Cybersecurity News Dataset and Exploratory Analysis for Early Detection of Threat

Md Faisal Ahmed*, Md. Tawhid Anwar*, Sifat Tanvir*, Ramkrishna Saha*, Syed Zamil Hasan*, Md. Sabbir Hossain* and Annajiat Alim Rasel*
*Department of Computer Science and Engineering, School of Data and Sciences, Brac University, Brangladesh
Email: *md.faisal.ahmed@g.bracu.ac.bd, *write2tawhid@gmail.com, *sifattanvirsami@gmail.com, *rk.saha@bracu.ac.bd, *shoumo.hasan@bracu.ac.bd, *md.sabbir.hossain@g.bracu.ac.bd, *annajiat@gmail.com

***Abstract*— Due to the rapid increase in the use of the internet and technology, issues of security have risen. Hence, it has become crucial to assist users who are potentially under attack and are looking out for help. The data that has been made available in this article, has been gathered and marked from the 'The Hacker News'. The total number of collected news articles is 3742 and the total word count of this corpus is 1408035. The dataset is built with the aim of developing a classification model that can read a news article about a hacking event and determine under which category of attack the article belongs. The features of the dataset created are the news article text and the label is the category of it. There are four categories: 'Cyber Attack', 'Malware', 'Vulnerability' and 'Data Breaches'. Furthermore, the dataset also provides the source of the news article from where it has been collected along with the title of the article. This dataset can be used by researchers for creating a model where they will be able to take input data from users and recommend the potential attack they might be under. The dataset is publicly available at "https://data.mendeley.com/datasets/n7ntwwrtn5".**

Keywords—Cybersecurity, NLP Dataset, Natural language processing (NLP), News Dataset, Text classification, Data Mining

I. INTRODUCTION

In recent years, there has been a surge in the number and types of cyber-attacks as a result of increased digitization and its processes. According to research conducted in the cyber scene in 2019, there has been a 67% increase in security breaches over the last five years [1], with consequences ranging from financial loss to reputational damage [2]. Thus, by incorporating novel risk prevention and mitigation measures, the cyber security sector is expected to grow dramatically [3]. As a result, by presenting the dataset for building a text classifier due to the field's expanding complexity, this work contributes to the state of the art.

This dataset may be used as open-source data to discover emerging vulnerabilities and decrease the chance of them being exploited. It is feasible to detect developing cybernetic risks by collecting, cleaning, and analyzing the gathered news articles. This method is effective in both public [4] and private [5] contexts. The effectiveness of such approaches in the realm of cybersecurity is dependent on developing optimum ways for transforming open source data into valuable intelligence, which helps minimize the danger of cybernetic vulnerabilities being exploited [6]. The amount, diversity, authenticity, and speed with which open-source data is published are all current challenges. As a result, the creation of a sufficient data collecting and processing pipeline, as well as the identification of relevant data sources, is required for the effectiveness of a similar approach for the early detection of cybernetic vulnerabilities.

As part of a system for automatic detection of early cybernetic threats, this paper presents the most recent dataset that may be utilized to identify developing cybernetic vulnerabilities in cybersecurity news articles. We introduce a corpus of 3742 labeled articles on which various machine learning and deep learning models can be trained for classifying cybersecurity articles.

A few recent works that have been done by the researchers, analyzed specialized articles as the primary source of information to detect potential cyber-attacks. Abdullah, et al. [7] devised a method for providing timely information on potential emerging vulnerabilities while addressing the issue of different styles and structures in online articles. Zhou, et al. [8] created a NER solution capable of identifying indicators of compromise in an article's body, followed by a classification solution that quantifies the extracted article's relevance. Husari, et al. [9] present a model that utilizes specialist articles and is trained by SVM to produce early alerts about cybernetic risks. Another study[10] proposed a similar SVM approach for identifying relevant articles based on factors including the density of cybersecurity words and the length of the article..

II. EXPERIMANTAL DESIGN

The data was obtained from The Hacker News website which is a widely-read infosec source of the latest hacking news, cyber-attacks, computer security, and cybersecurity. We scrapped the data using Selenium to automatically extract the data without any sort of human error. We captured the title, URL, news article, and its category while collecting the data. The category was defined by the authors or the journalist of that article. Finally, all the collected data was stored in an excel file, which we are proposing in this paper.

