IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DARK WEB MONITORING SYSTEM FOR EFFECTIVE ENUMERATION OF URL'S FOR DETECTING, ANALYSING AND FILTERING OF VARIOUS ILLEGAL ACTIVITIES.

[1]Munazirul Islam, [2]Rakesh Shetty, [3]Poojary Manish, [4]Aditya S Poojary, [5]Dr. Nagaveni V

[1, 2, 3, 4] Student, Department of Computer Science and Engineering, AMC Engineering College, Bengaluru, Karnataka, India.

[5] Professor and Head of Department (DS), Department of Computer Science and Engineering, AMC Engineering College, Bengaluru, Karnataka, India.

Abstract: Dark web monitoring is a crucial aspect of cybersecurity and law enforcement, focusing on the surveillance and analysis of activities within the dark web, a hidden network accessible only through specialized software like Tor. This hidden network provides a haven for various illegal pursuits, including cybercrime, drug trafficking, and illicit trade. Detection, filtering, and analysis of illegal activities are essential components of dark web monitoring. Advanced technologies and methodologies, including machine learning algorithms, facilitate the identification and categorization of illicit content and behaviours across forums, marketplaces, and communication channels. Cybercriminals make full use of the hidden network called as Dark Web and they are hidden from the normal search engines which are publicly facing, to sell a variety of illicit products and services. A highly automated modular system called D-Monitor is described to help with the gathering and analysis of massive volumes of unstructured data. The system relies on a Python module named OnionSearch to collect URLs based on keywords we already define in our tool. and search through various search engines like Ahmia, tor66, and torgle. These keywords are not limited to already defined keywords, but more number of keywords can be added if required and expand our scope of data collection. The dark web's anonymity services safeguard user information protection, but they also provide the ideal setting for illicit and criminal operations including drug and gun trafficking, people trafficking, and illicit information transfers. A Python dark web monitoring crawler which is built by utilizing the tor network was created to gather and store a huge number of dark web site URLs in order to gain more information about the websites by searching using keyword parameter and find suspicious information and track criminals utilising the dark web to carry out illicit operations. The tool takes screenshots using the tbselenium module and displays them on the frontend, preserving anonymity and avoiding the need to rely on the Tor browser User Interface.

Index Terms **-** DarkWeb Monitoring system, Crawling of DarkWeb, Detect anomalies on dark web, Effective URL enumeration, Onion crawling using python, tbselenium module to crawl dark web.

I. INTRODUCTION

The internet's evolution has revolutionized communication, information sharing, and business, but the dark web's secrecy has made it a hub for cybercriminals, leading to a complex landscape of illegal trade, drug trafficking, Child pornography, Arms dealing and Cyber Crime. The challenge of dark web monitoring lies in efficiently enumerating URLs associated with anonymous channels on the Tor network, which are not indexed by traditional search engines. Detection, filtering, and analysis of various types of malicious activities carried out by the criminals on the dark web are crucial components of an effective monitoring strategy.

The terms "web" and "Internet" are different but still share some common features. The Deep Web refers to material not indexed by search engines like Google or Firefox— information that can't be reached using regular search engines because it's not cataloged. In contrast, the web provides access to such data through these search engines. This material is further divided into the Dark Web— inaccessible through regular web browsers and deliberately hidden.

Users are able to exchange data on the Dark Web with minimal risk as website owners remain anonymous and concealed. The US Naval Research Laboratory's Onion Routing (TOR) project — initiated in 2002 — supports online anonymous communication by routing encrypted user traffic through a network offering more reliable infrastructure capabilities than others called Invisible Internet Project (I2P). The vast infrastructure and myriad networks comprising Internet connect millions of computers together: irrespective of this immense reach however, certain specific types of material on it can only be accessed anonymously due their secretive nature.

Anonymous and decentralised nodes of certain network groups (such as TOR or I2P) may be used to access the dark web. With TOR, users may visit websites anonymously by spreading data over public networks through virtual tunnels. It creates "relays" on computers carrying information through its tunnels worldwide, with encrypted information placed between them. The final relay, called the "exit relay," has an IP address that resembles the TOR traffic source, allowing software to hide user addresses.

