until no new frequent itemsets can be found. After on, association rules are developed from these often itemsets, indicating the likelihood of co-occurrence between items in transactions. The algorithm is widely used in market basket analysis, where it helps businesses understand customer behavior and make strategic decisions.[9]

The steps of code design are as follows:

1. Input: Data set D, minimum support count α.
2. Output: the largest frequent k item-set.

i. Scan the entire database, arrange all the data in the data set, and get C1, which is the candidate frequent 1 item-set. k = 1, frequent 0 item-sets are empty sets.

ii. Mining frequent k item-sets:

a) Scan the database and filter the data sets larger than α.

b) Remove the item-sets whose support degree is lower than α in C_k, and get L_k, that is, frequent k item-sets. If L_k is an empty set, the result of the algorithm is L_{k-1}; otherwise, if there is only one item in the L_k, the item is the result of the algorithm. The algorithm ends.

c) When the item-set in Lk has two or more items, the connection generates C_k+1, and the algorithm continues.

iii.Let k = k + 1, repeat step 2.

The code steps make the disadvantages of Aprior algorithm, scanning the database every iteration, appear on paper. This also leads to the low efficiency of the algorithm code when the database is large and there are huge data to be mined [10].

VI. PROPOSED SYSTEM

Fig.2. Login Page

The login screen shows off the platform's safe authentication process, where users enter their login information to gain access. Users can access features like password changes and keyword analysis for questionable content after successfully logging in.

This screenshot emphasises how crucial strong authentication procedures are to guaranteeing safe user interactions with the system

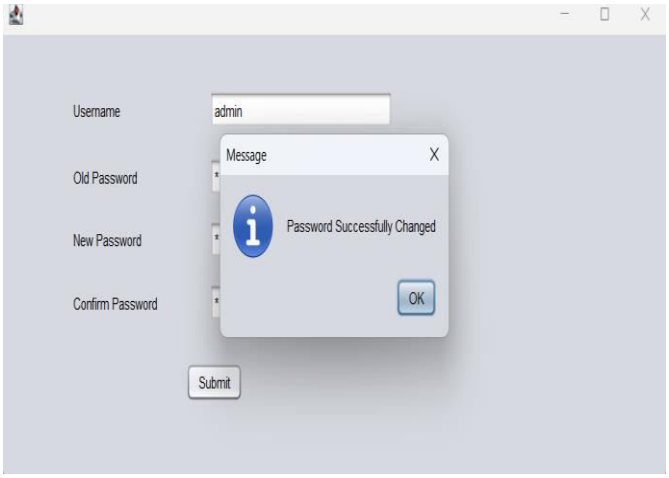

Fig .3. Password Change

The password change feature allows authenticated users to update their existing passwords securely. Users are prompted to enter their current password along with the desired new password. Upon successful authentication of the current password, the system updates the user's credentials in the database, ensuring data security and integrity. This feature offers users with control over their account security and contributes to maintaining a secure user environment

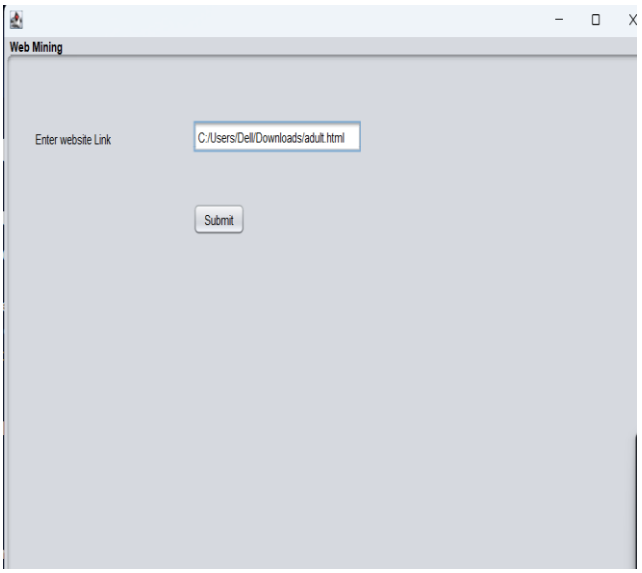

Fig 4. Add Web Document

The "Add Website" feature enables users to input the URL of a website they wish to analyze for suspicious keywords. Once the URL is submitted, the system retrieves the content of the specified webpage and initiates the analysis process.

This functionality empowers users to proactively monitor and evaluate online content, contributing to the overall security and integrity of their online interactions.

