Cybersecurity Scientometric Analysis: Mapping of Scientific Articles using Scopus API for Data Mining and Webscrapping

Carlo H. Godoy Jr.
Bayan Academy
Quezon City, Philippines
jhayar11090408@gmail.com

Nicole Jehru R. Diego
Bayan Academy
Quezon City, Philippines
njrdiego@dswd.gov.ph

Russel E. Tagumasi
Bayan Academy
Quezon City, Philippines
russeltagumasi@gmail.com

Jordan C. Lerit
Bayan Academy
Quezon City, Philippines
jordanlerit@gmail.com

Jefferson A. Costales
Eulogio "Amang" Rodriguez Institute
of Science and Technology
Manila, Philippines
jacostales@earist.ph.education

2022 5th International Conference on Data Science and Information Technology (DSIT) | 978-1-6654-9868-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/DSIT55514.2022.9943876

***Abstract*—In terms of computers, cybersecurity has experienced massive technological and operational transformations in recent years, with data science at the forefront. Cybersecurity in the Philippines is still believed to be in its infancy. Though it is commonplace in other areas of the world, just a few scholars in the Philippines have dared to concentrate in this field. As a result, a need for assistance in promoting cybersecurity in the Philippines must be identified. Because it is difficult to persuade scholars in the Philippines to work on a certain field of computing without figures, Scientometric Analysis is required to persuade scholars in the Philippines to handle cybersecurity. The findings of this study revealed that there is no link between the frequency with which authors publish cybersecurity and their chances of receiving a high number of citations, either by year or overall. The study also found that having a lot of publications does not ensure a higher Google Scholar ranking. According to the study, the most interesting themes in the field of cybersecurity are to connect it with another discipline, thus an author should choose an attractive topic in this field to acquire a higher rank in Googlescholar and especially in Scopus. Machine Learning and IoT are the two fields that most people want to see integrated into cybersecurity. Hence, this topics can be used to encourage cybersecurity experts in the Philippines to used this topic on their research or speaking engagements.**

Keywords—cybersecurity, data mining, webscrapping, scientometric analysis, scientific mapping (*key words****)***

I. INTRODUCTION

In the context of computing, cybersecurity has undergone tremendous technological and operational changes in recent days, with data science driving the transition. The key to making a security system automated and intelligent is extracting security event patterns or insights from cybersecurity data and constructing a data-driven model to match. Data science is the application of numerous scientific methodologies, machine learning techniques, processes, and systems to comprehend and analyze actual occurrences with data [1]. On the other hand, Data mining is the process of looking through a huge pre-existing database to find new information. A strong technology has a lot of potential for assisting firms in focusing on the most important data in their data warehouses. Data mining technologies and approaches will help businesses foresee future trends by allowing them to make more proactive, knowledge-based decisions. Data mining techniques could help answer business-related queries that were previously too time-consuming to answer [2].

In the Philippines, cybersecurity is still considered to be in its baby stage. Though in other parts of the world it is already part of their computing practices, in the Philippines, there are only a few scholars who are brave enough to specialize on this field. Hence, a need for support to promote cybersecurity in the Philippines must be established. Since everyone knows that in the Philippines it is hard to encourage scholars to work on a certain area of computing without numbers, Scientometric Analysis is needed to convince the scholar in the Philippines to tackle cybersecurity. Scientometrics is identified as the "statistical study of science, scientific interaction, and policy analysis," and it includes research impact measurement, exploring the impact of organisations and journals in a specific area of research, and providing a deeper knowledge of science citation [3]. In this study, scientometric analysis will be used as a sort of data-driven management tool to encourage more scholars in the Philippines to be interested in cybersecurity.

