Web mining for suspicious keyword

1Miss Vaishnavi Deshmante, 2Mr. Bhavin Shah, 3Mr. Anuj Mutha, 4Miss Rupal Agrawal, 5Prof G.B. Katkade

1,2,3,4Student, 5Head of Department, Computer Technology, K.K. Wagh Polytechnic, Nashik, India

Abstract: With the rapid development of Internet, we have entered an era information explosion there is a lot of redundant information in the network. As there is increased growth in terrorism, and other activities there is the need of the hour to put a full stop to these types of suspicious activities with lead to the destruction of the youth. It has become very important to detect suspicious keyword patterns from online messages, websites and pages to stop the activity. The project aim at web mining for finding out density of selected keywords in order to check its keyword prominence on those web pages. Web mining technology came into being, and saved out of the human from the information ocean. Web mining can be termed as an information mining method to naturally search, collect and organize data from indexed online records, which might be in various organized, unstructured or semi-organized structure. We usually use web-mining techniques in order to assess the viability of a specific web page/entity in order to figure out various factors related to it. This project consolidates the best-researched mechanisms from the semantic web and synaptic web at low entropy in order to build structural engineering of Semantic-Synaptic web mining. Our proposed projects aim at web mining for finding out density of selected keywords in order to check its keyword prominence on those web pages.

Index Terms: Suspicious keywords, Word detection, Web mining, Websites, Keyword prominence

I. INTRODUCTION

This project brings a new way for future security enhancement to detect apprehensive words from a respective website, webpage or any web portal. Concept of this project is to detect suspicious keyword/statement prominence from the website/webpage or a portal and to count the number occurrence to verify the harmlessness of the software. As increase in technology where users are surrounded by social media networks sites, it is very effortful job to track the suspicious words. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information onto a comprehensive structure.

Web mining is the application of data mining techniques to discover patterns from the World Wide Web. As the name proposes, this is information gathered by mining the web. It makes utilization of automated apparatuses to reveal and extricate data from servers and web reports/pages. The big challenge is the recognition of messages containing suspicious words or some other web content. Web pages created in different platforms can be tracked using this portal. To address the drawback in investigation technique we are proposing a data mining-based framework model, based on web mining, which will be able to work with larger amount of data. This system will check web pages whether the web pages is promoting terrorism or any suspicious activity. Our system can be integrated with vast range of social media platform such as Facebook, Instagram, you tube’s comment section area that can easily find out for any abusive/suspicious/offensive or vulgar statement by matching the phrases or keyword in particular from the portal’s library [of suspicious word] with actual system that is being analyze and the same word or phrase will be added to spam list. In same way, the portal can also be integrated with the search bar of any browser, consider Microsoft edge, if any user is searching for offensive words then he will get certain result in form of web links and he/she will visit any one/more of them. If the portal carries the rights to access the database of that web site the portal is able to find out that user by legal use of database. This is an important factor in various fields in order to check the prominence of specific keywords and flag websites having over prominence of those keywords. For example, this can also be used to detect websites inspiring terrorism. The system can be used to check prominence of a particular keyword commonly used in terrorist convincing strategies on various websites. So, this system can be used to point out websites having more prominence of such keywords. Thus, this system can also be used in several different domains for a variety of important purposes using smart web mining approach. Figure 1 Shows the steps performed in the search suspicious data discovery
Traditional Techniques
In an attempt to combat this problem, computer security researchers have developed tools and techniques to help users identify and avoid malicious content. Static methods, such as blacklists, often have a tremendous success rate when it comes to identifying attacks, though they are often difficult to maintain due to the short lifespan of malware and phishing sites.

II. LITERATURE SURVEY
There are certain technologies, methods and algorithm available to detect the keyword prominence but they have certain disadvantages and limitations.

Michael Robertson, Yin Pan and Bo Yuan [1] that explained about the social approach to detect malicious content in web technologies for Facebook with security heuristics is limited to identify malicious URL links or websites.

John Resig, Ankur Teredesai [2] in their paper explored Framework for IM and various data mining issues and how they relate to Instant Messaging and current Counter-Terrorism efforts. Moreover, this paper detects does not tend to fully detect suspicious keyword, and not even their prominence and rank.

S. Gowari, G.S. Anandha, G. Divya [3] explained the paper that there are tools available to detect word but they analyze the network level data and head level data but the tools used are outdated. This paper explains the need to detect suspicious keyword pattern with efficiency.