Using email, web chats, or similar platforms housed in TOR may greatly enhance security, anonymity, and privacy while communicating online. Overall, the Internet and the Dark Web provide different approaches to navigating the internet and maintaining privacy.

Since the 2000s, research on the Dark Web has primarily focused on terrorists and extremist groups, but you can find very less information about the collective gathering of various types of illegal activities on the Dark Web, where majority of the illicit activities and malicious activities are not covered. There are several activites on the darkweb that take place. These markets launch new markets in various majors, offering a wide range of products and services, including drugs, weapons, pornography, malware, software exploits, hacking tutorials, botnet rentals, trading documents, fake IDs, Government leaked data, stolen credit cards, and hitmen. Accessing the Dark Web, data extraction and structure, data cleansing and transformation, and data mining are the numerous steps needed to generate useful information from dark web resources. Dark web monitoring involves a systematic approach to efficiently enumerate URLs, detect, filter, and analyse illegal activities. the first stage involves using web crawling and scraping techniques to identify and catalog '.onion' URLs, explore known entry points, forums, and marketplaces, and categorize based on keywords. The second stage is to verify these collected URL and data by whether they are already up and running, because the data URL which we collect in the first stage contains mix of both alive and dead onion sites. The third stage is to take screenshots of whole web page using tbselenium module using python algorithm and store it in the backend and matching with the corresponding URLs. The fourth stage involves displaying these content on the front end using react js and HTML, CSS and javascript.

Content filtering mechanisms are developed using keyword analysis as well as already provided keyword to the system. Anomaly detection algorithms are employed to identify deviations from established norms, such as sudden spikes in activities or changes in communication patterns. A continuous monitoring system is implemented to keep pace with the dynamic nature of the dark web, updating the URL database, refining detection algorithms, and adapting filtering mechanisms. Legal and ethical compliance is ensured, respecting privacy rights, adhering to jurisdictional regulations, and prioritizing ethical handling of information.

II. LITERATURE SURVEY

The article compares and contrasts the surface web, deep web, and dark web, focusing on their fundamental differences and interdependencies. Additionally, it provides technical details on the technology that supports the three most prominent darknets—Freenet, Tor, and I2P—and examines the idea of the Dark Web. Furthermore, it examines the potential impact of law enforcement actions on the surface web on the Dark Web, the "dilemma" of usage that anonymity technologies pose, and the challenges that law enforcement agencies face in their efforts to combat and prevent crime and terrorism on the Dark Web. [1]

The Surface Web is the part of Internet made up of websites that can be accessed via search engines like Google and Firefox. The Internet, on the other hand, is a network of computer networks and infrastructure. There is also an unindexed portion of the Internet known as the "Deep Web" which cannot be accessed by traditional search engines. A portion of the Deep Web known as the "Dark Web" is accessible via TOR and operates covertly and anonymously. Three things are offered by specialised browsers like TOR and I2P: anonymity, privacy, and non-detection. This essay explores the Dark Web's impact on several facets of society, including the quantity of everyday anonymous users in Kosovo and throughout the globe. Although the Dark Web's anonymity isn't entirely guaranteed, TOR makes a point of offering anonymous activities. [2]

Content analysis on the Dark Web is essential for spotting cyberattacks and locating important players. Whether integrated or independent, this field's research aids in the fight against cybercrime. Examining contemporary research on Dark Web content analysis for Cyber Threat Intelligence (CTI), this study discusses methodologies, methods, tools, strategies, and constraints. It emphasises the value of researching various Dark Web sites and guides novice researchers through cutting-edge techniques. Future developments in the field, ethical issues, and technological difficulties are also included in the review. [3]

Tor hidden service protocol is used by dos website on the dark web to host dangerous content. Detecting and monitoring these sites is important for law enforcement and computer security. Dos website content is accessed and analyzed using an electronic engine called LIGHTS. Dos Tag Automation Tool (ATOL) is a service that many people can use to perform thematic analysis of content on a website. The tool has three main components: keyword search mechanism, classification framework, and category. The results show ATOL's performance compared to other datasets such as the LIGHTS dataset and the 20 Newsgroups dataset. [4]

This paper investigates the methodology used to acquire onion addresses for exploring the Tor network, which are vital for visiting anonymous websites. According to the study, repository and Tor crawling are considered as the most common methods employed while injecting relay, repositories and Tor crawling are found to be the most efficient ones. The study also highlighted the limits of onion collection and advises future research to give more representative datasets for dark web investigation, since most earlier efforts studied a limited section of the Tor networks.[5]