II. SCIENTOMERIC ANALYSIS, SCIENTIFIC MAPPING, DATA MINING AND WEB SCRAPPING

A. Scientomeric Analysis

Scientometrics is a branch of science that focuses on science and has its own methodology. The term has gained in use and prominence in recent decades, particularly since Tibor Braun founded the specialized Journal of Scientometrics in 1978. It is a term that is used to define the study of science, including its growth, organization, interrelationships, and productivity. Scientometrics can offer existing trends backed up by quantitative data and provides a full picture of research activity in the topic [4]. Because the existing instruments for Scientometric analysis have diverse capacities and strengths, a full examination of any topic requires the employment of numerous tools for various types of study. The features and limits of various Scientometric tools were examined in order to identify the tools needed for the current investigation. VOSviewer, BibExcel, CiteSpace, CoPalRed, and Sci2 are examples of software. The results of the analysis of VantagePoint and Gephi led to the selection of VOSviewer, CiteSpace, and Gephi are three applications

that can be used to view video. VOSviewer is an acronym for VOSviewer. is a free application that allows you to see similarities in different ways. the fundamental capabilities for viewing Scientometric networks. CiteSpace is a scientific literature mapping tool that can visualize multiple network layouts, discover clusters, and emerging trends, including rapid changes in the scientific literature, as well as provide cluster and time-zone views. Gephi is an open source network graph and analysis tool that is at the vanguard of the network visualization and analysis revolution, and can be used to gain a comprehensive understanding of the information available from a given network [5].

B. Scientific Mapping

Scientific breakthroughs are made by linking previously unconnected pieces of knowledge in novel and creative ways. Understanding the processes of scientific output requires mapping the relationships and structures of scientific knowledge. Science mapping, also known as bibliometric mapping, is a hot issue in the field of bibliometry and scientometric analysis. This method, also known as scientific mapping or bibliometric mapping, is a hot topic in bibliometry research. It's a study of how different disciplines, fields, experts, and particular papers or authors interact with one another. In essence, this study focuses on keeping an eye on the field and establishing research boundaries in order to determine the field's cognitive structure and evolution. Science mapping, in other words, tries to depict the structural and dynamic features of scientific research [6].

C. Data Mining

Data mining is a technique for extracting patterns or models from large amounts of data. This technique is used in a variety of settings, including biology, education, and finance, as well as business, law enforcement, and political processes. There are a variety of approaches used in data mining, including rule induction and decision trees. According to numerous research, are among the most commonly utilized [7].

Data mining is also defined as a method of extracting knowledge from a database. This information can be categorized into different rules and patterns that can assist a user or organization in analyzing aggregate data and predicting decision processes. Data warehouse is a centralized database for any organization, where all data is housed in a single large database. Data mining is a technique that organizations employ to extract meaningful information from raw data. Software is used to search for desired patterns in a large amount of data (data warehouse), which can assist businesses in learning more about their customers, predicting behavior, and improving marketing efforts [8].

D. Web Scrapping

The Internet is the world's largest database of data ever created. It includes a wide range of self-explanatory materials in a variety of formats, including audio/video, text, and others. The badly designed data that mostly fills the Internet, on the other hand, is difficult to extract and use in an automated process. Web scraping eliminates the need for manual data extraction and organization by providing an easy-to-use method for collecting data from webpages, converting it to a desired format, and storing it in a local repository. Many firms utilize numerous methods to collect relevant information from the web, owing to the wide range of uses of Web scraping, which range from lead generation to reputation and brand management, from sentiment analysis to data augmentation in machine learning [9].

Big Web Data has three characteristics namely: volume, variety, and velocity, given this information, individual researchers or even huge teams of academic researchers or commercial data professionals will struggle to collect and organize this data manually. As a result, researchers frequently use a variety of technologies and applications to automate some or all elements of data collecting and management on the Web. Web scraping is the term for the developing activity of automatically extracting and organizing data from the Web for the purpose of subsequent analysis [10].

III. FROM DATA MINING TO SCIENTOMETRIC ANALYSIS

Structured, semi-structured, and unstructured quantitative and qualitative data in the form of Web pages, HTML tables, Web databases, emails, tweets, blog posts, images, and videos are all available on the Internet. To make use of Big Web Data, a number of technical difficulties connected to the volume, variety, velocity, and authenticity of data on the Internet must be addressed [11]. In line with this, the proponent will be following a certain process from data mining using scopus API, to web scrapping using publish and perish desktop application back to data mining and data cleansing thru excel and MS SQL to Scientific Mapping of the data collected and established to the MS SQL databased. The conceptual framework as shown in Figure 1 defines the summary of the process that will be followed by the proponents.

