Design and Implementation of Web Extraction-based Cyber Threat Intelligence System on Reddit

Fathan Ananta Nur
School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia
E-mail: 18219008@mahasiswa.itb.ac.id

***Abstract*— Cyber Threat Intelligence (CTI) has become a social problem as cyber threats continue to increase. Organizations have invested significantly in cybersecurity defense, and CTI has emerged as a mechanism to combat these threats. CTI encompasses factual data that includes context, threat methods, indicators, and potential mitigation strategies. Hacker communities and forums can provide substantial proactive CTI. Among various platforms, forums offer comprehensive metadata, permanent data, and openly accessible tools, techniques, and procedures (TTPs). Reddit is one such platform that accommodates hacker communities and forums. A model or prototype is needed to simulate how information can be extracted from Reddit. Reddit allows access to essential data that can be converted into CTI. Professionals can make informed decisions and take future actions based on CTI analysis conclusions presented visually. Therefore, data collected from Reddit can be processed and visualized in context to assist organizations in preparing preventive measures against cyber threats.**

Keywords*— *CTI, Reddit, web scraping, data visualization

I. Introduction

The evolution of cybercrime and the increased connectivity of computers have led organizations to connect their infrastructure to the internet for improved efficiency. This has raised concerns about network vulnerabilities and cyber threats, prompting the development of Cyber Threat Intelligence (CTI). CTI involves collecting and analyzing data from internal systems to gain insights into emerging threats and protect assets. While reactive CTI requires specific events, proactive CTI focuses on preventive measures. Social media platforms like Reddit serve as valuable sources of information, and web data scraper systems can collect relevant content [1]. Implementing proactive CTI on Reddit can provide insights into threat actors and models.

This paper aims to address and provide answers to the problems related to proactive CTI. The objectives of this article are as follows: a) to build a web scraper model that can be applied to retrieve and process data enabling the compilation of proactive CTI; b) to explore valuable information and insights that can be obtained from Reddit using the developed web scraper model; c) to assess the performance of the web scraper model on Reddit while considering implementation challenges in continuously gathering information; d) to create interactive visualizations presented to academics and CTI practitioners for investigating the collected exploitation for proactive CTI.

II. Preliminaries

A. Fundamental Concept

1) Cyber Threat Intelligence (CTI): According to the SANS Institute, cyber threat intelligence (CTI) relates to computer, network, and information technology threats. CTI includes various attributes that serve as intelligence information [2]. These attributes can include descriptions of threat actors, maneuvers, motivations, and Indicators of Compromise (IoC) that can be shared with trusted stakeholders [3]. IoCs are easily actionable CTI attributes and are the focus of many CTI models. IoCs are typically applied in applications such as Intrusion Detection Systems (IDS), website blocking, blackholing, identifying compromised hosts and malware [4]. Big Data technology is utilized to store CTI attributes and establish connections between historical and new indicators [5]. CTI attributes tend to focus more on corporate IT and may overlook emerging fields like the Internet of Things (IoT), Industrial Internet of Things (IIoT), and the automotive sector. However, these technologies, also known as embedded systems, are connected to the back-end and can benefit from and analyze CTI attributes aimed at corporate IT.

2) Data Extraction: Web scraping, also known as web data extraction, involves extracting a large amount of data from the World Wide Web, a global information space accessible through the Internet. In the field of journalism, web scraping has been utilized to assess the significance of international news by counting how often it is mentioned on social media [6]. The research on information extraction, which explains how unstructured or semi-structured content can be processed to achieve specific information goals, has enabled this type of processing. In the early years, web scraping was manually performed through human copy-pasting. However, due to the dynamic nature and technological advancements of the web, traditional methods like manual copy-pasting became ineffective. Automation became essential in the web scraping process. By leveraging the hierarchical and structured nature of Hypertext Markup Language (HTML), the programming language used to display modern web pages, known as semantic web, automated data extraction from the web became possible [7].

