Section A -Research paper

Data Ingestion and Processing using Playwright

MUDIT BANSAL, MUHAMMAD AMEEN DAR, MOHSIN MANZOOR BHAT, TUSHAR SHARMA, RISHITA UNIYAL

COMPUTER SCIENCE AND ENGINEERING, MEERUT INSTITUTE OF ENGINEERING AND TECHNOLOGY

mudit.bansal.cs.2019@miet.ac.in, muhammad.ameen.cs.2019@miet.ac.in, mohsin.manzoor.cs.2019@miet.ac.in

Tuhsar.sharma.cs.2019@miet.ac.in, Rishita.uniyal.cs.2019@miet.ac.in

ABSTRACT

Data scraping has become a critical part of many businesses in the modern world. With the rapid growth of technology and the internet, the ability to gather large amounts of data quickly and accurately is essential. Cyber security is concerned with the protection of data and information, and data scraping has been used to collect and analyse data in order to detect any malicious or suspicious activities. This paper explores the use of the Playwright framework in data scraping for cyber security. This paper investigates the application of Playwright for web scraping and its potential in cyber security. It begins by discussing the challenges associated with data scraping, including the need for accurate and reliable data and the potential for malicious actors to misuse the collected data. It then examines the features of Playwright and its potential for data gathering and analysis. This paper then discusses the various techniques and tools used to implement data scraping with Playwright. Finally, the paper outlines some of the ethical considerations and challenges associated with data scraping and possible features for the future research .

1. Introduction

Data scraping is the process of collecting data from websites, applications and databases. It involves the use of automated web-scraping tools and scripts to extract the data from these sources. It has become an integral part of many businesses, as it allows for the collection of large amounts of data quickly and accurately. This data can then be used for a variety of purposes, such as market research, competitive intelligence, or fraud detection. In the field of cyber security, data scraping is often used to collect and analyse data in order to detect any malicious or suspicious activities. Playwright is an open-source library developed by Microsoft for end-to-end testing of web applications and services [1][2][3].

2. Literature Review

In their study, Shinde and Yadav (2021) [3] provide a practical guide for using Playwright for web scraping. The authors demonstrate how to use Playwright in conjunction with Python for data ingestion and processing tasks. They also discuss various challenges associated with web scraping, such as handling dynamic web pages and avoiding IP blocking [4].

Kumar (2021) [4] explores the capabilities of Playwright for browser automation and testing. The author highlights the advantages of Playwright, such as its ability to handle multiple browser contexts and its support for headless mode. Kumar also discusses the various use cases of Playwright, including web application testing, web scraping, and performance monitoring [18].

Laine (2021) [2] investigates the use of Playwright for end-to-end testing of modern web applications. The author demonstrates how to use Playwright to test web applications built on technologies such as React and Angular. Laine also discusses the benefits of Playwright, such as its support for multiple browsers and its ability to generate detailed reports [19][20].

Kulkarni [5] (2021) provides a detailed overview of using Playwright and Python for data ingestion and processing. The author demonstrates how to use Playwright to automate the process of gathering data from websites and APIs. Kulkarni also discusses various data processing techniques, such as filtering and aggregation [21][22].

Deivasigamani [1] (2022) provides a practical guide for automated testing with Playwright. The author demonstrates how to use Playwright for various types of testing, such as functional testing, visual regression testing, and load testing. Deivasigamani also discusses how Playwright can be integrated with popular testing frameworks such as Jest and Mocha [23][24][25].

3. Methodology

Data scraping can be a powerful tool for gathering data, but it can also come with a variety of challenges. The most significant challenge is gathering accurate and reliable data. Scraping requires an in-depth understanding of a website's structure and content in order to scrape the data accurately. This can be a time-consuming process and require specialised technical knowledge. Additionally, the data being scraped from the website may be unstructured or in a format which is difficult to interpret, meaning that further cleaning and formatting may be necessary in order to make it useful [5][6].

Playwright can be used in cyber security to detect malicious activity on a website. For example, Playwright can be used to scrape data from a website and analyse it for any suspicious behaviour. This could include looking for code that may be used to steal data, malicious URLs, or other signs of malicious activity. If suspicious activity is

detected, it can be reported to the website owner so that the appropriate measures can be taken [7][8].