Fig 1: Conceptual Framework

A. Data Mining and Web Scrapping using Scopus API

To start the data mining process in the Scopus database the proponent will go to Elsevier Developer Portal. After going to the portal, the proponent will then click I want an API key. Since the proponent has no existing API Key, the proponent will be clicking on create API Key. This API key will then be stored and will be unique whenever the proponent will be mining data from the Scopus database.

The next process is to open the Publish or Perish installed on the proponent's computer. The application will show the

different databases available that can be datamined and web scrapped. In this regard, the proponent will be clicking Scopus. After clicking Scopus, Search bars for the following will be shown: 1) Authors; 2) Years; 3) Affiliations; 4) Publication Name: 5) Title Words; and 6) Keyword. Other visibility will be a table consisting of the different information of the web scrapped publications under the Scopus database. In regards to the proponent's study, the proponents will be using the title/word field and will be typing the word cybersecurity.

Once the proponent's starts to click on the search button, 200 articles will be selected but prior to the result, the application will be asking for the API key. Hence, the proponent will need to go back to the Elsevier Development Portal and copy the API Key and paste it to the perish or publish application.

The proponents will let the perish or publish application to choose the publication. Since it will be 200 per web scrapping process, the proponents will be generating 200 publication per year from 2018 to 2022 which will result to a total of 1000 datasets. After downloading the CSV file, the data will now be ready for data cleansing and analysis using MS SQL.

B. *From Data Cleansing to Scientific Mapping*

- The proponents will encode the data to MS SQL to prepare for data cleansing. Search queries will be used to locate duplicate data to ensure that data are clean to provide a better data analysis result.
- Once the data is cleaned, it is now time to perform data mapping. The proponent will be using coded command to perform scientific mapping.
- The following will be considered for the scientific mapping process: 1) Study about Cybersecurity With Highest Number of Citations; 2) Study about Cybersecurity With Highest Number of Citations Order by Year; 3) Study about Cybersecurity With Highest Number of Citations Per Year; 4) The Most Active Authors in the Field of Cybersecurity published in Scopus; and 5) Study group per type, published year and total citations.
- The tables will then be downloaded to an excel file for the preparation of the graph that will be needed for the Data Visualization.

C. *Preparation of the Data Visualization*

After scientifically mapping the information needed by the proponents, the five tables downloaded from MS SQL will then be prepared for the Data Visualization Method. The proponents will be using excel to convert the tables into graphs so that the proponent would be able to come up with a better data analysis. This visualization will be safekeep and will be the sole basis of the recommendation that the proponents will be making.

To start, the proponent will be importing the CSV file into the python using Pandas Data frames. After which, the graphs will be created with Matpotlib using Pandas Dataframes.

D. *Data Analytics*

Analytics is the science of evaluating raw data with the goal of deriving inferences and applying the information to make decisions. Pictures were used to communicate ideas, goals, and history before the formal written language was developed. The majority of our knowledge of our forefathers comes from these cave or monument paintings. Visualizations such as bar charts, scatter plots, and dashboards are important tools in business intelligence today because they help managers assimilate information and make quick choices. Managers benefit greatly from dashboards because they can view several charts and graphs that provide the most recent information about sales, returns, market share, and other topics, keeping them informed about the company's current developments [12].

In this process, the proponents will now then device the top 5 based on the data analysis of the 5 tables as visualize on the acquired table and graphs. This is where the proponents will be getting their final say on the recommendation and conclusion of the study.