3) Reddit Internet Forum: Internet forums are online message boards created for members to communicate with each other by discussing various topics [1]. While there are many framework structures for forums, most of them have a similar tree structure. A forum can include multiple subforums, each focusing on different topics. Users can create posts or discussions relevant to specific subtopics within a subforum. By posting, users engage in conversations within a thread. Each post includes at least the username, post date, and text. Forums automatically archive all posts, making them accessible unless deleted by forum moderators or the original poster. One of the most well-known internet forums is Reddit. Figure 1 illustrates the content structure within Reddit. Registered users can submit posts or submissions containing titles, external links, or self-written content, which are available for all Reddit users to vote and comment on [8].

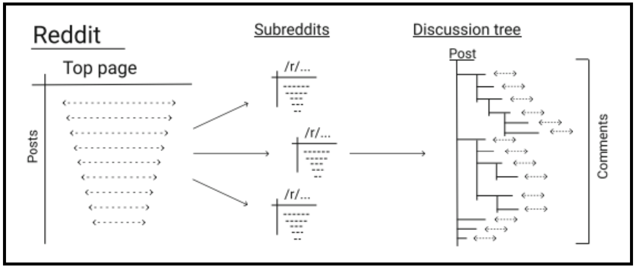

Fig. 1. Content Structure Schematic Diagram on Reddit [8]

B. Previous Works

Previous works in this topic mainly talk about frameworks used in gathering data from various source for certain use.

Samtani et al, 2017 [9] develops a new framework for CTI (Cyber Threat Intelligence) by utilizing automated mining approaches for web, data, and text, with the aim of collecting and analyzing a large amount of source code from hacker communities, tutorials, and attachments found in the international underground hacker community. The framework enables researchers to identify numerous malicious assets freely available in underground hacker forums, such as crypters, keyloggers, SQL injections, and password crackers, some of which may be the root causes of recent breaches against organizations can also determine the key individuals behind these assets by using social network analysis techniques and metrics. This approach can be generalized to any hacker forum, regardless of subforum structures. The research has practical implications for organizations seeking to enhance their cybersecurity posture. Assuming an organization is aware of the system they want to protect, they can apply this framework to their chosen forums to identify relevant hacker assets for their system.

Proferes et al, 2021 [10] explain how Reddit is utilized by researchers as a data source. Firstly, the study highlights the increase in the number of studies using Reddit data over the past decade. Most processes involve generating research articles using computational techniques such as web scraping. The topics studied using Reddit data are highly diverse, and at times, researchers gather data from Reddit communities that may encompass vulnerable populations. The findings indicate that only a few researchers share the knowledge they generate on Reddit, but nearly 30% of the research articles in this dataset related to Reddit appear on Reddit itself. This demonstrates a widespread interest in Reddit for research about Reddit. However, further exploration is needed to better understand the value created by such engagement and knowledge sharing.

III. Proposed Method

A. Framework Overview

The requirements for a CTI system can be explained using the CTI 8-step model depicted in Figure 2 [11].

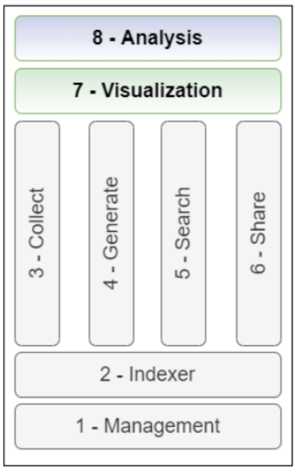

Fig. 2. CTI Model [11]

Step 1 - Management is responsible for managing user interactions with each functionality offered by the application and the interactions among them. This step requires controlling data flow and access through permission management. Step 2 - Indexer is supported by a storage structure capable of handling the input and output data volume in the CTI system. Step 3 - Collect is responsible for providing external data collection and insertion into the CTI system. Step 4 - Generate can provide internal data normalization using the same patterns from Steps 2 and 3, allowing absorbed data to be transformed into feeds for future use in the CTI system. Step 5 - Search can provide mechanisms and methods for effective data manipulation and exploration, enabling fast data indexing and full visibility of stored data. Step 6 - Share allows for the sharing of internal data such as feeds, collections, and threat indicators among users, as well as sharing with third-party tools. Step 7 - Visualization offers CTI data visualization in a temporal format to create threat indicators, allowing individuals to have a comprehensive overview of threat traces, IoCs, and any enriched and shared useful information from other stakeholders. Step 8 - Analysis, which is an intrinsic part of Step 7, can implement functionality that enables data analysis and manipulation to extract the best information from available threat data.

