Cyber Threat Intelligence Discovery using Machine Learning from the Dark Web

Azene Zenebe
Bowie State University[1]

[1]Department of Management Information Systems (MIS), Bowie State University, MD

ABSTRACT

Cyber threat intelligence (CTI) is an actionable information or insight an organization uses to understand potential vulnerabilities it does have and threats it is facing. One important CTI for proactive cyber defense is exploit type with possible values system, web, network, website or Mobile. This study compares the performance of machine learning algorithms in predicating exploit types using form posts in the dark web, which is a semi- structured dataset collected from dark web. The study uses the CRISP data science approach. The results of the study show that machine learning algorithms which are function-based including support vector machine and deep-learning using artificial neural network are more accurate than those algorithms which are based on tree including Random Forest and Decision-Tree for CTI discovery from semi-structured dataset. Future research will include the use of high-performance computing and advanced deep-learning algorithms.

Keywords:
Cyber Security, Cyber Threat Modeling, Machine Learning, Deep Learning, Dark Web

INTRODUCTION

The Web has three types: the Surface Web, the Deep Web and the Dark Web. The Surface Web contains all indexed web pages on the Internet and accessible to the public. The Deep Web is not indexed or searchable by common search engines and only users with access rights can login and access them. The Dark Web is a subset of the Deep Web and requires special software such as Tor and trusted connections. It allows user anonymity and is known for a place where malicious actors meet, communicate, and share unethical hacking techniques, tools and information for exploits or attacks. The amount of data and information on the dark web is big data and estimated to be about 5% of all content on the Web, which is about 100 of zettabytes, where a zettabyte is about a trillion gigabytes (Saleem, Islam, & Kabir, 2022).

Cyber threat intelligence (CTI) is an actionable information or insights on current and possible attacks that can be used to alert, protect, and counteract cyber threats. On the Dark Web, CTI discovery techniques include monitoring forums, marketplaces, websites, and traffic, as well as the use of honeypots (Saleem, Islam & Kabir, 2022). Machine learning is training machines using data and algorithm to build a model for prediction. Cyber threat intelligence discovery using big data and machine learning provides a proactive approach to identify cyber threats early before they become problems that bring high risks for cyber-attacks to organization. This paper focuses on Cyber Threat Intelligence Discovery using Big Data and Supervised Machine Learning from the Dark Web. The paper has five sections. Section 2 presents the background, followed by the methodology in Section 3. Section 4 presents the results and discussion, and Section 5 presents the conclusion.

BACKGROUND

As the popularity and adoption rate of big data and machine learning increases, its applications in various sectors of business and industry continues to expand beyond the scope it traditionally defined. One of its more extensive applications is in the field of cybersecurity. The cyber security industry has evolved rapidly in the last few years, owing largely to the incidence and proliferation of online threats, risks, and security saboteurs across all spheres of the web.

Machine learning has proven to have numerous cyber security applications, including the improvement of available antivirus software, proactive identification of cyber vulnerabilities, and fighting machine learning-dependent cybercrime. Cybersecurity mainly safeguards the hardware, software and data from external threats as well as prevent unauthorized access to their databases and systems.

.

Some of the many kinds of cyber-attacks include phishing, spear phishing, drive-by attack, password attack, denial of service, and ransoms.
Threat intelligence, or cyber threat intelligence (CTI), is actionable information or insight an organization or individual uses to understand the threats that have, will, or are currently targeting the organization. This insight can be used to anticipate, avert, and pinpoint where the cyber threat is coming from and who is trying to take hostage of resources that are needed to sustain business operations and support decision making and strategic planning. CTI is important because the information collected can be used to understand and protect people and organization from current and future cyberattacks. Hacker forum posts are blogs that hackers use to post instructions and tools that they have used to infiltrate systems, databases and networks. By analyzing these posts, we can see how hackers think and behave, and giving us enough information to predict and prevent the cyber-attack from happening.
Evaluating hacker forum post on dark web can give valuable insights. Tavabi and his colleagues (Tavabi et al., 2019) studied the characteristics of 80 forums on the dark web for a year using the Latent Dirichlet Allocation (LDA) technique by looking at their most significant words. They found that the topics are categorized as either talking about Security, Gaming,
Vending and others. Clustering of the forums also led to forums talking about cyber hacking (Cluster 1), dark web marketplace (cluster 2), hacking PlayStation (Cluster 3), and white hat hacking (Cluster 4). Tavabi and his colleagues (Tavabi, et al., 2018) developed an efficient algorithm that uses ANN to predicate vulnerability exploit using data from dark web. This study is unique as it uses visual analytics to analyze forums from dark web to describe in depth the players on the dark web, and supervised machine learning or classification techniques to build predicative models for determining the exploit types from the forum posts on the dark web.