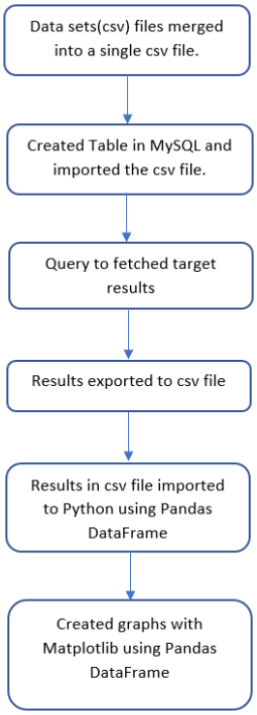

Fig 2: Data Visualization Process Summary

IV. DISCUSSION OF THE SCIENTOMETRIC RESULT

The bibliographic data for this study was gathered from the Scopus database. It is the world's largest multidisciplinary database of peer-reviewed articles, and it is widely recognized and used for quantitative research. However, the metadata retrieved from Scopus is ambiguous due to multiple representations of author names, author affiliations, titles, journals, volume and issue numbers, and page numbers in the original manuscript. As a result, building a clean data set necessitated a significant amount of manual labor in order to assure the data's integrity and the correctness of the study. The following are the result using the scientometric analysis.

A. *Cybersecurity Topics and Authors rank by GS Rank*

After the data cleaning, the proponents did a data visualization to be able to know the different authors who wrote cybersecurity topics and is rank thru its Google Scholar Rank. Figure 6 shows all of the publication with a cybersecurity topic and its corresponding author.

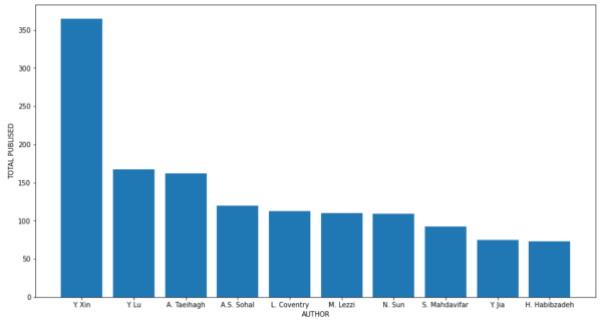

Fig 3: Cybersecurity Authors rank by their GS Rank

After performing a data visualization, the proponent selected the top 5 cybersecurity topic with the highest GS rank which is shown in Table I.

TABLE I. TOP 5 CYBERSECURITY TOPICS RANK BY GOOGLE SCHOLAR(GS) RANK

GS Rank	Publication Details			
	Author	***Title***	***Cite Count***	***Year***
3	R.K. Sakthivel	Core-level cybersecurity assurance using cloud-based adaptive machine learning techniques for manufacturing industry	7	2022
3	P. Dixit	Deep Learning Algorithms for Cybersecurity Applications: A Technological and Status Review	28	2021
3	A. Corallo	Cybersecurity in the context of industry 4.0: A structured classification of critical assets and business impacts	56	2020
3	N. Sun	Data-Driven Cybersecurity Incident Prediction: A Survey	109	2019
3	L. Coventry	Cybersecurity in healthcare: A narrative review of trends. threats and ways forward	113	2018

B. *Cybersecurity Topics and Authors rank by Number of Citations Per Year (CPY)*

Another important observation that has made by the proponents after performing data visualization is the rank of authors and their corresponding publications which has been rank thru the number of citations per yer (CPY). Figure 7 shows the data visualization for the ranking of the authors by their number of citations per year.

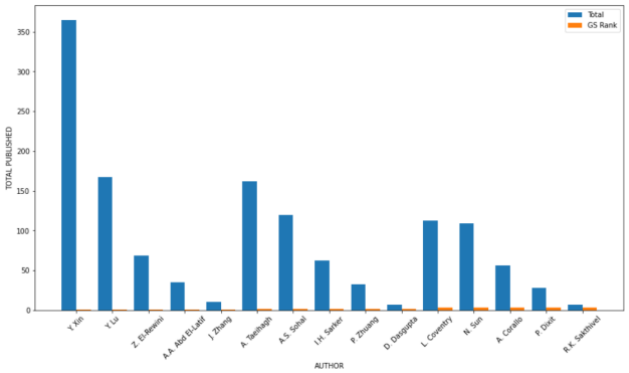

Fig 4: Cybersecurity Authors rank by their number of citations per year (CPY)

As mentioned before, another important thing that the data visualization shows the proponents are the different topics that has been published by the authors rank in Figure 7. To be able to trim down the list, the proponents selected the top 5 topics that has been published by the top 5 authors from the rank specified in Fugure 7. In this way, the proponents would be able to follow its goal in finding the best topic that can be suggested for future researchers or those who want to talk about cybersecurity in their future seminars. Table II show the top 5 cybersecurity topics that has been published by the top 5 authors who were rank by their number of citations per year.