III. DATA DESCRIPTION

The dataset[11] attached with this article is an excel file that contains title, URL, news article, and category or label. The total amount of collected articles is 3742 and the total word count of this corpus is 1408035.

Table 1: Variables of dataset

Variable	Type	Description
Title	Text	Title of the news article
Link	URL	URL of the news article
Article	Text	Text of the news article
Label	Categorical	Category of the news article

Here, table 1 shows the 4 types of variables used in the dataset, and the data type and description of each of the variables. Title, link, and article here are independent variables and label is the only dependent variable. The titles have been extracted from the list of titles of the news articles, the texts have been extracted from corresponding news articles, the links are also collected which is basically the URL using which the news can be directly accessed. Finally, the labels are the categorical values that are assigned to each of the articles based on their nature.

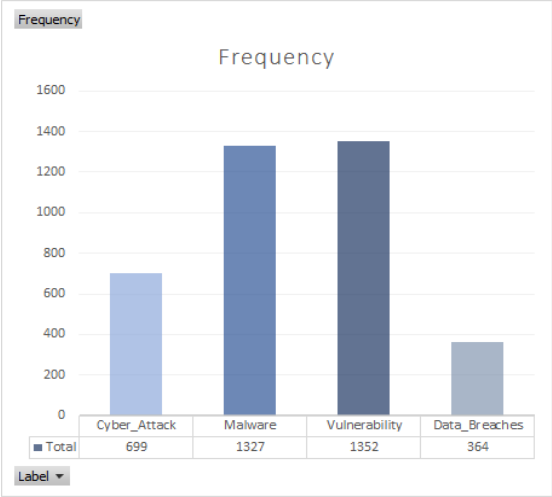

Fig. 1 Frequency Chart

Fig.1 shows the frequency of each category of news articles. The category with the most number of data is Vulnerability with 1352 entries. Malware is the second most frequent category with 1327 entries. Cyber_Attack comes in third place with 699 entries. Data_Breaches has the least amount of data, 364 entries.

Table 2: Frequency Table

Label	Label	Label
Cyber_Attack	699	18.68%
Malware	1327	35.46%
Vulnerability	1352	36.13%
Data_Breaches	364	9.72%

Table 2 demonstrates the percentage of each of the categories present in the dataset. Vulnerability and Malware are very close with 36.13% and 35.46% frequency percentages respectively. Data_Breaches has the lowest percentage with 9.72% frequency and Cyber_Attack has the second-lowest with 18.68% frequency.

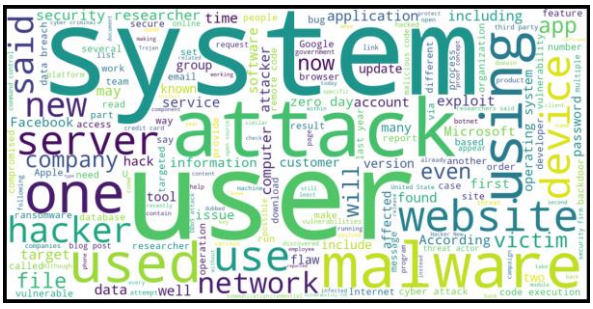

Fig. 4 Wordcloud of the corpus

A word frequency diagram has been shown in fig. 2. The figure shows some of the most occurring words found in the dataset. The word security has occurred 9080 times which is higher than any other word. Malware comes next with the occurrence of 7468 times.

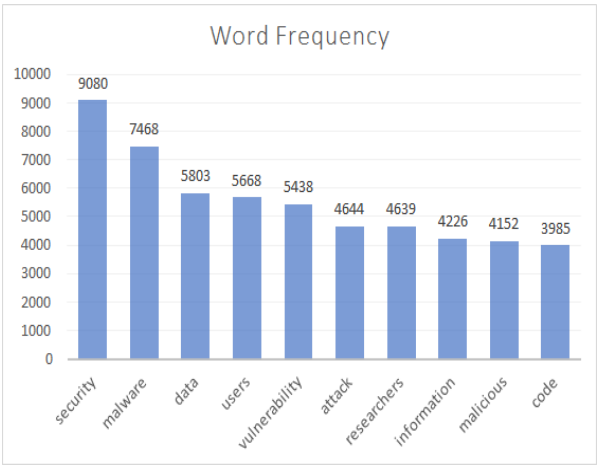

Fig. 2 Word Frequency Diagram

Table 3 gives an idea about the relationship between the word count and the categories. The malware had the most number of words, 614548 in a total of 1327 articles it was featured in whereas Data_Breaches had the least with the 179926-word count in 364 articles. The vulnerability had 577449 words in 1352 articles. Finally, Cyber_Attack had 361112 words in 699 articles. Fig. 3 gives a graphical representation of table 3.