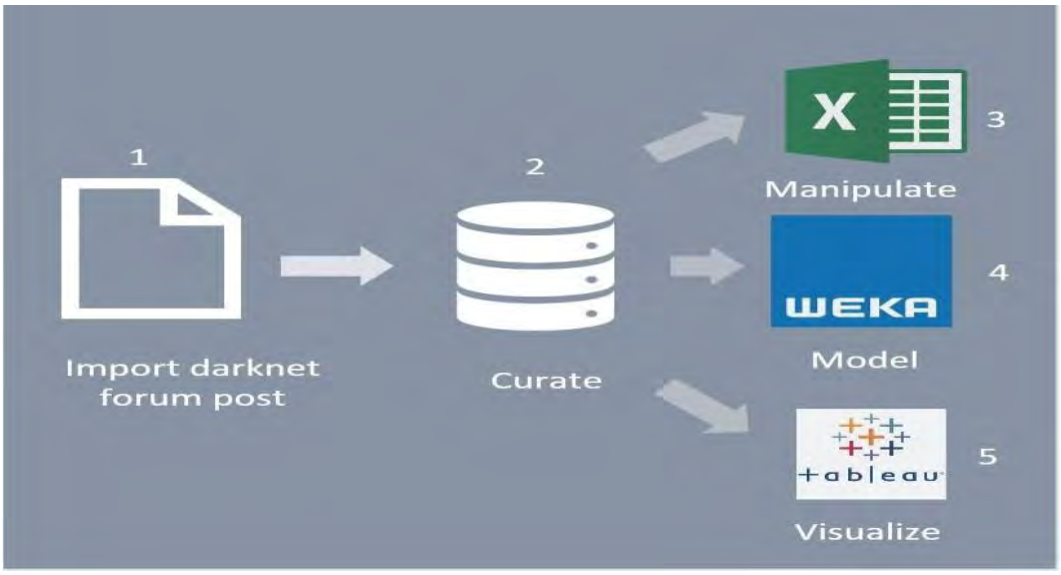
METHODOLOGY

This research attempts to answer the question how accurate are machine learning (ML) algorithms using semi-structure data to discover Cyber Threat Intelligence from the dark web to improve cyber defense?
Using the CRISP (Cross Industry Standard Process) for data mining that serves as base for data science process, and forum posts data from dark web, the study discovers patterns using visual analytics, and builds models using supervised machine learning algorithms: the different algorithms that are tree-based and function-based in determining type of exploits talked about and planned.
The data science process followed includes the problem understanding, data understanding, data preparation, modeling, evaluation, and deployment of model

.

created. Figure 1 shows the steps taken in this research. In step 1, hacker forums data from the dark web (Alsayra, 2011) was used. In step 2, the data is gathered, curated, and cleaned. In step 3, data is converted into a usable form with Excel. Step 4, models are built with decision tree (J48), random forest, Naïve Bayes, support vector and artificial neural network machine learning algorithms using the WEKA machine learning software tool. Furthermore, in step 5, using Tableau, visuals are created to describe the dataset and to characterize the players in the different forums and determine patterns.