TABLE II. TOP 5 CYBERSECURITY TOPICS RANK BY NUMBER OF CITATIONS PER YEAR (CPY)

CPY	Publication Details			
	Author	***Title***	***GS Rank***	***Year***
91	Y. Xin	Machine Learning and Deep Learning Methods for Cybersecurity	1	2018
56	Y. Lu	Internet of things (IoT) cybersecurity research: A review of current research topics	1	2019
54	A. Taeihagh	Governing autonomous vehicles: emerging responses for safety. liability. privacy. cybersecurity. and industry risks	2	2019
36	N. Sun	Data-Driven Cybersecurity Incident Prediction: A Survey	3	2019
35	A.A. Abd El-Latif	Quantum-Inspired Blockchain-Based Cybersecurity: Securing Smart Edge Utilities in IoT-Based Smart Cities	1	2021

C. *Cybersecurity Topics and Authors rank by their Cites Count*

After checking the number of citations per year, the proponent decided also to identify the ranking of the authors by their overall citation counts. In this manner, the proponents would be able to select the best of the best among all the publications that has been written from 2018 to 2022. Figure 8 shows the ranking of all the authors that has been identified after data cleansing.

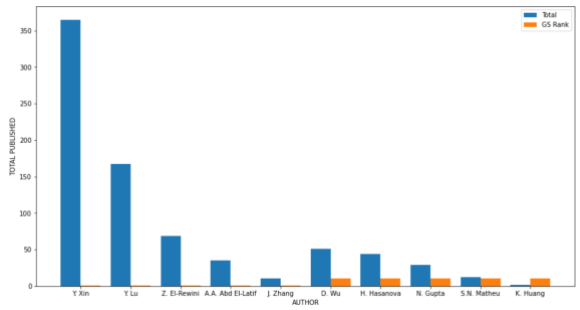

Fig 8: Cybersecurity Authors rank by their cites count

After ranking the authors by their cites count, the proponents also decided to select the top 5 authors and their respective publications. Table III shows the data of the top 5 cybersecurity topic gathered from the authors who were rank by their cites count.

TABLE III. TOP 5 CYBERSECURITY TOPICS RANK BY CITES COUNTS

Cites Count	Publication Details			
	Author	*Title*	*GS Rank*	*Year*
365	Y. Xin	Machine Learning and Deep Learning Methods for Cybersecurity	1	2018
167	Y. Lu	Internet of things (IoT) cybersecurity research: A review of current research topics	1	2019
162	A. Taeihagh	Governing autonomous vehicles: emerging responses for safety. liability. privacy. cybersecurity. and industry risks	2	2019
120	A.S. Sohal	A cybersecurity framework to identify malicious edge device in fog computing and cloud-of-things environments	2	2018
113	L. Coventry	Cybersecurity in healthcare: A narrative review of trends. threats and ways forward	3	2018

D. *The Most Active Authors in the Field of Cybersecurity published in Scopus*

After ranking the authors by their GS Rank, Citations per Year and their Cites count, the proponent also decided to double check the most active authors in the field of cybersecurity which is shown in Figure 9. The main reason for doing this is to check whether those authors who has the highest rank are also the ones who are active in publishing in the field of cybersecurity.

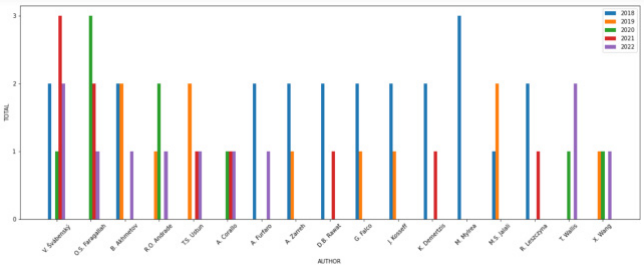

Fig 9: Most Active Authors in the Field of Cybersecurity published in Scopus

After seeing the data visualization for the authors, the proponents then again captured the top 5 authors which is shown in Table IV to be able to compare it by the top 5 authors who were found in the ranking by their GS Rank, Citations per Year and their Cites count. Results has shown that the most active authors in the field of cybersecurity were also not the ones who published in this field often.