B. System Requirement

System requirement analysis is necessary to understand the issues by examining the initial overview of the system and its capabilities. Requirement analysis can encompass both functional and non-functional aspects. Table 1 and Table 2 provide an analysis of functional and non-functional requirements.

TABLE I. FUNCTIONAL REQUIREMENT

ID	Requirement	Description
FR-1	The system has the capability to send API requests to Reddit.	The system utilizes the GET method to send API requests to Reddit in order to retrieve data related to the content within a specific subreddit based on predetermined links.
FR-2	The system has the capability to filter the required data.	The system can filter the obtained data from the Reddit API response based on desired parameters to facilitate processing.
FR-3	The system can provide visualizations of the obtained data.	Visualization is performed to present data in an easily understandable form, facilitating analysis. Therefore, the system can display the transformed data in graphical form.
FR-4	The system can provide indicators to denote the threat level of the analyzed subreddit.	The system can assign a flag to content obtained through Reddit web scraping. This flag serves as a means for the CTI to provide significance to relevant parties and draw special attention to the flagged content.

TABLE II. NON-FUNCTIONAL REQUIREMENT

ID	Requirement	Description
NFR-1	The system aims to ensure that 95% of the response time for API calls to Reddit is within 500 ms.	The system ensures that the elapsed time between sending an API GET request to Reddit and receiving the data does not exceed 500 ms in 95% of all instances.
NFR-2	The system is capable of storing a minimum of 1000 crawled data per cycle.	The system ensures that a minimum of 1000 data is stored per cycle of web crawling. This limitation is in accordance with Reddit's rules, which allow a maximum of 1000 content items to be scraped.
NFR-3	All levels of system access can be used by users.	The system can be fully utilized by users. Currently, access levels have not been specialized as the system has a single main function, so every usage has the same purpose.

C. The Model

The overall system architecture for web scraping on Reddit can be represented using an architectural diagram, illustrating the different components involved in the process, from fetching data through API requests to processing, storing, analyzing, and visualizing the scraped data. The system aims to provide a comprehensive solution for extracting and utilizing data from Reddit, allowing users to effectively gather, process, and present the scraped information.

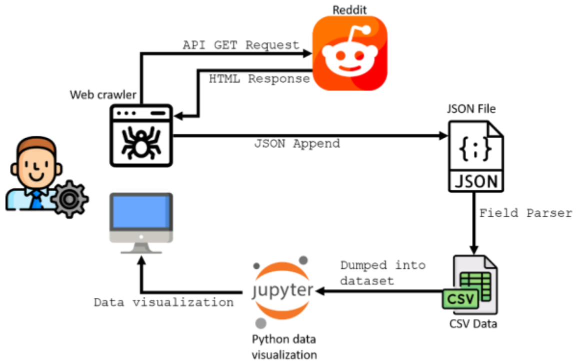

Fig. 3. Architecture Diagram of CTI System Based on Web Scraping on Reddit

Figure 3 illustrates the architecture of the CTI system based on a web scraper for Reddit. The system utilizes a web crawler that actively calls Reddit's API endpoint. The web crawler is built using Python and implements a specific library called PRAW (Python Reddit API Wrapper) to interact with Reddit. The web crawler program/application sends a GET request to Reddit, which responds by sending content related to the specified subreddit link. The received content from Reddit includes comprehensive metadata for each type. However, for CTI visualization purposes, only essential data such as comment content, comment timestamps, and usernames are required. Therefore, a subset of metadata is selected and dumped to obtain the dataset. This dataset can be visualized using Python data visualization techniques and presented to relevant stakeholders.