Figure 1.
Data Curation, Patterns and Models Discovery Process



Dataset

The dataset used in the research was provided by University of Arizona's Artificial Intelligence Lab (Alsayra, 2011). The dataset contained forum posts from the dark web that had attachment. The data also comes from a variety of forums and different languages. The datasets were in MySQL database, and then exported into an HTML file and converted into a CSV file as well as ARFF (Attribute-Relation File Format) file format needed by WEKA, the machine learning software. The dataset was cleaned for duplicates using the duplicate function of excel. The time frame for the dataset was 2003 to 2016. A sample of the rows from the dataset,

.

i.e., examples of forum posts, are presented in Table 1. The data structure or attributes of a dark web forum post are:

- ID -the numerical value attached to each post
- postID - the unique number for each post
- forumName - the name of the forum collected from
- authorName - the name of the author of the post
- threadTitle - Name of the thread in which the post is in
- postdatetime - Date and time that the post was submitted
- attachmentName - Name of the attachment
- flatContent - Text written by the author in the forum post
- URL - The link to the post
- language - Programming language, not valid in this dataset
- exploitType - Aspect of computer/cyberspace the attachment is exploiting
- assetType - Type of asset, default value of attachment
- assetName - the types of tools used
- attachmentURL - The link to the attachment
- contentwithhtmltag - Text written by the author in the forum post with html tags for formatting

Table 1.
Examples of Form Posts

postID	forum Name	author Name	threadTitle	flatContent	exploitType
2179	Opensc		Avi webcam & screencap	This component will write an AVI file containing a single video stream of any number of TBitmaps etc., plus a single audio stream with one WAV file.get this shit now , example includedgreets	System
7283	Opensc	LttCoder	Zombie's Mailer	Keywords: Mail notification, send e-mail Author:Zombi e Web: http://fre enet.am/~zombie ICQ:51962597 Description: This little program sends mail to people using a specified SMTP server.	Network

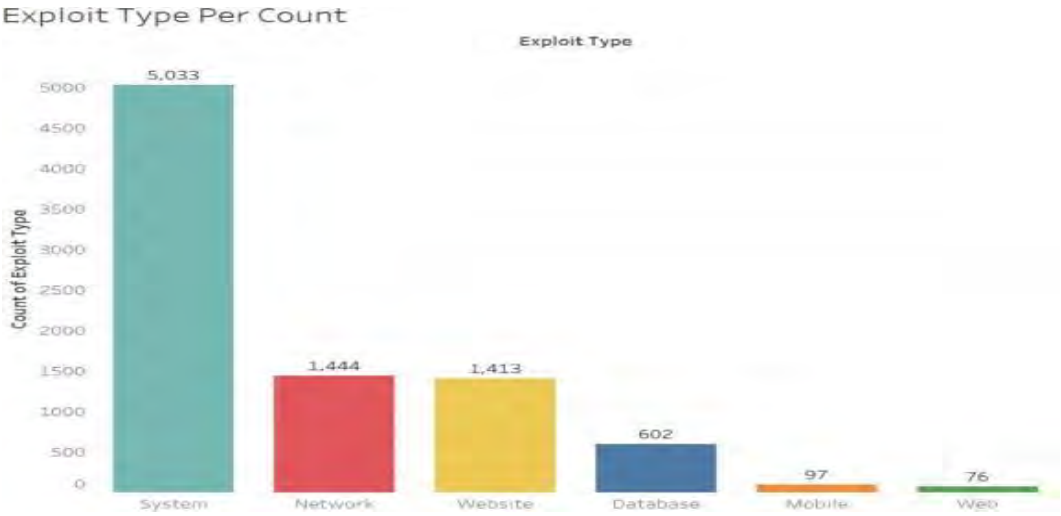
.

RESULTS AND DISCUSSION

Visual and Descriptive Analytics

Various visuals are created using Tableau. The visual in Figure 2 presents the count of posts by exploit type. It shows that most posts were about system exploits and the least posts was about the Web exploits. This shows that the dataset has imbalanced classes problem where some classes have more instances than others. Figure 3 shows the count of the six exploit types in the posts over the years. Using this line graph, you can see that over the years posts for the different exploits are growing with the peaks being between year 2010 to 2013. System, Website and Network exploit types are the top three with highest number of posts throughout 2009 to 2016.