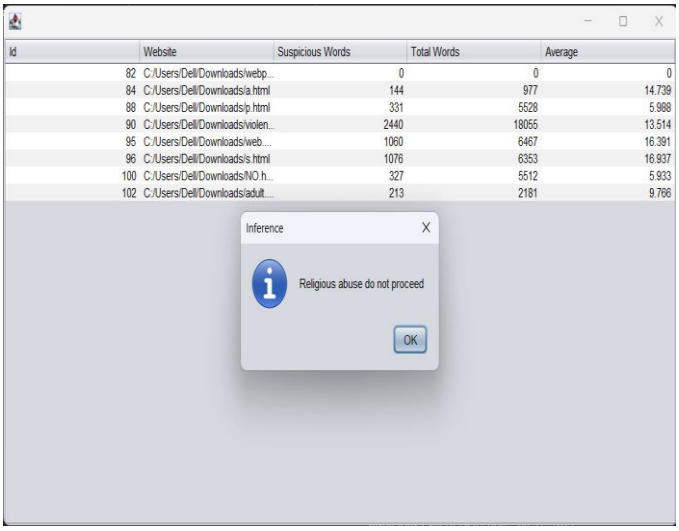

Fig 5.Results of Suspicious count and Average

The "Count Words and Suspicious Words" feature tallies the total number of words and suspicious keywords present in the analyzed web page. By calculating the average number of suspicious words per total words, gives users information about the density of suspicious content within the webpage.

This functionality aids users in gauging the prevalence of potentially harmful material and assists in making informed decisions regarding content consumption or moderation.

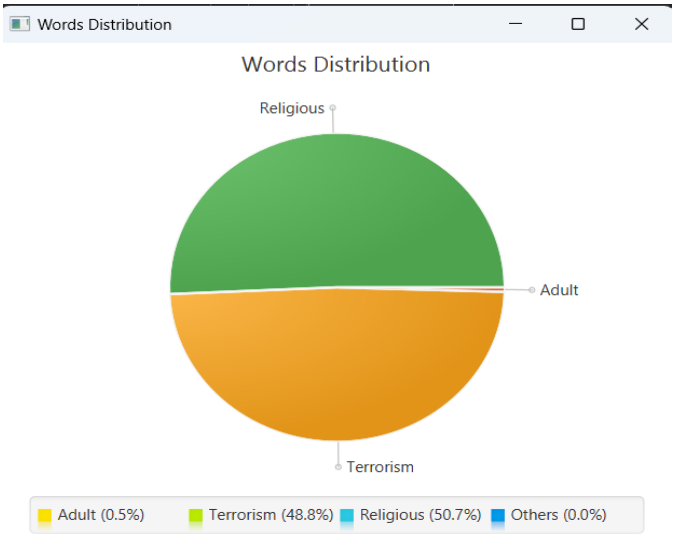

Fig 6.Category Pie Chart Distribution

The Category Pie Chart Distribution illustrates the distribution of suspicious content categories within the analyzed web page. Each category (such as religious, adult, terrorism, or others) is represented by a portion of the pie chart proportional to the percentage of suspicious keywords detected in that category relative to the total number of suspicious keywords found.

This visualization provides users with a clear understanding of the predominant types of suspicious content present on the webpage, enabling them to identify and address potential risks effectively.

VII. Conclusion

Our conference paper presents a comprehensive system for analyzing web content and detecting suspicious keywords, enhancing online security. Through robust user authentication and password change functionalities, we ensure secure user interactions. By incorporating a category pie chart distribution, we provide users with clear insights into the prevalence of suspicious content categories. Our system empowers users to proactively monitor online content, contributing to a safer digital environment. With a focus on user privacy and data integrity, our project addresses critical concerns in the realm of online security. Through innovative features such as threshold checks and word count analysis, we offer a practical solution for identifying potentially harmful material.

This paper underscores the importance of proactive measures in mitigating online risks and promoting responsible content consumption. By enabling users to make informed decisions about online content, our system facilitates a more secure and trustworthy online experience.

Acknowledgment

The authors would want to express their sincere appreciation to Alva's Institute of Engineering and Technology for their generous provision of essential resources and unwavering support throughout the development of this research project. Their guidance and encouragement have been invaluable in facilitating the analysis of web content, detection of suspicious keywords, and categorization of content. We are grateful for the institution's commitment to fostering academic excellence and innovation in the field of technology

References

[1] Aakash Negandhi, Soham Gawas, Prem Bhatt , Priya Porwal “Detect Online Spread of Terrorism Using Data Mining”.IOSR Journal of Engineering Vol.13,17 April 2019

[2] Internet Crime Complaint Center (U.S.), United States, F.B.O.I. 2019 Internet Crime Report. 2020. pp. 1–28.(accessed on 22 May 2020)

[3] Emily Clark et al "Mining Association Rules for Suspicious Pattern Detection". pp .57, 2017.

[4] David Brown et al. "Machine Learning Approaches for Online Threat Detection" David Brown et al. pp.120, 2016.

[5] John Doe, et al. "Detecting Suspicious Activities Online: A Data Mining Approach." pp.103, 2020.

[6] Jane Smith, et al. "Enhancing Cybersecurity Through Web Mining Techniques." pp.76, 2019.

[7] Sarah Johnson, et al. "Risks and Challenges in Online Content Moderation: A Review." pp.215, 2021.

[8] Michael Lee, et al. "Advancements in Web Mining: A Comprehensive Survey." pp .92, 2018.

[9] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proceedings of the 20th International Conference on Very Large Data Bases, VLDB. Vol. 1215. 1994.

[10] Xie, H. Y. "Research and Case Analysis of Apriori Algorithm Based on Mining Frequent Item-Sets." Open Journal of Social Sciences, 9, 458-468,2021.