IV. Implementation, Testing, and Evaluation

The objective of this chapter is to demonstrate the extent to which the proposed solution outlined in the previous section can address the main problems identified in the background. The methodology employed involves conducting tests based on constructed scenarios to validate the requirements outlined in the preceding chapter.

A. Environmental Setup

The system will be implemented on the Google Cloud Platform, covering storage and operating system aspects. Python programming language will be utilized as it offers the PRAW library for retrieving data from the Reddit API. Additionally, Python notebooks will be used for data processing and visualization purposes. Figure 4 below shows how system deployed on the environment.

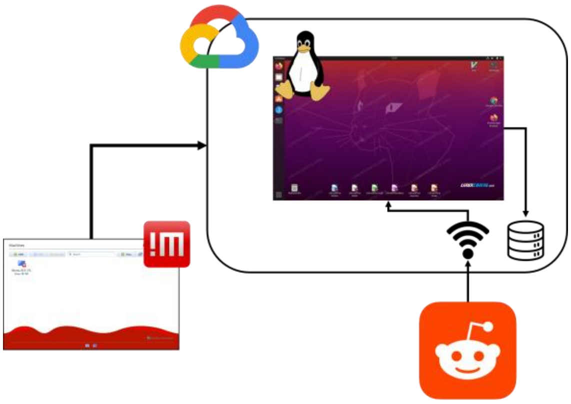

Fig. 4. System Environment Setup

B. System Testing

1) Testing of Processing Time Based on the Network: The testing compares the processing time for a specific scenario using different network conditions to extract data from the Reddit API. The tests are conducted with the same command to ensure consistent data. The extraction is done in batches and laps, with each batch representing a limit for retrieving submissions and each lap representing the number of batch calls. The aim is to have the same quantity of extracted items by using the same batch and lap values.

TABLE III. Processing Time Based on Network

Internet Provider	Proxy/VPN	Download speed (Mb/s) and latency (ms)	Average time (minute)	Total time (minute)
Biznet	-	82, 11	Failed	Failed
	ITB VPN	57, 40	0,46	1,4
	Cloudflare 1.1.1.1	88, 256	0,86	2,6
Firstmedia	ITB VPN	28, 23	0,16	0,5
	Cloudflare 1.1.1.1	94, 266	0,13	0,4
Eduroam	-	15, 19	0,26	0,8
Cloud network	-	1100, 8	0,03	0,1

Based on the testing scenarios, it was found that the fastest process for obtaining certain rows of data is achieved using a cloud network. It can be concluded that the cloud network has better download speed and ping values compared to other network testing scenarios. Therefore, the use of a cloud network will be implemented in the subsequent data extraction process testing.

2) Testing Item Quantity Based on Processing Time: The purpose of this testing is to compare the processing time required to complete a specific scenario under different conditions. The testing is conducted using a cloud network, as explained before. The tested conditions involve the quantity of data requested from the Reddit API. Different commands are used for each request, and the testing includes variations in the number of laps and batch scrapings.

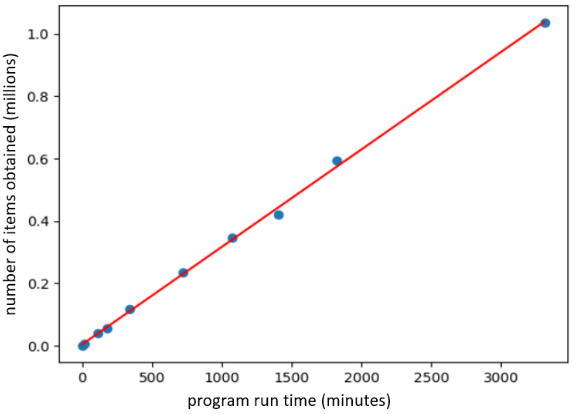

Fig. 5. Item Quantity Based on Processing Data Graph

Figure 5 shows a plot of the total number of items obtained from Reddit based on the time (in minutes) spent. The obtained regression equation values are provided.

$$f(x) = 312{,}8x$$

3) Data Size Testing Based on the Number of Retrieved Items: The testing aims to compare the data size in storage under different scenario conditions. The testing is conducted using the same cloud network but with varying numbers of laps, resulting in different numbers of retrieved items.