Anonymity networks, also known as darknets, are privacy-enhancing technologies aimed at avoiding censorship, preserving user privacy, and promoting freedom of speech. Tor, the most popular anonymization technology, caters to various users, including individuals, businesses, journalists, and activists. Users can safeguard online activity, keep data confidential, conduct competitive analysis, protect anonymous sources, and report abuses from dangerous areas. [6]

The Dark Web, a web-based platform for anonymous communication, has become a key route for selling dangerous information and illegal commodities owing to its high anonymity protection. This article discusses the present condition of Dark Web communication technologies, research methodologies, and obstacles, emphasising on the absence of thorough analysis and study in the domain, as well as the limits of existing website coverage and obsolete data. [7]

The deep web is the primary storage for large amounts of data, making it impossible to create perfect or complete crawlers for this data. To address this, a new crawler structure has been developed, The crawler is adapted to extracting data from forms, since most of deep web interfaces are form-based. We introduce some advanced components, such as mainframe extracting module, enhanced Bayesian classification algorithm, and AJAX form dealing. Moreover, we also introduce DOM Tree for helping a better and more visual analysis on downloaded web pages. [8]

A study on the Tor darknet, a widely studied system in computer security, found that Tor hidden services are organized in a bipartite core-periphery network. Most websites are part of a weakly connected core, with 10% belonging to the periphery. The study also found that 15% of onion domains appear on the Surface Web, and 20% include web pages embedded in the Surface Web via hidden Network. Despite Tor hidden services implementing best practices, web tracking is present in the Dark Web. [9]

III. PROPOSED METHODOLOGY

Methodology for DarkWeb Monitoring

Our methodology includes the process of setting up a VPN and Tor connection on the dark web, which anonymizes internet traffic which will hide our identity and server's identity from being exposed to the hacker on DarkWeb or exit nodes. We have randomized the user agent to maximize anonymity and connect to the DarkWeb. Our methodology includes best measure and techniques to stay anonymous while enumerating the DarkWeb. The methodology includes connecting to onion websites, which are not accessible by regular web browsers. In the next step, we use CURL to verify the collected URLs. After crawling, the collected information is verified and updated. The tool will automatically capture screenshots of verified .onion sites and save the final result, including links to relevant dark web pages. This helps protect users' identities while browsing the dark web.