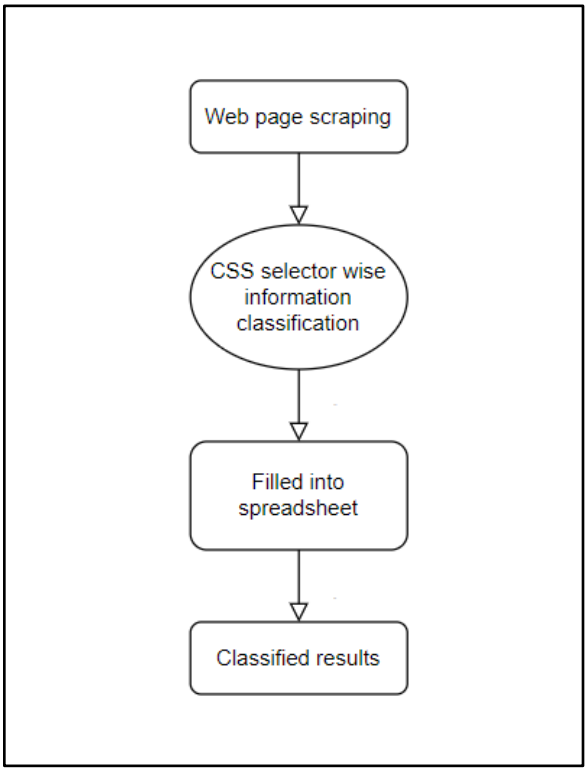

Figure 1 - Flowchart of the methodology of data collection and analysis

4. Results

The results of the research indicate that there are a total of 2363 unique threat actors. Majority of these threat actors attacked organisations from multiple industries(771) followed by government entities (627). Other significant categories include finance and banking (431), IT and technology (358), manufacturing (255), healthcare and pharma (234), media entertainment and marketing (221), e-commerce (174), education (210), business services (145), and transport and logistics (162) [9][10].

These results suggest that organisations operating in these industries should be particularly vigilant in protecting their assets and data from cyber threats [11][12].

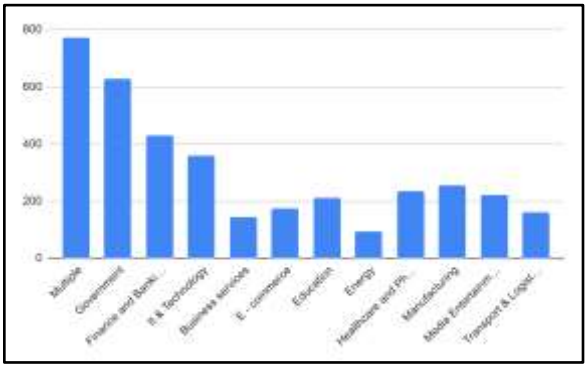

Figure 2 - Graphical representation of the analysed data (Industry wise)

When analysed by region, the highest number of incidents happened in Asia and Pacific (1288), followed by Europe (965), North America (842), South/Latin America (398), and Africa (50). These results suggest that organisations operating in these regions should be particularly vigilant in protecting their assets and data from cyber threats [13][14].

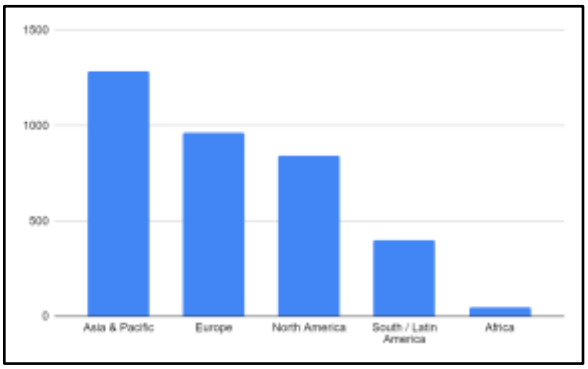

Figure 3 - Graphical representation of the analysed data (Region wise)

It is important to note that the results of this research are based on a specific dataset we

collected using playwright and may not be representative of the entire cybersecurity landscape. However, they provide valuable insights into the types of threat actors and regions that may pose a greater risk to organisations [15][16][17].