Figure 2.
Count of Exploit Types



In figure 4 the graph represents the frequency of posts by different Authors on the Forums by exploit types: databases, mobile devices, networks, systems, the web, and websites. This visual helps us recognize the authors or potential hackers that post exploits the most and what kind of exploits they are likely to post. With this insight one can determine potential type of attacks they may encounter. “Unline” is the author that posts the most exploits and is mostly known for posting system and network exploits. “Majdi-Hidden” is the author that posts the least of these authors

.

and is mostly known for posting website exploits. “Und3rg0und” has the highest posts on Network exploit. These authors can be further investigated and monitored.

Figure 3.
Exploit posted over the years by Exploit Types

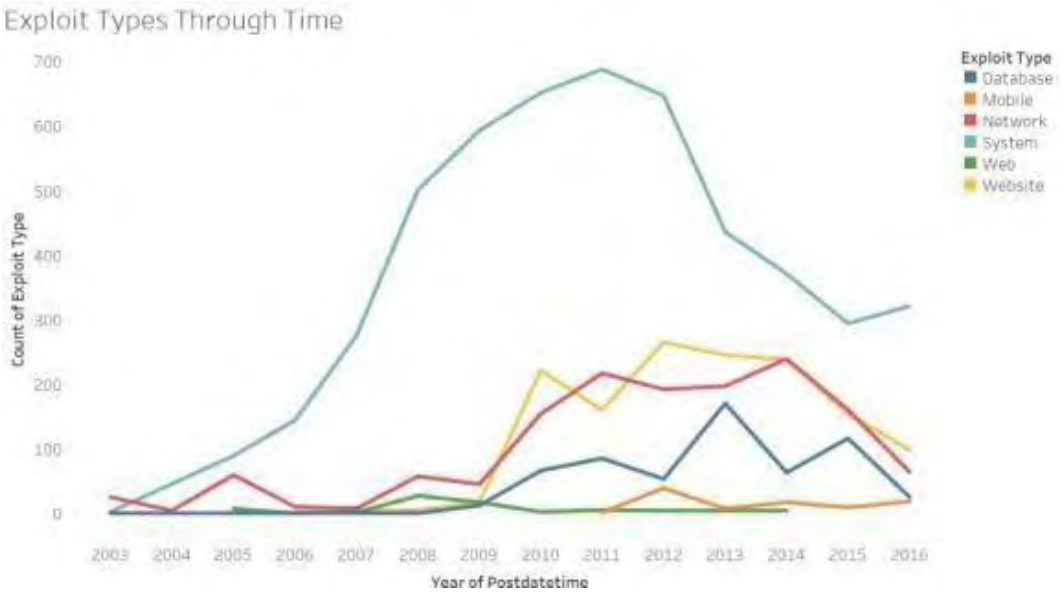
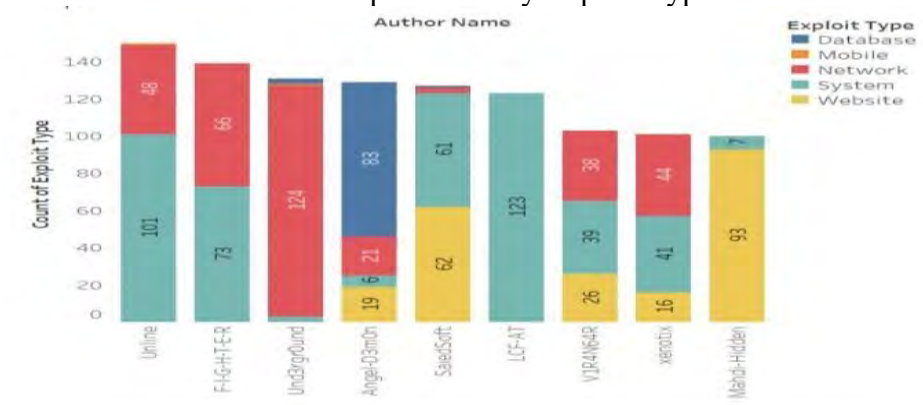


