

Machine Learning and Big Data for Security Incident Response

Roberto Omar Andrade¹^a, María Cazares²^b, Iván Ortiz-Garces³ and Gustavo Navas²^c

¹*Facultad de Ingeniería en Sistemas, Escuela Politécnica Nacional, Quito, Ecuador*

²*IDEIAGEOCA, Universidad Politécnica Salesiana, Quito, Ecuador*

³*Facultad de Ingeniería y Ciencias Aplicadas, Universidad de las Américas, Quito, Ecuador*

Keywords: Cybersecurity, Machine Learning, Cognitive Process.

Abstract: Cybersecurity attacks have grown exponentially. At present, cyberattacks have different attack vectors and techniques, generating a high impact on social and commercial worldwide. On the other hand, cybersecurity analysts need to process large amounts of data to detect patterns to make possible proactive security defences strategies. Incident response processes are based on detection tasks developed by a security analyst in the first stages of incident response. This work analyses the cognitive functions performed by cybersecurity analysts in the detection phase and combines big data and machine learning to enhance the detection processes of cyberattacks.


1 INTRODUCTION


According to the World Economic Forum (WEF), cyberattacks have been considered in the top ten list of threats jointly with natural disasters and extreme poverty. So, cyberattacks are considered a source of high impact in economic, social, and environmental domains, which impules organizations to establish strategies, policies, and guidelines to protect against cyberattacks. One security strategy is defining an Incident Response Process (IRP) in the organization. The University of Carnegie Mellon proposed a formal IRP during the Morris worm attack in 1988, which impules the creation of the first CERT (Computer Emergency Response Team) or CSIRT (Computer Security Incident Response Team). Today, there are some proposals for the incident response process by international organizations such as NIST, ISO and SANS that include the following phases (Andrade et al., 2018a): Preparation, Detection and Analysis, Containment, Eradication & Recovery, Post-Incident Activity.


The first and second phases of the incident response process require building a baseline of normal and abnormal patterns by collecting and analysing logs from firewalls, servers, computers, and

network devices. Using machine learning and bigdata could allow the augmentation of human capabilities needed during incident detection. For instance, security analysts could analyse hundreds of DNS to detect malicious used for spoofing attacks. However, the volume of logs could be so extensive, and it could oversaturate the human capabilities of security analyst (Andrade and Torres, 2018). Conclusion This context impules several academic and industry proposals that have been considered to develop proactive solutions to improve the baseline of normal patterns and outliers.

This work shows an overview of machine learning and bigdata in the cyber incident response process. Cognitive tasks developed for security analysts during the incident response process are described in Section 2. An overview of machine learning (ML) and bigdata for detecting security events is present in Section 3. Scenarios of the use of ML and bigdata for anomaly detection are discussed in Section 4. Finally, in Section 5, we discuss the relevant facts of ML and bigdata in the cognitive process of the incident response process.

^a <https://orcid.org/0000-0002-7120-281X>

^b <https://orcid.org/0000-0003-3407-9442>

^c <https://orcid.org/0000-0002-2811-0282>

2 BACKGROUND AND MOTIVATION

The first step for incident response is establishing a cybersecurity situation awareness based on identifying threats, vulnerabilities, risk, the impact of an attack, and the behaviour of attackers and users.

2.1 Cybersecurity Situational Awareness

Cybersecurity awareness has three layers from a cognitive security perspective model (Andrade et al., 2018b): a) Perception, generated by collecting the information of the elements such as the firewall router, switches, and servers. b) Comprehension determines the status of the situation based on the analysis of patterns. Finally, c) Projection establishes a prediction of the types of threats or attacks.