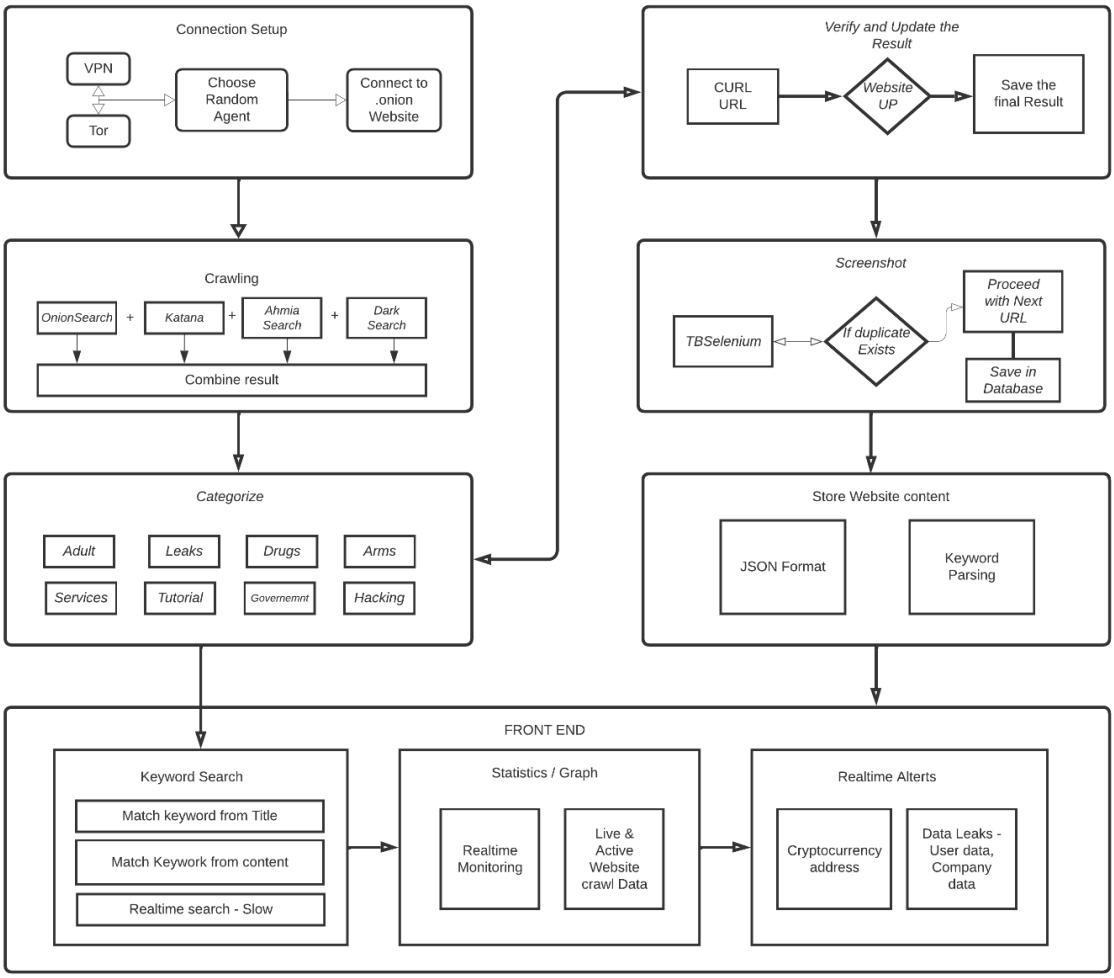

Fig 1: Flow Diagram of proposed method

3.1 Connection Setup

The first phase of our proposed methodology starts with the connection setup where we connect to the Virtual Private Network (VPN) and Tor connection which is necessary step towards maintaining anonymity and establish a tunnelled connection while mining and crawling the DarkWeb. This is essential for anonymizing internet traffic, especially when delving into the depths of the dark web. Utilizing VPN and Tor in always ensures that system's identities as well as user's identity are protected and their data encrypted throughout their browsing sessions.

Detailed Steps:

3.1.1 VPN Configuration: Our solution is integrated with VPN connection setup, where the user's connections are automatically routed through encrypted tunnels to VPN servers located in various countries, enhancing anonymity and thwarting potential surveillance.

3.1.2 Tor Integration: Tor is a costless set up that obscures your internet traffic. By using a network of volunteer relays, it directs your connection in a way that hinders the ability to trace your online activities and determine your location. This feature might be advantageous for safeguarding your confidentiality when navigating the internet. Our solution integrates the automatic connection with Tor, tailored and enhanced for our surveillance objectives. Users may easily initiate the Tor browser module, which will anonymize their traffic by routing it over the Tor network.

3.1.3 Choose Random Agent: This feature lets you pick a temporary, anonymous identity for browsing. It includes a fake IP address (like your online address) and disguises your user-agent, device's signature every time you crawl the URLs on the DarkWeb. This makes it much harder for someone to follow your tracks and identify you, giving you an extra layer of privacy.

3.1.4 Connect to Onion Website: Our platform's interaction with the Tor network allows for easy access to dark web material. Users may easily access ". Onion'' websites without the requirement for manual setup or specialised web browsers.

3.2 Crawling

Our methodology employs advanced web crawling techniques to systematically index dark web content. Using custom-built crawling algorithm using already available module in python called OnionSearch and we cover major activities while collecting various types of illegal activities. OnionSearch is a Python script that scrapes URLs on various ".onion" search engines, allowing users to scan websites for information gathering and vulnerabilities. It can target any domain and offers features like command completion and contextual help. To use OnionSearch, Python must be installed on Kali Linux. The tool supports various search engines, including Ahmia, Darksearchio, Onionland, Notevil, Darksearchenginer, Phobos, Onionsearchserver, Tordex, Tor66, Tormax, Haystack, Multivac, Evosearch, and Deeplink. The interactive console provides command completion and contextual help. To use OnionSearch, python must be installed on Kali Linux. The tool is available on GitHub and can be used on various search engines.