The above papers describe may use different technique to capture suspicious work but nowadays users are uses tricks like sending short words or code words which is not so easily understood by admin. Our proposed algorithm will help to detect these words from the websites and webpages effectively.

III. COUNTER TERRORISM AND DATA MINING
Since September 11, 2001, and the increased terrorist activity against the United States, the area of Counter Terrorist Data Mining has seen a surge of interest and papers relating to applications of old data mining techniques to a new field of study. Most papers attempt to utilize the study of Social Network Analysis in order to find potential links between suspicious groups of people. Two such works include the Mapping Networks of Terrorist Cells [7] and A CBR Approach to Asymmetric Plan Detection [8]. The first paper concerning Counter-Terrorist activity is that of Mapping Networks of Terrorist Cells [7] and directly concerns the social network analysis surrounding the Sept. 11 attacks. This paper attempts to reveal the difficulty in attempting to find direct correlating activity or suspicious words between a sparsely related group of people. Mentioned in this paper is the tracking of Task Data Sources, which includes the use of Chat Rooms and Instant Messaging. The paper ‘A CBR Approach to Asymmetric Plan Detection’ [8] attempts to coordinate a social network of people and places with links of seemingly trivial actions. The concerns of this work surround three major issues: Massive data sets, noise, and incomplete information. In our framework we deal with similar issues: • Develop algorithms for handling streams of data as it collects to perform clustering and averaging of the information being transmitted.

The research topics presented by current counterterrorist research represent the challenges that exist within this field. By adapting the algorithmic techniques (such as Social Network Analysis and Textual Topic Detection) presented as solutions, new techniques can be adopted.

IV. PROBLEM STATEMENT/CONCEPT
Web mining techniques are used in order to access the viability of a specific web page/entity in order to figure out various factors related to it. The system uses web mining algorithms to mine textual information on web pages and detect their relevance to suspicious activity. Website created on different platforms can be tracked using this application. Data mining is a technique used to mine patterns of useful data from large data served to make the most use of obtained results. Data mining as well as web mining...
are used together for efficient system development. Our system will detect patterns, keywords and relevant information in unstructured texts in websites using web mining as well as data mining. Web mining also consists of test mining methodologies that are used to scan and extract useful content from unstructured data. Government officials who work for security can also use this system. This project brings a new way for future security enhancement to detect apprehensive words from a respective website, webpage or any web portal. Concept of this project is to detect suspicious keyword/statement prominence from the website/webpage or a portal and to count the number occurrence to verify the harmlessness of the software. The big challenge is in automated message surveillance, websites and webpages is the recognition of messages containing suspicious words or some other web content. Our project aims at web mining for finding out density of selected keywords in order to check its keyword prominence on those webpages. This is an important factor in various fields in order to check the prominence of specific keyword. This system can be used to point out websites having more prominence of keywords. Thus, this system can also be used in several different domains for variety of important purpose using smart web mining approach. The word detection and graph generating will be as follows:

![Graph representation](image)

**V. WORKING**

The concept of this system is to detect suspicious activity over the web. For this purpose, the system provides Admin module and a user module where the admin can enter websites, scan them, detect words, disable website from viewing and disapproving the user.

**Admin**

The admin is the main handler on the system where all the decision will be taken by the admin only. The admin first determines the words based on the category and enter different categories.
Once the category as added the next work of admin is to create a database of keyword, which may be present in the website based on the categories. Admin adds different keywords and creates as database where the keyword will word as a reference to determine the rate of suspicious activity being performed. The keywords will be added based on the category created by the user.

One of the major jobs of admin is to add websites to check the vulnerable activities. The URL being added by the admin will be check for the keywords and the rate and graph will be generated accordingly.

The admin has the supreme rights to decide whether a website is suspicious or not for the admin can disable the websites for the user to stop spreading of any further. The admin or any forensic authority and have all the website detail in a glance :how many suspicious words it may contain, what is its occurrence where is the occurrence, it rate etc. in a tabular format.in just a glance of an eye the authority can disable the website from further viewing by just clicking on the disable button.

If the website is not suspicious, it is available for the user to be viewed and shared. On the other hand, the user can view the website along with the table of keywords and their count only if the admin has permitted it to be a non-suspicious website and safe to be shared.
VI. SYSTEM CONTEXT DIAGRAM

All the major activities like the user admission, User authentication, adding keywords to the library, adding the URL and managing the spam list. Added keywords will be detected based in their severity and their prominence. If a particular website has higher number of Suspicious, keywords then the user, which is logged in the system will not be able to view the website, as it will be blocked by the admin as it is detected as suspicious. The History module will maintain the history of the activities. The database will store all the suspicious words, their prominence and the date and place of occurrence the webpage.