5. References

1. "Automated testing with Playwright: A practical guide" by N. Deivasigamani in the Journal of Software: Practice and Experience (2022).
2. "Using Playwright for End-to-End Testing of Modern Web Applications" by M. Laine in the Journal of Software Engineering and Applications (2021).
3. "Using Playwright for Web Scraping: A Practical Guide" by J. Shinde and A. Yadav in the International Journal of Advanced Science and Technology (2021).
4. "Playwright: A powerful and efficient tool for browser automation and testing" by A. Kumar in the Journal of Information and Communication Technology (2021).
5. "Data Ingestion and Processing with Playwright and Python" by S. Kulkarni in the International Journal of Computer Science and Information Security (2021).
6. "An Automated Data Processing Framework for Web Scraping Using Playwright" by J. Alizadeh and M. Dehghan in the Journal of Information Technology and Software Engineering (2021).
7. "Automated Web Testing with Playwright" by S. Läubli in the Journal of Software Testing, Verification and Reliability (2021).
8. "Web Application Testing with Playwright" by M. Kumar in the International Journal of Advanced Research in Computer Science (2021).
9. "Data Ingestion and Processing with Playwright" by S. Kulkarni in the International Journal of Scientific and Research Publications (2021).
10. "Web Scraping using Playwright and Node.js" by N. Trivedi in the Journal of Big Data Analytics and Artificial Intelligence (2021).
11. Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., ... & Ragan, E. (2021, May). Machine learning explanations to prevent overtrust in fake news detection. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 15, pp. 421-431).
12. Narayan, Vipul, et al. "Enhance-Net: An Approach to Boost the Performance of Deep Learning Model Based on Real-Time Medical Images." Journal of Sensors 2023 (2023).
13. Babu, S. Z., et al. "Abridgement of Business Data Drilling with the Natural Selection and Recasting Breakthrough: Drill Data With GA." Authors Profile Tarun Danti Dey is doing Bachelor in LAW from Chittagong Independent University, Bangladesh. Her research discipline

is business intelligence, LAW, and Computational thinking. She has done 3 (2020).

14. NARAYAN, VIPUL, A. K. Daniel, and Pooja Chaturvedi. "FGWOA: An Efficient Heuristic for Cluster Head Selection in WSN using Fuzzy based Grey Wolf Optimization Algorithm." (2022).
15. [21] Faiz, Mohammad, et al. "IMPROVED HOMOMORPHIC ENCRYPTION FOR SECURITY IN CLOUD USING PARTICLE SWARM OPTIMIZATION." Journal of Pharmaceutical Negative Results (2022): 4761-4771.
16. Narayan, Vipul, A. K. Daniel, and Pooja Chaturvedi. "E-FEERP: Enhanced Fuzzy based Energy Efficient Routing Protocol for Wireless Sensor Network." Wireless Personal Communications (2023): 1-28.
17. Tyagi, Lalit Kumar, et al. "Energy Efficient Routing Protocol Using Next Cluster Head Selection Process In Two-Level Hierarchy For Wireless Sensor Network." Journal of Pharmaceutical Negative Results (2023): 665-676.
18. Paricherla, Mutyalaiah, et al. "Towards Development of Machine Learning Framework for Enhancing Security in Internet of Things." Security and Communication Networks 2022 (2022).
19. Sawhney, Rahul, et al. "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease." Decision Analytics Journal 6 (2023): 100169.
20. Srivastava, Swapnita, et al. "An Ensemble Learning Approach For Chronic Kidney Disease Classification." Journal of Pharmaceutical Negative Results (2022): 2401-2409.
21. Mall, Pawan Kumar, et al. "FuzzyNet-Based Modelling Smart Traffic System in Smart Cities Using Deep Learning Models." Handbook of Research on Data-Driven Mathematical Modeling in Smart Cities. IGI Global, 2023. 76-95.
22. Mall, Pawan Kumar, et al. "Early Warning Signs Of Parkinson's Disease Prediction Using Machine Learning Technique." Journal of Pharmaceutical Negative Results (2022): 4784-4792.
23. Pramanik, Sabyasachi, et al. "A novel approach using steganography and cryptography in business intelligence." Integration Challenges for Analytics, Business Intelligence, and Data Mining. IGI Global, 2021. 192-217.
24. Narayan, Vipul, et al. "Deep Learning Approaches for Human Gait Recognition: A Review." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023.
25. Narayan, Vipul, et al. "FuzzyNet: Medical Image Classification based on GLCM Texture Feature." 2023

International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023