Advantages:
Webdriver capabilities are essential for web applications, controlling properties like user agent or platform, and enabling features. These capabilities are browser-specific and could affect the privacy of the browser. Currently, tbselenium has the same capabilities as Firefox, with some defaults enabled. These include handlesAlerts, database enabled, JavaScript enabled, and browser connection enabled. These capabilities are crucial for web applications to function properly and maintain user privacy. However, the exact implications of these capabilities remain unclear.

3.2.1 Technical Implementation

The crawling process is powered by proprietary data collection scripts, designed to navigate the complex structure of the dark web efficiently. These scripts leverage CURL commands and other data transfer protocols to retrieve and index webpages while minimizing detection risk.
The bash script is designed for performing searches on the Tor network and run the onion search module using automation and every time, predefined keyword is passed and related URLs are collected.
Functions in our tool:
perform_onion_search(): This function takes a keyword as an argument, runs a search on the Tor network using the onionsearch command, and saves the results to a temporary file.
compare_and_update_files(): The above function handles responsibility for comparing the freshly retrieved results with a previous file containing URLs linked to the keyword. If there are any new URLs detected, they are added to the original file.
Main Loop: The script begins an indefinite loop (while true) and repeats over an array of keywords preset such as drugs, guns, government, leaks, government leaks and services etc. For each term, it performs perform_onion_search() to search for onion URLs relating to that keyword and then calls compare_and_update_files() to modify the main file with any fresh URLs.
Real-Time Updates: Our software continually analyses dark web sources for new information, giving users with real-time updates and insights into changing patterns and risks.

3.2.2 Screenshot

Our tool automatically takes screenshots of relevant webpages encountered during monitoring directly within the platform interface. This feature facilitates documentation and analysis, enabling users to preserve critical evidence or insights of the collected URLs during crawling process.

Screenshot Integration: Screenshots are seamlessly integrated into the platform's category function. And it will help to keep users informed about the activities happening on the dark web without manually visiting them.
The user can search through the different categories of web URLs. The user can enter a search term, such as an email address or a company name, and the system will search the dark web for mentions of that term. If the system finds any matches, it will return them to the user.

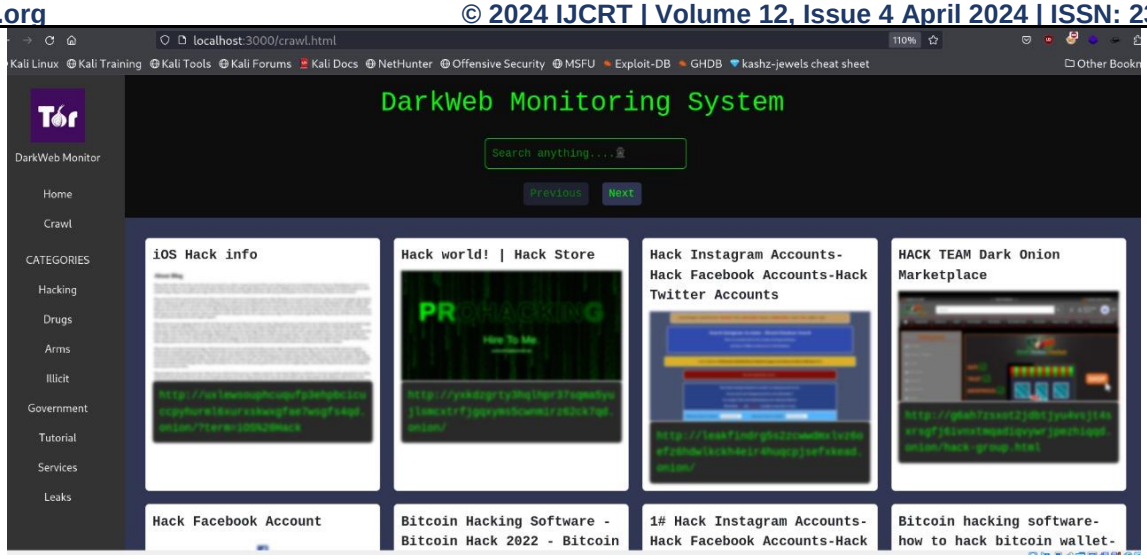

Fig 2: Screenshot Implementation.