Figure 4.
Top Authors by Exploit Types

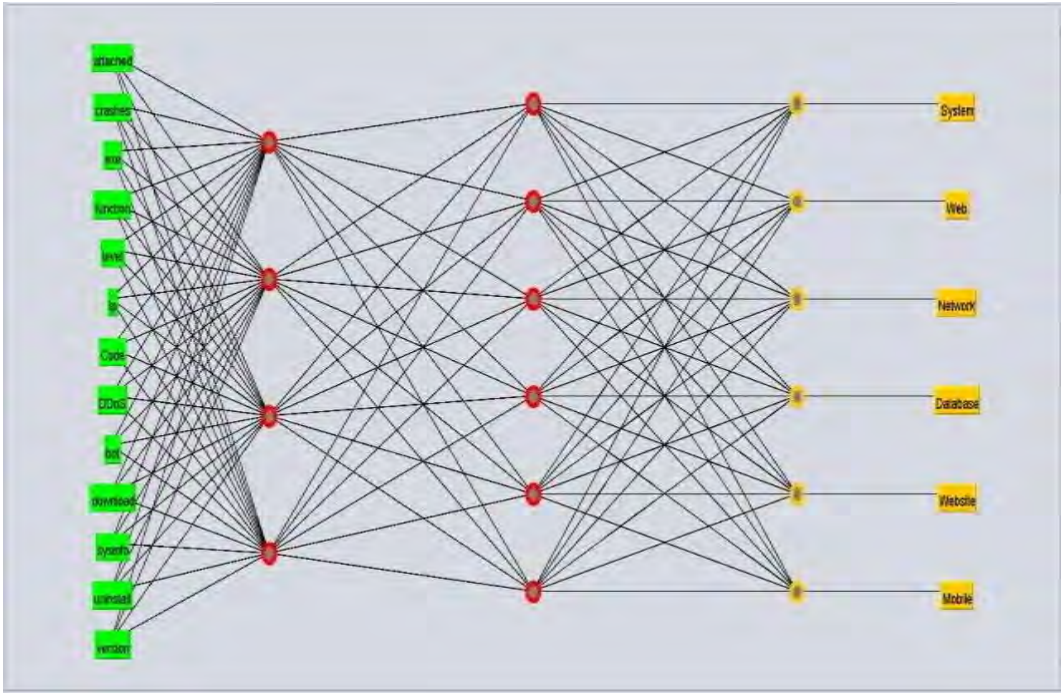


.

Prediction Models using Machine Learning

After pre-processing the input field flatContent using text to word vectors, machine learning using the different algorithms performed using the 10-flod cross validation techniques for training and evaluation. For measuring the performance of the algorithms, accuracy and MCC are computed. Accuracy measures percentage of correct predication. Matthews correlation coefficient (MCC) is used in machine learning as a measure of the quality of the classification. The MCC takes values between -1 and 1, and a closer value to 1 indicates close to perfect agreement between the actual and the predication.
Figure 5 present the deep learning model (Shrestha & Mahmood, 2019; Chen and et al, 2020) where we use two hidden, input and output layers. The results of the evaluation of the different algorithms are presented in Table 2 and Table 3.

Figure 5.
Model using Deep Learning (ANN) using text extracted from dark web



Input Hidden Layer 1 Hidden Layer 2 Output

.