2.2 Cognitive Security Tasks

To establish cybersecurity awareness, security analyst development activities to detect normal or abnormal behaviours on the network. According to NIST Cybersecurity Framework, the detection activity includes the following categories (Andrade et al., 2018b):

- Detect anomalies and events and their potential impact.
- Implement continuous monitoring capabilities and verify proactive measures.
- Maintain detection processes to provide awareness of anomalous events.
- The baseline for network operations and data flow expected by users and systems.
- Detected events are analysed to understand the methods and objectives of the attack.
- Data events are aggregated and correlated by multiple sources and sensors.
- The impact of events is determined.
- Thresholds of incident alerts are established.

To develop the activities of detection, security analyst executes the following cognitive tasks: Review incident data; Review the events by aspects of interest; Pivot in the data to find atypical values or outliers; Expand the search to find more data; Investigate the threat to develop experience; Discover new threats; Determine indicators of commitment in other sources; Apply intelligence to investigate the incident; Discover IPs potentially infected.

2.3 Source's Information

For developing the cognitive tasks for the detection process, security analysts could consider the following sources of information: Vulnerability Information; Security intelligence feeds; Topology information; URL connection details; Domain Name System (DNS) logs; Intrusion Prevention System (IPS) logs; Operation Systems logs; Syslog's servers. Analysing the sources of information, security analysts could identify some features of the information. The sources have different formats; The volume of data is high; New data is generating each second; Some source of information has an unstructured data type; Information needs to be correlated.

The human capabilities to process this information could be a big challenge. Cybersecurity Incident Response Team (CSIRT) needs to take advantage of new technologies to support the cognitive tasks of security analysts to improve the detection activity process.

2.4 Big Data and Machine Learning for Security Incident Response

CSIRTs need to respond to security incidents in a short time and with an effective process to reduce the impact of cyberattacks. During the incident response process, relevant time indicators are:

- MTTD: Time to detect an incident, and
- MTTR: Time to resolve an incident.

Based on these time indicators CSIRT could design indicators of priority. The indicator of priority could have levels such as: high, medium, and low. Based on these priority indicators, CSIRT can build incident response plans that are a set of steps to manage a security incident (IRM, 2021). We developed examples of priority indicators that are shown in Table 1. Low indicators need 24 hours to resolve the incident, but high need only 16 hours. This time is dependable on the security context of each organization (Tello-Oquendo et al., 2019). Security analysts should complete all cognitive tasks in this period to resolve the security incident. However, the amount of information from different sources could oversaturate human capabilities (Chockalingam et al., 2017).

Table 1: Indicators based on the priority assigned to cyberattacks impact.

Indicator 1	High Priority
Scope	Capability to resolve incidents of high impact.
Method	Time to resolve incidents of high priority.
Green umbral	90% of incidents are resolved in $T \leq 16$ hours.
Yellow umbral	80% to 90 % of incidents are resolved in $T \leq 16$ hours.
Red umbral	80% of incidents are resolved in $T \leq 16$ hours.
Indicator 2	Medium Priority
Scope	Capability to resolve incidents of medium impact.
Method	Time to resolve incident of medium priority.
Green umbral	90% of incidents are resolved in $T \leq 24$ hours.
Yellow umbral	80% to 90 % of incidents are resolved in $T \leq 24$ hours.
Red umbral	80% of incidents are resolved in $T \leq 24$ hours.
Indicator 3	Low Priority
Scope	Capability to resolve incidents of low impact.
Method	Time to resolve incidents of low priority.
Green umbral	90% of incidents are resolved in $T \leq 40$ hours.
Yellow umbral	80% to 90 % of incidents are resolved in $T \leq 40$ hours.
Red umbral	80% of incidents are resolved in $T \leq 40$ hours.

When security analysts, development cognitive tasks, there are three macro cognitive processes involved: perception, comprehension, and projection.

- Perception. The cognitive capability to collect information.
- Comprehension. The cognitive capability to generate knowledge based on the collected information.
- Projection. The cognitive capability to predict future events based on the collected information and the knowledge generated.

Perception is the first cognitive process development for humans for decision-making. In this process, security analysts collect signals or signs for development judgment if one event is good or bad. But the amount of information generated for logs of

firewall, servers, network devices, or other source's data used for security analysts to detect patterns is very high in volume and complexity. Additionally, the time to execute this process is limited for the indicators of priority established.