Save the Final Result: The system displays a dynamic live count of indexed links on the dark web, allowing users to filter search results by categories like hacking, drugs, arms, illicit content, government, leaks, tutorials, and services. It also features a search bar for specific keywords within the indexed dark web content. To conclude our process, users have the ability to preserve the final results of their monitoring efforts immediately inside the platform. This involves the preservation of hyperlinks to relevant dark web sites, together with their corresponding information and analysis.

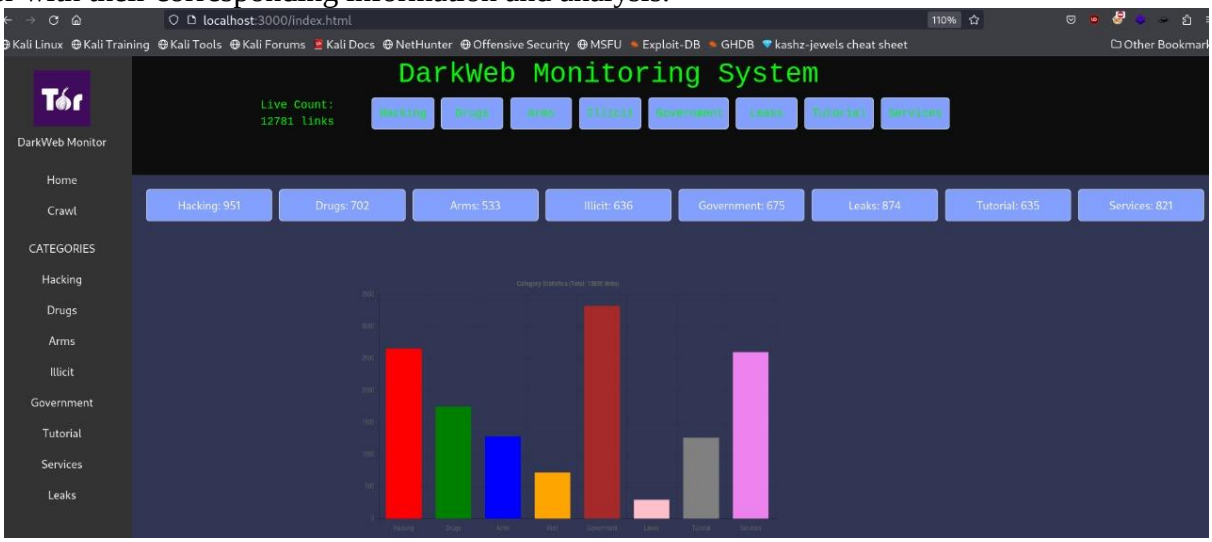

Fig 3: Charts are displayed on the index page for different categories of collected URLs.