TABLE IV. TOP 5 MOST ACTIVE AUTHORS IN THE FIELD OF CYBERSECURITY PUBLISHED IN SCOPUS

Author	Publication Details		
	Total	*Year*	*Nr of Publication per year*
V. Švábenský	8	2022	2
		2021	3
		2020	1
		2018	2
O.S. Faragallah	6	2022	1
		2021	2
		2020	3
B. Akhmetov	5	2022	1
		2019	2
		2018	2
R.O. Andrade	4	2022	1
		2020	2
		2019	1
T.S. Ustun	4	2022	1
		2021	1
		2019	2

E. *Type of publications that is being used to published cybersecurity topics*

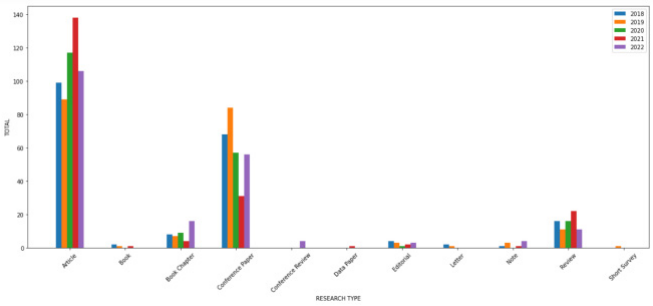

Fig 10: Type of publications that is being used to published cybersecurity topics

Figure 10 shows the most common type of publication platform being used by the different authors who published in the field of cybersecurity. As shown in Table V, the most utilized publication type from 2018 to 2020 is Article, Book Chapter, Conference Paper, Editorial and Review. It is also identified that it is the most consistent type that has been seen in the five-year range.

TABLE V. PUBLICATION TYPE BEING USED TO PUBLISHED CYBERSECURITY TOPICS

Publication Type	*Publication Details*

	Year	*Total Cities*
Article	2018 - 2022	549
Book	2018,2019,2021	4
Book Chapter	2018 - 2022	44
Conference Paper	2018 - 2022	296
Conference Review	2022	4
Data Paper	2021	1
Editorial	2018 - 2022	13
Letter	2018,2019	3
Note	2018,2019,2021, 2022	9
Review	2018 - 2022	76
Short Survey	2019	

V. CONCLUSION AND RECOMMENDATION

The Internet may be used to access email, do schoolwork, buy books, and negotiate online transactions that need personal information in a variety of ways. However, this type of Internet reliance has its own set of problems and consequences. Students who use the Internet frequently face a variety of risks, including exposure to inappropriate or potentially dangerous information, disclosure of sensitive and personal information, online purchase scams, and enticement by cyber-predators who want to meet them in person for nefarious purposes. According to studies, there are a variety of factors that influence people's cybersecurity habits. While technological controls are important for computer information security, cybersecurity also rely on an individual's security opinion [13]. Future studies should consider the individual's psychological behavior as prior research have highlighted its significant influence on technology adoption [14,15,16] and internet behavior [17,18,19]

Results have shown that there is no significant relationship between the authors who published cybersecurity often and the possibility of them getting a high number of citation count either by year or in totality. The study have shown as well that it will not guarantee a higher rank in google scholar if the author published often. In the field of cybersecurity, it is better for an author to choose an interesting topic in this field to get a higher rank in Google scholar and specially in Scopus.

According to the study the most interesting topics in the field of cybersecurity is to integrate it with another field. The most common field that interest the people the most to be integrated in cybersecurity is Machine Learning and IOT.

ACKNOWLEDGEMENT

The author would like to thank Bayan Academy for giving us the opportunity to be trained in advance data analytics and as a result we were able to come up with this novel study.