ADMIN: Admin module provides setting option for administrator for listing out of keyword, which are to be scanned. Admin will analyze scanned URL to find suspicious keywords. Admin module also authenticate by receiving admin username and password. This module also adds keywords to the library in the database. Admin module of the duty id to change old password with the new password and confirm the new password. Add keywords, remove keywords and match keywords is the work of the admin module. Another major responsibility of the admin is matching keywords, matching URL and sorting URL. Admin module also monitors all the process in the system like add keywords, removing keywords, matching keywords, specifying URL, matching URL and sorting them. Adding and confirming new password. The most important task of the authenticate and authorize the user.

ADD KEYWORD: This module allows Administrator to add or admit new suspicious/offensive/abusive keywords into library of suspicious words. Such keywords will also be considered while comparing the keywords in the process of data mining from website. Added keywords will be detected based in their severity and their prominence. If a particular website has higher number of Suspicious, keywords then the user, which is logged in the system will not be able to view the website, as it will be blocked by the admin as it is detected as suspicious.

ADD URL: This module requires a URL to analyze and make respective decision depending on suspiciousness of the website. If the report says that the URL specified is suspicious then the module add such URL into spam list and further the user will not be allowed to access the content of URL.

HISTORY: The purpose of history module is to maintained record about recently verified URL. Record includes time and date of verification activity of each URL. The history of the activities will be maintained future suspicious words, their prominence and the date and place of occurrence the webpage.

All these modules are clearly depicted in the following System Context Diagram:

VII. SEQUENCE DIAGRAM

Sequence diagram consist of three objects i.e. user, admin and database. When user will perform registration admin object will detect, validate and further save the details in the database.

Later on, when the user will attempt to login then admin will again validate the details and provide acknowledgement to the user.
In next part if there is requirement for addition of new keywords that represents suspiciousness or some offensive activity, then such keywords can be added into the database (suspicious keyword library). Here the admin will pick the keywords and add manually in the database.

Now, Suppose user feels that the website which he or she desires to use is suspicious than he can confirm about its nature by just inserting the URL of the website in the respective field. Here the URL will be Analyzed by the admin, meanwhile scanning the website if admin detects the keyword is suspicious by comparing the data with keyword library then the detailed Report will be visible to the admin.

Depending upon the reports or statistics generated while detecting or analyzing the website if reports says that the occurrence of the offensive keywords reaches up to a certain level of tolerate then the webpage associated with URL will be added to the spam list. In addition, if not the detected as suspicious then website will be considered as normal webpage.

The following diagram simply depicts interaction between objects in a sequential order i.e. the order in which the interaction will take place:

![Sequence Diagram](image)

**VIII. ADVANTAGES**

1. Security secures the internet from illegal activities. The system is able to provide security from activities.
2. This system can be used in several different domains for a variety of important purposes by integrating them with other system by using smart web mining approach.
3. Keep an eye on Activity Helps the authority to keep an eye on unusual activities and illegal activities.

**IX. DISADVANTAGES**
The only disadvantage of the project is that it still marks the websites suspicious even if the words are not marked in that manner.

**X. APPLICATIONS**

1. Efficient use of web mining the project use web-mining algorithms to mine textual information on webpage and detect their relevancy to suspicious activity. Tracking of information of data.
2. Tracking of information of data. Website created in different platform can be tracked using the application.
3. Web sorting this system will classify the web pages into various categories and sort them appropriately i.e. normal website and the websites, which promote suspicious activities.
4. Integrated with vast Range of social media our module can be integrated with vast range of social media platform such as Facebook, Instagram, you tube’s comment section area that can easily find out for any abusive/suspicious/offensive or vulgar statement by matching the phrases or keyword in particular from the portal’s library.
XI. **FUTURE WORK/SCOPE**
We aim to update our system that will automatically scan the websites which in now done manually. We are also working to detect codes from the websites as well, the data being transferred may not be textual format always we are working to detect and encode different symbols as well.

XII. **CONCLUSION**
To curb the menace of suspicious/vulnerable/apprehensive activities, this project is developed to detect and terminate websites, which are suspicious, which spread vulnerable/harmful contents to used radicalizing youth. By deploying this software, we have been able to bring the word detection mechanism in its simplest from for the authorities to detect and stop the unwanted growth of suspicious activities. This application is to develop to stop suspicious activities by preventing radicalization of the people through websites.

**REFERENCES**