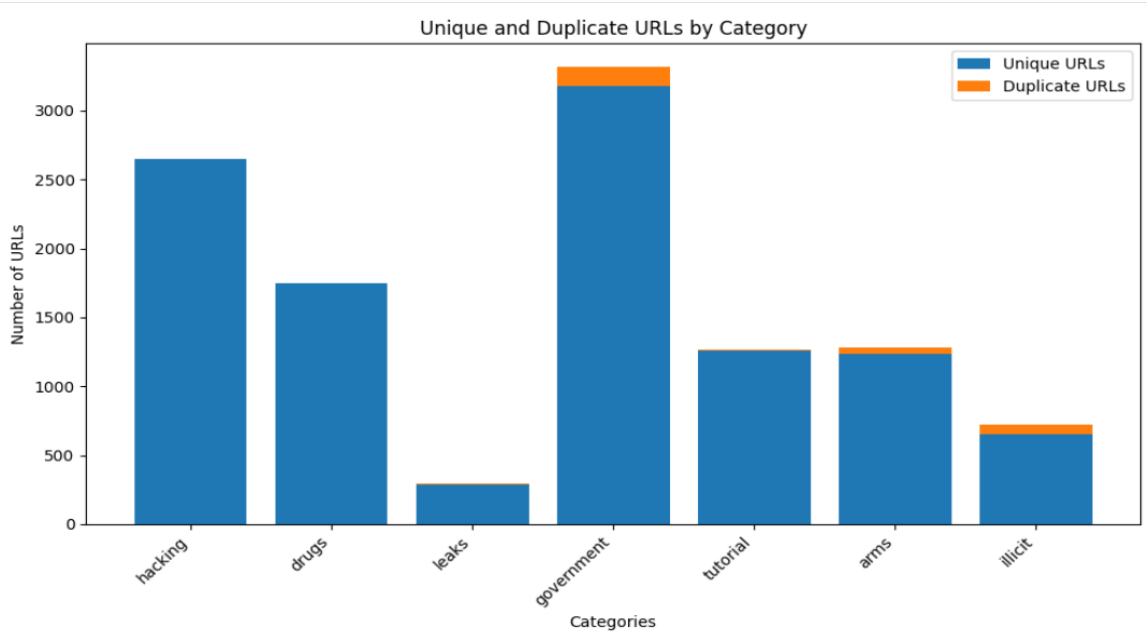

Fig 4: Graphical representation of Unique and Duplicate URLs crawled and collected.

The above graph shows that, in the government, arms and illicit categories, contains duplicate URLs which are very less in number and this shows the efficiency of our crawling system.

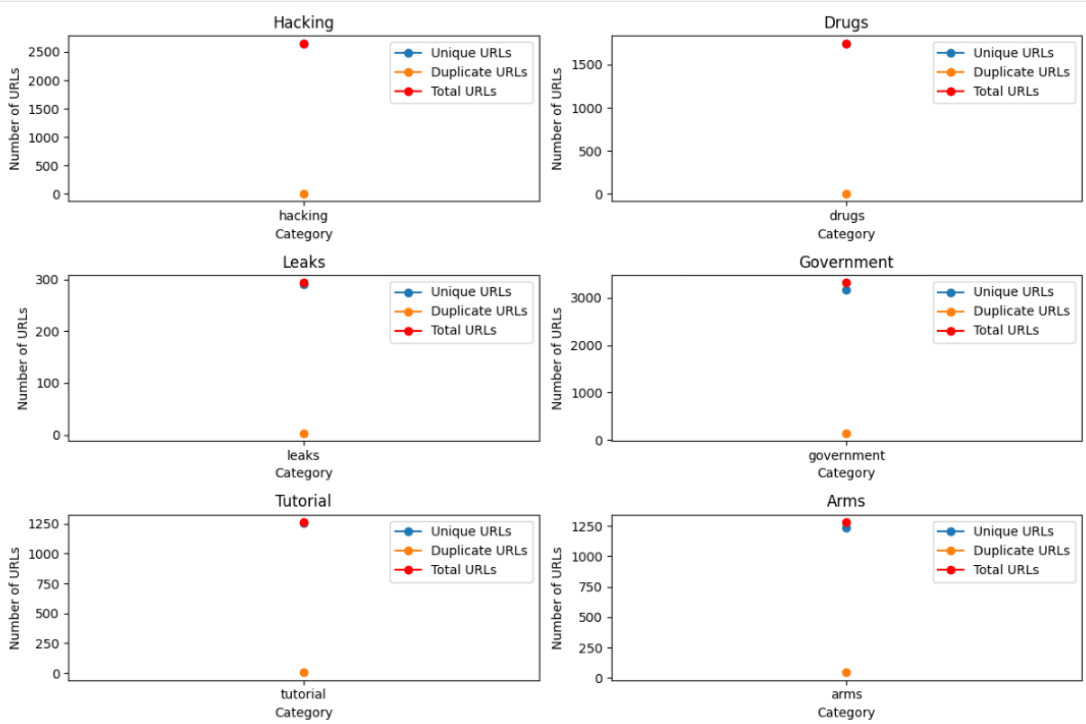

Fig 5: The above chart displays list of duplicate URLs, Unique URLs and Total URLs of each category.

Our DarkWeb monitoring approach includes a wide range of advanced tools and strategies that enable users to safely and effectively navigate the hidden online world. Through solid connection protocols, smart web crawling algorithms, and automated verification procedures, we allow users to find useful information while safeguarding their privacy and security.