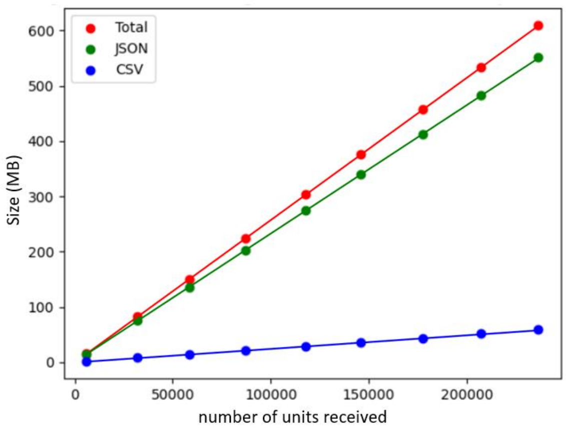

Fig. 6. Data Size Based on Number of Retreived Item Graph

Based on Figure 6, the graph illustrates the storage size required to store the extracted data based on the number of units received. The experiment results were used to determine the prediction of storage size for the extracted data, which includes the combined size of CSV and JSON data. The regression equation for the number of units received versus the total size was also obtained.

$$g(x) = 0{,}0026x$$

Based on the regression equations f(x) and g(x) from sections before, a new regression equation h(x) can be derived to establish the relationship between the runtime of the extraction process (in minutes) and the required storage size (in MB). The linear regression equation g(x) can be determined as follows.

$$h(x) = \ 0{,}81328x$$

For example, when performing the extraction process for a full week (168 hours), the minimum required storage size during the process is 8197.86 MB (megabytes).

4) Testing JSON to CSV and Database Conversion: The testing aims to determine the optimal data format for CTI processes, including data retrieval and analysis. Two common formats, CSV (Comma Separated Value) and database (SQLite), are chosen for the testing. These formats are widely used for representing data. The testing will compare the conversion from the JSON format (the default format of the Reddit API) to CSV and database formats.

TABLE IV. JSON CONVERTION TO CSV AND DATABASE

Number of rows	CSV time (second)	Database time (second)
3	0.00400090217590332	0.23458552360534668
5	0.003999948501586914	0.23686671257019043
10	0.00500178337097168	0.22804474830627441
25	0.013000011444091797	0.23559308052062988
50	0.013001680374145508	0.27994370460510254
80	0.01099967956542968	0.2850363254547119
200	0.015003204345703125	0.2646205425262451

Based on the conducted tests, it can be concluded that the average conversion time from JSON to CSV is shorter compared to the conversion from JSON to a database format. This reduces the time required for selecting fields and parsing data from the raw crawled data. Although SQL is a more efficient programming language for querying data, CSV format provides more capabilities for analyzing a database table. The Pandas library is an ideal choice for users who need to perform simple commands and analyze structured data.

C. *Implementation on Data Vizualisation*

From the extracted data, various visualization models can provide valuable information and insights for cyber threat intelligence analysis. Below are some examples of visualizations based on the extracted data from the tested subjects, which were also used in previous chapters.

1) Most Appearing Words: The image below, Figure 7, shows the bar plot results for the most frequently occurring words in the dataframe. The most common words are limited to comments (body) with a score greater than 250.

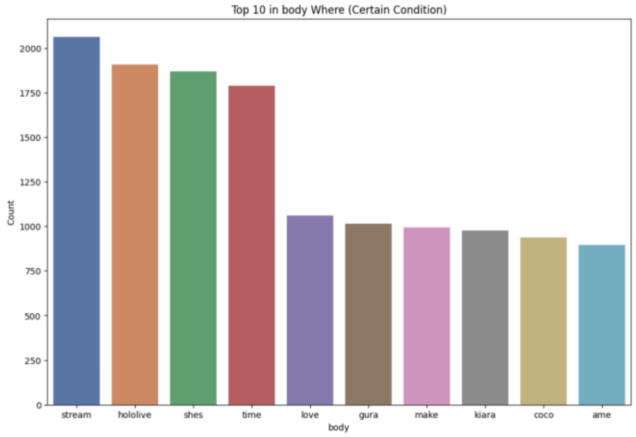

Fig. 7. Most Appearing Word Graph

2) Most Commented Author: The image below, Figure 8, shows the bar plot results for the authors who have made the most comments in the dataframe. The authors with the highest comment count are limited to those with a score greater than 250. There is a special case with the author name “nan”. The name “nan” indicates an empty value or an author who has not shared their username publicly. Therefore, in further analysis, the name “nan” can be disregarded or considered as an additional factor for anonymous user cases.