Table 2.
Performance of Different Machine Learning Algorithms

	Naive Bayes		Random Forest		Decision Tree (J48)	
Exploit Type	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
System	78.3	0.480	96.1	0.681	91.9	0.631
Web	60.3	0.157	0.0	0.001	20.5	0.295
Network	39.5	0.375	51.7	0.562	50.0	0.465
Database	65.4	0.584	63.2	0.728	57.9	0.641
Website	67.5	0.515	72.1	0.731	64.6	0.639
Mobile	52.0	0.482	40.0	0.601	58.0	0.723
Weighted Average	64.7	0.468	83.1	0.66	79.4	0.603

Table 3.
Performance of Function-based Machine Learning Algorithms

	Support Vector Machine (SVM)		Artificial Neural Network (ANN)	
Exploit Type	Accuracy (%)	MCC	Accuracy (%)	MCC
System	95.6	0.698	99.9	0.069
Web	72.8	0.774	0.00	
Network	58.4	0.633	0.00	
Database	77.3	0.832	0.00	
Website	70.6	0.737	0.023	0.13
Mobile	60.0	0.158	0.00	
W. Average	85.0	0.695	66.4	

The Random Forest algorithm outperforms the Naïve Baye and Decision Tree (J48 in Weka) algorithms. In all algorithms the classification accuracies for system type

.

of exploit are high, 78% and above. Furthermore, when data has less cases for the exploit type Web, other than the Naïve Bayes algorithm, all perform very low but with low MCC.

The SMV outperforms the ANN algorithm both in higher the accuracy and MCC values. In both algorithms the classification accuracies for system type of exploit are very high, 95.6% and 99.9%. Furthermore, when data has less cases for the exploit type Web, SMV accuracy
(73%) is the highest with high MCC value of 0.8. Other than the class Mobile Exploit Type, SVM have a high-quality predication for the model as well as for all classes as their MCC values are close to 1.

CONCLUSION AND FUTURE RESEARCH

The focus of this research was to access data from the dark web and determine how effective machine learning (ML) is in discovering Cyber Threat Intelligence (CTI) to be used to improve cyber defense. The SVM algorithm had the highest overall accuracy rate of 85.0%. The SVM model also produced the highest accuracy rates for Network, Database, Website, and Mobile exploit type classifications. When using the ANN, it contained the best results for System exploits while Naive Bayes had the best for Website exploit. Also, the Random Forest model does not have any single high accuracy rate but had the 2nd highest Average Accuracy Rate (83.1%).

Overall, the machine learning algorithms are very effective in predicting exploit types that could lead to potentially cyberattacks using dark web forum posts. Overall, the functionbased models' performance is better than the three-based models. Future research will include the use of high-performance computing and advanced deep-learning algorithms.

Acknowledgement: This research is partially supported by the NSF Grant # 1818669, and NSF Grant # 1757924. Furthermore, undergraduate students Asmamaw Mersha, Maurice Bugg, Jason Smith, and Cion Sandidge who were in summer research training program provided support in data curations and analysis.

.

REFERENCES

Alsayra. (2011, April 5). *AZSecure Hacker Assets Portal*. Retrieved from Intelligence and Security Informatics Data Sets: https://www.azsecure-data.org/hacker- assetsportal.html

Chen, M.-Y., Chiang, H.-S., Lughofer, E., & Egrioglu, E. (2020). Deep learning: Emerging trends, applications and research challenges. *Soft Computing*, *24*(11), 7835–7838. https://doi.org/10.1007/s00500-020-04939-z

N. Tavabi, P. G. (2018). DarkEmbed: Exploit Prediction with Neural Language Models. *AAAI , page 7849-7854. AAAI Press, (2018)* (pp. 7849 - 7854). AAAI Press.

Nazgol Tavabi, N. B. (2019). Characterizing activity on the deep and dark web. *The 2019 World Wide Web Conference* (pp. 206-213). San Francisco, CA: ACM.

Saleem, J., Islam, R., & Kabir, M. A. (2022). The Anonymity of the Dark Web: A Survey. *IEEE Access*, 1-33.

Shrestha, A., & Mahmood, A. (2019). Review of Deep Learning Algorithms and architectures. *IEEE Access*, *7*, 53040–53065. https://doi.org/10.1109/access.2019.2912200

.