In this context, BigData and ML could be one alternative to support this cognitive process (Elastic, 2021; Fahim et al., 2006; Wickham and Bryan, 2021; Hamerly and Elkan, 2004). The inclusion of Bigdata and Machine Learning could be a great support to resolve some issues related to managing the sources of information and its processing (Andrade et al., 2018b; Olukanmi et al., 2018). Technologies of BigData used in the cybersecurity context are shown in Table 2, while ML in the cybersecurity context is shown in Table 3.

Table 2: Use of BigData technologies in a cybersecurity context.

Cybersecurity context	Technologies
Anomaly detection	Hadoop/Apache spark
Network analysis	Hadoop/Apache spark
Alert correlation	Hadoop/Apache spark
Intrusion detection	Hadoop/Apache spark
Network monitoring	Hadoop
DDoS detection	Hadoop/Apache spark
Phishing detection	Hadoop/Apache spark
Cyber threat intelligence	Hadoop
Security events correlation	Hadoop/Apache spark

Table 3. Machine learning applied to incident response.

Scope	Technologies
Anomaly detection	Random forest, SVM, ANN (Salman et al., 2017; Subba et al., 2016; Muna et al., 2018).
Network analysis	Decision tree, KNN, SVM (Salman et al., 2017; Subba et al., 2016; Muna et al., 2018).
Alert correlation	SVM, Random Forest (Salman et al., 2017; Subba et al., 2016; Muna et al., 2018).
Cyber threat intelligence	ANN (Lee et al., 2019; Mishra et al., 2018).
Intrusion detection	ANN, KNN (Mishra et al., 2018; Malik and Khan, 2018).
Network monitoring	Decision tree. KNN, SVM (Malik and Khan, 2018).
DDoS detection	Random forest, NN (Alswailem et al., 2019; Wankhede S, and Kshirsagar, 2018).
Phishing detection	Random Forest (Malik and Khan, 2018).

3 RESEARCH METHODOLOGY

Cybersecurity analyst needs to process a large of data for information systems. The volume of data generated for IT systems grows with each new device connected to the network. Under this context, the inclusion of BigData and ML is a good tool, especially for anomaly detection and decision support. The objective of this work focuses on identifying how the cognitive process of security analysts is applied to the detection of security incidents when using machine learning and bigdata solutions. The research methodology develops two case studies related to security incidents. The first one is detecting fake news and the second one is the identification of malicious DNS. The two scenarios have been selected based on the premise of extensive data generated, as in the case of DNS and unstructured data handling as fake news. The datasets used are open access and have a malicious or regular data label to facilitate the implementation process in this work.

3.1 Fake News

Social media have become an essential part of our lives and for the entire world. As (Internet Live Stats, 2021) says, social media also plays an important part in the development process of young adults. Social media has also started to play an essential role in obtaining information and news (Vermeer et al., 2020). Under these circumstances, the detection of fake news is more important than ever; however, given the amount of information circulating on the internet daily (Muna et al., 2018), this detection process needs to be automated to process all the data. With this motivation, we decided to start developing an automated model for fake news detection. Initially, we need to obtain the data to perform the analysis. For this analysis, the data was obtained from some web pages that serve their checks on news circulating the Internet. These web pages function in the United States and perform a fact check on the most popular news. The pages used for collecting the data were: politifact.com, snopes.com, truthorfiction.com and checkyourfact.com. The news from all of these pages was collected using a technique called web scraping, which performs a download of the page and automates the obtention of the information needed. We used the library beautiful soup (Richardson, 2020) to perform the data collection in python. We can format the web pages we are inspecting and navigate them to obtain the required information with this library. After we find