IV. CONCLUSION AND FUTURE WORK

Dark web monitoring systems are not perfect and may not find all information available on the dark web. Additionally, using a dark web monitoring system can be risky due to the presence of malicious websites. Malware, which is software designed to harm a computer system, can infect a user's computer if visited. Phishing scams, which trick users into giving up personal information, are common on the dark web. Law enforcement agencies may also monitor the dark web for illegal activity, potentially flagging users using a dark web monitoring system. Therefore, it is crucial to weigh the risks and benefits of using a dark web monitoring system carefully and be aware of the security risks involved. This approach uses various data mining techniques on extracted data, with the design of the crawler and data mining operations varying depending on the website's nature and quality. Results may vary over time due to keyword differences. The system is designed to investigate dark Web sites anytime and anywhere. Future developments include integrating search feature which will reach for content of the website besides the title which is currently integrated. The future work also includes using of AI algorithm to classify and get more accurate results about the suspicious activities and drop all false positive onion addresses and add more keywords based on opensource data, Alert system in a Realtime for Data Leaks such as User data, Company data and analyse the extracted Crypto currency addresses. Expanding experiments to include keyword searches in multiple languages, and involving preprocessing techniques like corrections for misspelled words and mixed-language characters thus it will help us add more scope to the project. The system will also consider other methods of data mining, such as classification integrated with clustering and preprocessing techniques.

REFERENCES

[1] Kavallieros, D., Myttas, D., Kermitsis, E., Lissaris, E., Giataganas, G., Darra, E. (2021). Understanding the Dark Web. In: Akhgar, B., Gercke, M., Vrochidis, S., Gibson, H. (eds) Dark Web Investigation. Security Informatics and Law Enforcement. Springer, Cham. https://doi.org/10.1007/978-3-030-55343-2_1

[2] Beshiri, A. and Susuri, A. (2019) Dark Web and Its Impact in Online Anonymity and Privacy: A Critical Analysis and Review. Journal of Computer and Communications, 7, 30-43. doi: 10.4236/jcc.2019.73004.

[3] Randa Basheer, Bassel Alkhatib, "Threats from the Dark: A Review over Dark Web Investigation Research for Cyber Threat Intelligence", Journal of Computer Networks and Communications, vol. 2021, Article ID 1302999, 21 pages, 2021. https://doi.org/10.1155/2021/1302999
[4] Shalini Ghosh, Ariyam Das, Phil Porras, Vinod Yegneswaran, and Ashish Gehani. 2017. Automated Categorization of Onion Sites for Analyzing the Darkweb Ecosystem. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery, New York, NY, USA, 1793–1802. https://doi.org/10.1145/3097983.3098193
[5] Javier Pastor-Galindo, Félix Gómez Mármol, Gregorio Martínez Pérez, On the gathering of Tor onion addresses, Future Generation Computer Systems, Volume 145, 2023, Pages 12-26, ISSN 0167-739X, https://doi.org/10.1016/j.future.2023.02.024.
[6] M. Simioni, "Investigative Techniques for the De-Anonymization of Hidden Services," in IEEE Security & Privacy, vol. 19, no. 2, pp. 60-64, March-April 2021, doi: 10.1109/MSEC.2021.3050245.
keywords: {Privacy;Writing;Censorship;Business},
[7] H. Zhang and F. Zou, "A Survey of the Dark Web and Dark Market Research," 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 2020, pp. 1694-1705, doi: 10.1109/ICCC51575.2020.9345271. keywords: {Communication systems;Conferences;Market research;Encryption;Data mining;component;tor;hidden service;dark web;attack and defense of anonymous;dark web crawler;dark web data mining;understanding dark jargons},
[8] W. Ma, X. Chen and W. Shang, "Advanced Deep Web Crawler Based on Dom," 2012 Fifth International Joint Conference on Computational Sciences and Optimization, Harbin, China, 2012, pp. 605-609, doi: 10.1109/CSO.2012.138. keywords: {Feature extraction;Crawlers;Data mining;Bayesian methods;XML;HTML;Web pages;AJAX;Dom Tree;Deep Web;Form},
[9] Iskander Sanchez-Rola, Davide Balzarotti, and Igor Santos. 2017. The Onions Have Eyes: A Comprehensive Structure and Privacy Analysis of Tor Hidden Services. In Proceedings of the 26th International Conference on World Wide Web (WWW '17). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1251–1260. https://doi.org/10.1145/3038912.3052657