REFERENCES

[1] Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big Data, 7(1).

[2] Osman, A. S. (2019). Data Mining Techniques: Review. 2(1), 1–4. https://www.educba.com/7-data-.

[3] Martinez, P., Al-Hussein, M., & Ahmad, R. (2019). A scientometric analysis and critical review of computer vision applications for construction. Automation in Construction, 107(May).

[4] Ramy, A., Floody, J., Ragab, M. A. F., & Arisha, A. (2018). A scientometric analysis of Knowledge Management Research and Practice literature: 2003-2015. Knowledge Management Research and Practice, 16(1), 66–77.

[5] Hosseini, M. R., Martek, I., Zavadskas, E. K., Aibinu, A. A., Arashpour, M., & Chileshe, N. (2018). Critical evaluation of off-site construction research: A Scientometric analysis. Automation in Construction, 87(July 2017), 235–247.

[6] Yildiz, T. (2019). Examining the Concept of Industry 4.0 Studies Using Text Mining and Scientific Mapping Method. Procedia Computer Science, 158, 498–507.

[7] Viloria, A., Acuña, G. C., Franco, D. J. A., Hernández-Palma, H., Fuentes, J. P., & Rambal, E. P. (2019). Integration of data mining techniques to postgresQL database manager system. Procedia Computer Science, 155(2018), 575–580.

[8] Hamid Mughal, M. J. (2018). Data mining: Web data mining techniques, tools and algorithms: An overview. International Journal of Advanced Computer Science and Applications, 9(6), 208–215.

[9] Nigam, H., & Biswas, P. (2021). Web Scraping: From Tools to Related Legislation and Implementation Using Python BT - Innovative Data Communication Technologies and Application (J. S. Raj, A. M. Iliyasu, R. Bestak, & Z. A. Baig (eds.); pp. 149–164). Springer Singapore.

[10] Krotov, V., & Silva, L. (2018). Legality and ethics of web scraping. Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018, May.

[11] Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and ethics of web scraping. Communications of the Association for Information Systems, 47(1), 539–563.

[12] Sharma, A. M. (2020). Data Visualization. In S. Kumari, K. K. Tripathy, & V. Kumbhar (Eds.), Data Science and Analytics (pp. 1–22). Emerald Publishing Limited.

[13] Caluza, L. J. B., Quisumbing, L. A., Verecio, R. L., & Tibe, D. S. (2018). International Journal of Social Science and Economic Research VIEWS ON CYBERSECURITY PRINCIPLES AND PRACTICES : THE CASE OF BS INFORMATION TECHNOLOGY STUDENTS OF LNU ,. 01, 289–296.

[14] R. Ebardo, J. De La Cuesta, J. Catedrilla, and S. Wibowo, "Peer Influence, Risk Propensity and Fear of Missing Out in Sharing Misinformation on Social Media during COVID-19 Pandemic," in Proceedings of the 28th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education, 2020, pp. 351–359.

[15] J. Catedrilla et al., "Loneliness , Boredom and Information Anxiety on Problematic Use of Social Media during the COVID-19 Pandemic," in Proceedings of the 28th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education, 2020, pp. 52–60.

[16] R. Ebardo, "Predictors of Cyber-plagiarism : The Case of Jose Rizal University," in Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education, 2018, pp. 778–783.

[17] R. Ebardo and M. T. Suarez, "Older Adults and Online Communities: Recent Findings, Gaps and Opportunities," Int. J. Web Based Communities, vol. 17, no. 1, p. 1, 2021.

[18] R. Ebardo, J. B. Tuazon, and M. T. Suarez, "We learn from each other: Informal learning in a facebook community of older adults," ICCE 2020 - 28th Int. Conf. Comput. Educ. Proc., vol. 2, no. December, pp. 594–601, 2020.

[19] J. de la Cuesta, J. Catedrilla, R. Ebardo, L. Limpin, C. Leaño, and H. Trapero, "Personality traits of future nurses and cyberchondria: Findings from an emerging economy," in ICCE 2019 - 27th International Conference on Computers in Education, Proceedings, 2019, vol. 2