the information, we can save it in a file with the corresponding rating given by the page, i.e., if the corresponding news are false, true, or a mixture. The data obtained was available in the mentioned pages as of February 25 of 2021. An example of the code used for scrapping one of the pages can be seen in Figure 1. Where the data is obtained from the page "politifact.com". As we can see, obtaining the data is simple, only needing to define the page's structure and save it using the "pandas" library.

```
newsData = list()
for i in range(1,647):
    url = "https://www.politifact.com/factchecks/list?page="+str(i)
    page = requests.get(url,headers=header)
    print(page,end="\n")
    soup = BeautifulSoup(page.text, "html.parser")
    titles = soup.find_all(class_ = "m-statement--is-medium")
    for title in titles:
        newsName = title.find(class_="m-statement__name").text[1:-1]
        newsDesc = title.find(class_="m-statement__desc").text[1:-1]
        newsQuote = title.find(class_="m-statement__quote").text[2:-2]
        newsRate = title.find(class_="m-statement__meter").div.picture.img["alt"]
        newsRate = "mostly-false" if newsRate=="barely-true" else newsRate
        news = [newsName+" "+newsDesc+" "+newsQuote,newsRate]
        newsData.append(news)

df_politifact = pd.DataFrame(newsData,columns=['News', 'Rating'])
df_politifact.to_csv('data/politifact.csv',index=False)
```

Figure 1: Web Scrapping in politifact.com.

Once the data has been collected, we can begin the analysis. First step in the analysis is to perform the pre-processing of the data, and this is done with the library nltk (NLTK Project, 2021) in python. This library allows us to start by removing the stop words found in the text; as (Richardson, 2020; NLTK Project, 2021) defines, a stop word is a word that doesn't add meaning to the sentence and only is there to make it more "human" readable, like the word "at", for example. After removing the stop words, we removed the punctuations inside the words and some other stop words, like dates. Finally, we can analyse the data by extracting the information of the most common words for each news class. As we can see in Figure 2, the most common words for fake (false) news are post, Facebook, and trump.

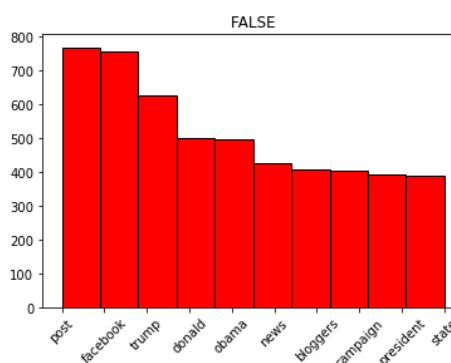


Figure 2: Most Common Words in Fake(False) News.

3.2 Malicious DNS Domains

The infrastructure used in this example is the spark for two main reasons: attackers could use DNS for executing cyber-attacks. In the case of phishing, attackers acquired a similar DNS domain to impersonate some organizations and try to trick their victims. Today, there are many DNS around the world, and the work of cybersecurity analysts is classified as the bad ones in blacklists. This process needs a lot of time and could be exhausting for cybersecurity analysts, but ML and machine learning could support this process. Spark provides the option for working with streaming data; Spark provides the opportunity for working with R and python for data analysis. The first step is preparing the dataset and define the set of variables to be used. The variables are built on basis of part of URL domains. Following we describe the code used in this step:

```
getSubdomain_udf = udf(lambda url:
tldextract.extract(url).subdomain, StringType())
getDomain_udf = udf(lambda url:
tldextract.extract(url).domain, StringType())
getSuffix_udf = udf(lambda url:
tldextract.extract(url).suffix, StringType())
getNoDigits_udf = udf(lambda text: len(re.sub("[^0-9]", "", text)), StringType())
```

Our interest, in this case, is to define the country and SSL certificated used for bad DNS domains. Following, we describe the code used for this step and Figure 3 shows the result.

issuer-CN	total
	2459769
cPanel, Inc. Certification Authority	745144
Sectigo RSA Domain Validation Secure Server CA	159254
Go Daddy Secure Certificate Authority - G2	19242
Actalis Domain Validation Server CA G3	3465
Sectigo RSA Organization Validation Secure Server CA	2978
Microsoft IT TLS CA 4	2090

Figure 3: Clusters visualized in 2 dimensions.

