EPiC Series in Computing

Volume 64, 2019, Pages 174–183

Proceedings of 28th International Conference
on Software Engineering and Data Engineering

Cyber Threat Discovery from Dark Web

Azene Zenebe[1], Mufaro Shumba[2], Andrei Carillo[1], and Sofia Cuenca[3]

[1] Bowie State University, Bowie, Maryland
[2]University of Maryland College Park, College Park, Maryland
[3]Farmingdale State College, Farmingdale, NY
azenebe@bowiestate.edu, mufaroshumba@gmail.com,
carillo.andreiwilson@gmail.com, cuens@farmingdale.edu

Abstract

In the darknet, hackers are constantly sharing information with each other and learning from each other. These conversations in online forums for example can contain data that may help assist in the discovery of cyber threat intelligence. Cyber Threat Intelligence (CTI) is information or knowledge about threats that can help prevent security breaches in cyberspace. In addition, monitoring and analysis of this data manually is challenging because forum posts and other data on the darknet are high in volume and unstructured. This paper uses descriptive analytics and predicative analytics using machine learning on forum posts dataset from darknet to discover valuable cyber threat intelligence. The IBM Watson Analytics and WEKA machine learning tool were used. Watson Analytics showed trends and relationships in the data. WEKA provided machine learning models to classify the type of exploits targeted by hackers from the form posts. The results showed that Crypter, Password cracker and RATs (Remote Administration Tools), buffer overflow exploit tools, and Keylogger system exploits tools were the most common in the darknet and that there are influential authors who are frequent in the forums. In addition, machine learning helps build classifiers for exploit types. The Random Forest classifier provided a higher accuracy than the Random Tree and Naïve Bayes classifiers. Therefore, analyzing darknet forum posts can provide actionable information as well as machine learning is effective in building classifiers for prediction of exploit types. Predicting exploit types as well as knowing patterns and trends on hackers' plan helps defend the cyberspace proactively.

1 Introduction

Current techniques for dealing with cyber breaches are reactive, meaning once a breach occurs then cyber professionals take actions. This is no longer acceptable because breaches are only detected on average after about 6 months and only 10% of breaches are detected in the first 24 hours [1][2]. In that amount of a time a lot of damage can be done to an entity. Secrets and information can be leaked, and damage can be done to the entity's system. The adaptation of cyber threat intelligence is crucial in keeping ahead of attackers. Cyber threat intelligence is any actionable information, insights, and knowledge about threats that can help in preventing security breaches in cyberspaces. The discovery of cyber threat intelligence will help keep security measures to be proactive. Threats may be recognized before they become a problem.

When dealing with hackers it is important to note that hackers spend a lot of time-sharing information on online communities. One of these communities is the darknet. The darknet is a network with restricted access where people can stay anonymous for legal and illegal reasons [3].

There are two different types of forums. There are the WhiteHat forums which are easy to access and useful in teaching people in becoming hackers, it has tutorials and it is more accessible since many are not on the Darknet. On the other hand, there are the BlackHat forums that are illegal, have encrypted URLs and sell and share the codes for hacking as well as stolen data. BlackHat forums are harder to access since they are on the darknet and have encrypted URLs. The hacker communities in the darknet are usually in the form of forums. These forums may have restricted access. They may have encrypted URLs and an invite and/or money may be required to be able to register. In these forums, hackers share malware and other tools that can be used to exploit a computer and network systems. Analyzing the contents of forum posts can give an insight on understanding hackers' identity, motivations, strategy, targets, tactics, etc., which in turn helps prevent next cyberattacks.

Every day, humans create 2.5 quintillion bytes of data [4]. Unstructured data including form posts are part of this big data. This large amount of data can assist the goal of preventing security breaches in cyberspaces in proactive way. This research uses analytics to extract actionable information and machine learning to build models to predicate exploit types. An exploit is a software or codes that finds and takes advantages of security weakness to breach or attack systems, networks and hardware. This paper is organized into several sections. Section 2 presents literature review. Section 3 presents methodology, followed by results and discussion in Section 4. Section 5 presents conclusion and future research.