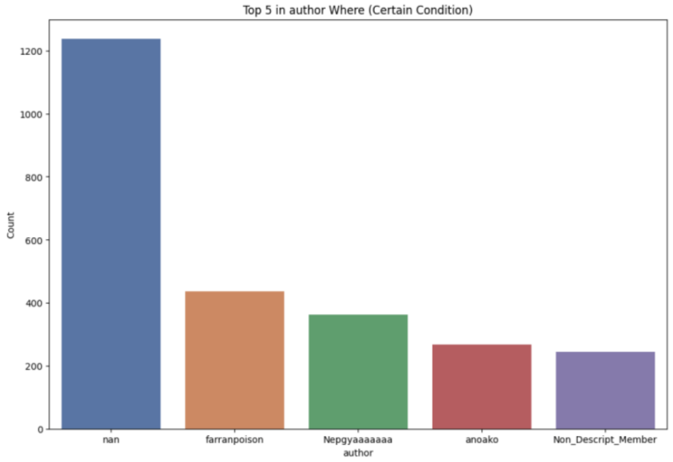

Fig. 8. Most Commented Author Graph

3) Number of Appearances of Certain Words Per Month: The image below, Figure 9, displays the bar plot showing the accumulated count of specific words appearing in comments per month.

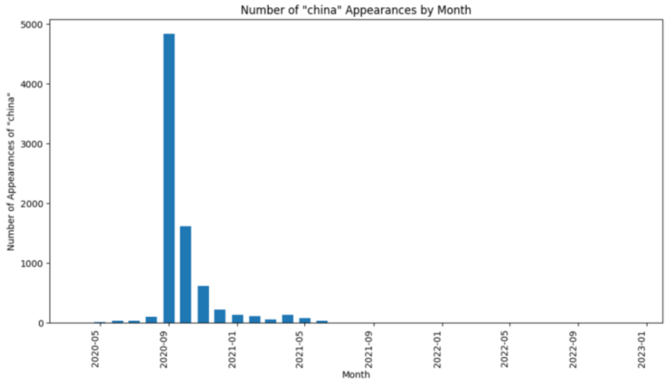

Fig. 9. Apperances of Certain Words Graph

V. Conclusion

The web scraper model can gather and process data for proactive Cyber Threat Intelligence (CTI) by extracting information from specific subreddits on Reddit. It provides benefits in understanding cybersecurity trends and patterns by analyzing structured datasets of submissions and comments. These datasets offer valuable insights into topics, user activity, and sentiments. The web scraper is efficient and effective in fulfilling the requirements outlined in System Requirement. Visualizing the data further enhances its understanding, although real-time and broader data sources would provide more accurate and up-to-date information. Overall, the web scraper model is a valuable tool for obtaining quality CTI information and improving readiness against cybersecurity threats.

Processing data in Cyber Threat Intelligence (CTI) is crucial for accurate and reliable security trend prediction. By applying machine learning techniques during data processing, organizations can generate more precise and timely security trend predictions for proactive threat mitigation. Machine learning algorithms can extract patterns and trends from processed data, providing more accurate and valid results. Integrating visualization in the data processing stage enhances understanding and decision-making regarding emerging security trends. Effective visualizations assist security analysts in extracting important information and making prompt and informed decisions. Therefore, implementing machine learning and visualization in CTI plays a vital role in generating effective and efficient security predictions for organizations. Additionally, there is a significant knowledge gap between the industry and academia when it comes to providing cybersecurity solutions through CTI in organizations. To address this gap, the creation of prototype systems that can be used by practitioners to test cybersecurity strategies before implementation is recommended. Further research on the cybersecurity needs of organizations is necessary to ensure focused and effective research and development efforts. Bridging the knowledge gap between industry and academia is expected to result in CTI solutions that better reflect real-time conditions and effectively address the increasingly complex and dynamic cybersecurity threats.