4 DISCUSSION

The cognitive process applied for a security analyst to detect anomalies is perception, comprehension, and projection. Security analysts use spatial and temporal reasoning to execute this mental process in information systems. According to Wickham (Wickham, 2014), temporal and spatial

representations are not independent between them, and different modalities are sensitive to time and space processing. Wickham analyses the hypotheses on the symmetric and asymmetric relationships of temporal and spatial information and how they affect the decision-making process (see Figure 4). Anomaly detection is based on the analyst's experience, so there is a high degree of subjectivity. Furthermore, the filtering process requires the use of the methodology for processing collected data, including their cleaning and profiling, such as CRISP-DM, which has been widely used in Data-Mining projects. Security analysts start from uncertainty and lack of relevant information and must build the necessary decision-making knowledge see Figure 5. Machine learning approaches could support the process of generating knowledge.

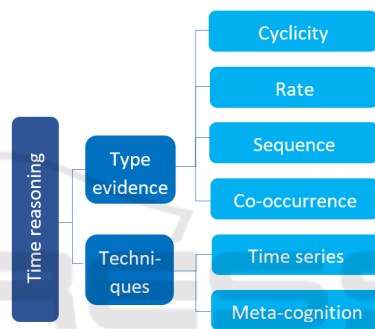


Figure 4: Temporal features for anomaly detection.

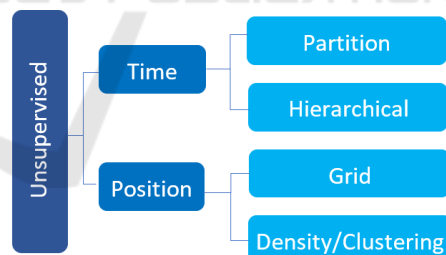


Figure 5: Space (position) and time feature applied to anomaly detection.

Based on the scenario of DNS malicious, we can observe that the inclusion of BigData and machine learning could support the cognitive process of the security analyst to generate knowledge and execute decision-making. We can summarize in Figure 6 the cognitive process to generate knowledge from a data-analytic process.

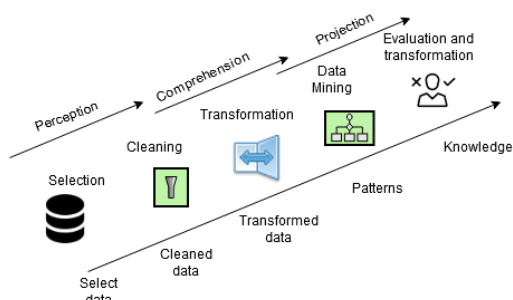


Figure 6: Cognitive process based on data analysis process..