Table 3: Total word count

Label	Number of articles	Word count
Data breaches	364	179926
Malware	1327	614548
Vulnerability	1352	577449
Cyber Attack	699	361112

The word cloud in Fig 4. represents the graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text.

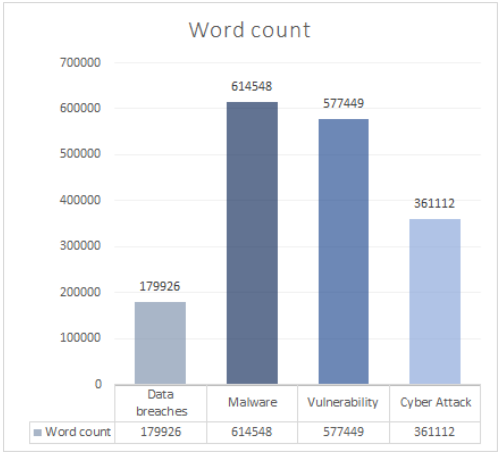

Fig. 3 Word Frequency Per Category

IV. Conclusion and Future work

This dataset was created with great care, and all of the data has been anonymized. These data can aid researchers in identifying possible cyber threats before it's too late. Various machine learning and deep learning models can be created with promising results based on the generated corpus of 3742 articles. Moreover, on this vast collection of articles from the cybersecurity domain, researchers may experiment with several language models, including a BERT model. Using the given dataset, future work will focus on integrating a larger system addressing the automatic identification of early cybernetic threats.

References

[1] K. Bissell, R. Lasalle, and P. Cin, "Ninth Annual Cost of Cybercrime Study," Ponemon Institute 2019.

[2] I. Ponemon, "Costs of Data Breach Study," Ponemon Institute, June 2017.

[3] L. Columbus, "2020 Roundup Of Cybersecurity Forecasts And Market Estimates", 2020. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2020/04/05/2020-%20roundup-of-cybersecurity-forecasts-and-market-estimates/?sh=7b34a04c40e2. [Accessed: Sep 3rd 2021]

[4] C. Andrew, R. J. Aldrich, and W. K. Wark, Secret Intelligence: A Reader: Routledge, 2009. [5] D. R. Hayes and F. Cappa, "Open-source intelligence for risk assessment," Business Horizons, vol. 61, pp. 689–697, 2018.

[6] L. Rosa, M. Freitas, S. Mazo, E. Monteiro, T. Cruz, and P. Simoes, "A Comprehensive Security Analysis of a SCADA Protocol: From OSINT to Mitigation," IEEE Access, vol. 7, pp. 42156–42168, 2019.

[7] M. S. Abdullah, A. Zainal, M. A. Maarof, and M. Nizam Kassim, "Cyber-Attack Features for Detecting Cyber Threat Incidents from Online News," in 2018 Cyber Resilience Conference (CRC), 2018, pp. 1–4.

[8] S. Zhou, Z. Long, L. Tan, and H. Guo, "Automatic Identification of Indicators of Compromise using Neural-Based Sequence Labelling," arXiv:1810.10156 [cs], 2018.

[9] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and X. Niu, "TTPDrill: Automatic and Accurate Extraction of Threat Actions from Unstructured Text of CTI Sources," 2017, pp. 103–115

[10] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the IOC Game: Toward Automatic Discovery and Analysis of OpenSource Cyber Threat Intelligence," in ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 755–766.

[11] Ahmed, Md Faisal; Anwar, Md. Tawhid; Tanvir, Sifat; Saha, Ramkrishna; Shoumo, Syed Zamil Hasan; Hossain, Md Sabbir; Rasel, Annajiat Alim (2021), "Cybersecurity News Article Dataset", Mendeley Data, V1, doi: 10.17632/n7ntwwrtn5.1Writer's Handbook. Mill Valley, CA: University Science, 1989.