Acknowledgment

I would like to express my sincere gratitude to all individuals and organizations who have contributed to the successful completion of this research. First and foremost, I would like to thank my supervisor, Mr. Budi Rahardjo, for his guidance, expertise, and continuous support throughout the project. His valuable insights and suggestions have greatly contributed to the quality of this work.

I would like to extend my appreciation to Bandung Institute of Technology for providing the necessary resources and facilities that have enabled me to conduct my experiments and gather the required data. I would also like to thank the participants of this study for their cooperation and willingness to share their insights.

Furthermore, I acknowledge the contributions of previous researchers in the field whose work has paved the way for my study. Their groundbreaking research and publications have served as a valuable foundation for my investigations.

Lastly, I would like to express my gratitude to my family and friends for their unwavering support, encouragement, and understanding throughout this research journey. Their patience and belief in me have been instrumental in my success.

I acknowledge that this research would not have been possible without the collective effort and support from all these individuals and entities. Thank you all for your invaluable contributions.

References

[1] Williams, R., Samtani, S., Patton, M., and Chen, H., “Incremental Hacker Forum Exploit Collection and Classification for Proactive Cyber Threat Intelligence: An Exploratory Study,” 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 94-99, 2018, doi: 10.1109/ISI.2018.8587336

[2] Feng, B, “Threat intelligence sharing: What kind of intelligence to share?” August 2021, from https://www.concordia-h2020.eu/blog-post/threat-intelligence-sharing/

[3] Wagner, T. D., Mahbub, K., Palomar, E., and Abdallah, A. E., “Cyber threat intelligence sharing: Survey and research directions,” Computers & Security, 87, 2019, https://doi.org/10.1016/j.cose.2019.101589

[4] Ciobanu, C., Dandurand, L., Grobauer, M., Kacha, B., Kaplan, P., Kompanek, A., and Van Horenbeeck, M., “Actionable Information for Security Incident Response,” ENISA, Heraklion, Greece, 2014.

[5] Chismon, D., and Ruks, M., “Threat intelligence: Collecting, analysing, evaluating,” MWR InfoSecurity Ltd, 3(2), 36-42, 2015.

[6] Marres, N., and Weltevrede, E., “Scraping the social? Journal of Cultural Economy,” 6(3), 313–335, 2013, https://doi.org/10.1080/17530350.2013.772070

[7] Parvez, M. S., Tasneem, K. S., Rajendra, S. S., and Bodke, K. R., “Analysis of different web data extraction techniques,” 2018 International Conference on Smart City and Emerging Technology (ICSCET), 2018, https://doi.org/10.1109/icscet.2018.8537333

[8] Medvedev, A. N., Lambiotte, R., and Delvenne, J.-C., “The Anatomy of Reddit: An Overview of Academic Research,” Springer Proceedings in Complexity, 183–204, 2019, doi:10.1007/978-3-030-14683-2_9

[9] Samtani, S., Chinn, R., Chen, H., and Nunamaker, J. F., “Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence,” Journal of Management Information Systems, 34(4), 1023–1053, 2017, doi:10.1080/07421222.2017.1394049

[10] Proferes, N., Jones, N., Gilbert, S., Fiesler, C., and Zimmer, M., “Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics,” Social Media + Society, 7(2), 2021, https://doi.org/10.1177/20563051211019004

[11] Amaro, L. J. B., Azevedo, B. W. P., de Mendonca, F. L. L., Giozza, W. F., Albuquerque, R. de, and Villalba, L. J. G., “Methodological framework to collect, process, analyze and visualize cyber threat intelligence data.” Applied Sciences, 12(3), 1205, 2022, https://doi.org/10.3390/app12031205.