REFERENCES

- Andrade, R., Torres, J., and L. Tello-Oquendo, 2018. Cognitive Security Tasks Using Big Data Tools 2018 International Conference on Computational Science and Computational Intelligence (CSCI) pp. 100-105, doi: 10.1109/CSCI46756.2018.00026.
- Andrade, R and Torres, J. 2018. Self-Awareness as an enabler of Cognitive Security 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) pp. 701-708, doi: 10.1109/IEMCON.2018.8614798.
- Andrade, R Torres, J and Tello-Oquendo, L. 2018. Cognitive Security Tasks Using Big Data Tools 2018 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 100-105, doi: 10.1109/CSCI46756.2018.00026.
- Certsocietegenerale/IRM. 2021. Retrieved April 26 2021, from <https://github.com/certsocietegenerale/IRM/tree/master/EN>
- Tello-Oquendo L, Tapia F, Fuertes W, Andrade R, Erazo N, Torres J, and Cadena A, 2019. A Structured Approach to Guide the Development of Incident Management Capability for Security and Privacy. In Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 2: ICEIS, ISBN 978-989-758-372-8 ISSN 2184-4992, pages 328-336. DOI: 10.5220/0007753503280336
- Chockalingam S, Pieters W, Teixeira A, van Gelder P, 2017. Bayesian Network Models in Cyber Security: A Systematic Review. Secur. IT Syst. 2017 105–122. doi:10.1007/978-3-319-70290-2.
- Elastic. Eland: DataFrames and Machine Learning backed by Elasticsearch - eland 7.10.0b1 documentation. Available online: <https://eland.readthedocs.io/en/7.10.0b1>. (Accessed on January 11 2021)
- Fahim A.M, Salem A.M, Torkey F.A, Ramadan M.A., 2006. An efficient enhanced k-means clustering algorithm. J. Zhejiang Univ. Sci. A, 7, 1626–1633. doi:10.1631/jzus.2006.A1626.
- Wickham H, Bryan J., 2021. Read Excel Files. Available online: <https://readxl.tidyverse.org>. (accessed on January 11 2021)
- Hamerly G, Elkan C. 2004. Learning the k in k-means. Adv. Neural Inf. Process. Syst. 17, 281–288.
- Olukanmi P.O, Nelwamondo F, Marwala T, 2018. k-Means-Lite: Real Time Clustering for Large Datasets. 2018 5th Int. Conf. Soft Comput. Mach. Intell. pp. 54–59. doi:10.1109/ISCMI.2018.8703210.
- Salman T., Bhamare D., Erbad A., Jain R., and Samaka M. 2017. Machine Learning for Anomaly Detection and Categorization in Multi-Cloud Environments. 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud). doi:10.1109/csccloud.2017.15
- B. Subba, S. Biswas, and S. Karmakar, 2016. Enhancing performance of anomaly-based intrusion detection systems through dimensionality reduction using principal component analysis, in 2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), IEEE, pp. 1-6.
- A.-H. Muna, N. Moustafa, and E. Sitnikova, 2018. Identification of malicious activities in industrial Internet of things based on deep learning models, Journal of Information Security and Applications, 41, 1-11
- J. Lee, J. Kim, I. Kim, and K. Han, 2019. Cyber Threat Detection Based on Artificial Neural Networks Using Event Profiles, IEEE Access, 7, 165607-165626
- P. Mishra, V Varadharajan, U Tupakula, E. S. J. I. C. S. Pilli, and Tutorials, 2018. A detailed investigation and analysis of using machine learning techniques for intrusion detection, 21(1), 686-728
- J. Malik and F.A.J.C.C. Khan, 2018. A hybrid technique using binary particles swarm optimization and decision tree pruning for network intrusion detection, 21(1), 667-680
- Alswailem A, Alabdullah B, Alrumayh N, and Alsedrani A, 2019. Detecting Phishing Websites Using Machine Learning. 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS). doi:10.1109/cais.2019.8769571
- Wankhede S, and Kshirsagar D, 2018. DoS Attack Detection Using Machine Learning and Neural Network. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). doi:10.1109/iccubea.2018.8697702
- Internet Live Stats. 2021. Total number of websites Home Page. Available online: <https://www.internetlivestats.com/watch/websites> (accessed on January 11 2021)
- Vermeer S, Trilling D, Kruikemeier T, and De Vreese C, 2020. Online News User Journeys: The Role of social media, News Websites, and Topics, Digital Journalism, 8(9), 111431141
- Richardson, L. 2020. Beautiful Soup Documentation — Beautiful Soup 4.9.0 documentation, Crummy.com, 2020. [Online]. Available: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. [Accessed: 30- Apr-2021].
- NLTK Project, 2021. Natural Language Toolkit — NLTK 3.6.2 documentation, Nltk.org, 2021. [Online]. Available: <https://www.nltk.org/>. [Accessed: 30- Apr-2021].
- Wickham H 2014. Tidy data. The Journal of Statistical Software, 14(5), Retrieved from: <http://www.jstatsoft.org/v59/i10>.