2 Literature Review

The literature review covers important sub-areas: the cyber threat intelligence and its importance, understanding hacker forums, and machine learning models that are useful in analyzing unstructured text data.

2.1 Cyber Threat Intelligence

The world is becoming more and more connected, automated and run by computers and smart devices forming the cyberspace. Therefore, it is crucial to monitor, collect and analyze data related to security threats and vulnerabilities in cyberspace. The understanding and knowledge of these threats is called cyber threat intelligence (CTI). CTI is still in its beginning of development and lacks structure and a common understanding throughout the community. In addition, security professionals are beginning to be one step behind of attackers. Attackers are constantly adapting their tools to go around the security measures. The adaptation of cyber threat intelligence is important in keeping ahead of attackers. In cyber security, monitoring and data collection is the first step in efforts to defend from attacks. This step generates big data associated with security events. Currently cybersecurity teams have had challenges in making use of security related large data and are looking towards artificial intelligence/machine learning to help them parse and analysis this data and discover insightful information about the threats and possible attacks [5].

Mavroeidis and Bromander [5] presented two models for detecting cyber threats: Detection Maturity Level Model (DML) and Cyber Threat Intelligence model (CTI). The CIT model focuses on representation of the types of information about threats and potential attacks. The first element of the

CTI model is the identity of a threat actor such as a person, an organization, or a nation state to determine any strategy, tactics, techniques, and procedures expected to be used [5]. The other elements are motivations, goals, strategy, TTPs (Tactics, techniques, and procedures), indicators of compromise (IOCs), target, and courses of action [5].

2.2 Hacker Forum Analysis

Most hacker forums are found on the Darknet. The Darknet is an area of the internet where everyday users are not able to access [3]. The sites that are in the darknet do not show up on search engine like Google or Bing. In order to get into the Darknet, the user would need to use The Onion Router (TOR) [3]. For more security the group would use a Virtual Private Network (VPN) to spoof the IP address. Most websites URLs in the Darknet start with an encrypted address and end in dot onion.

Forum posts are written in natural language, so it is not easy for a computer to analyze out sensible information without human assistance. Analyzing the contents of forum posts can give an insight on understanding hackers' identity, motivations, strategy, targets, tactics, etc. that can help security professionals prevent attacks. One method of analyzing forum posts is sentiment analysis which is a process of determining the positive or negative tone of a piece of text [6]. Running forum posts through sentiment analysis can provide an extra feature in a data set and show hackers who constantly show negativity through their posts. In addition, keyword searches can be done to analyze a forum post. Analysis of keywords such as virus, worms, and malware can help detect a potential threat [1] [6]. Another approach for CTI discovery is to use analytics and machine learning.

2.3 Machine Learning for CTI discovery

Machine learning is the process of teaching machines to see patterns in data. There are 2 main types of machine learning algorithms: supervised and unsupervised learning. Supervised learning algorithms are applied to labeled datasets and are used to predict or classify the data. Unsupervised learning algorithms are unlabeled datasets that take the data and attempts to cluster them according to similarities. There are also instances where a mix of supervised learning and unsupervised learning known as semi-supervised learning which is mainly utilized when the data is not completely labeled.

The machine learning algorithms used in past research include decision tree, support vector machine and naïve Bayesian, and were successful in analyzing textual dataset [7, 8, 9]. Furthermore, Convolutional Neural Networks is up there with these models, and a natural language processing approach, analyzing character n-grams vs word n-grams is an interesting point to explore [10].

3 Methodology

This section presents the method and procedures followed in the research.

3.1 Research Framework

Figure 1 presents a general framework for CTI discovery system in terms of the input, output and process. The input could be direct data that come to the system such as network traffic as packets, event data generated by OS, network and security devices, and/or indirect data on Facebook, forums, twitter, etc. This data is then run through a process of cleaning and transformation followed by summarization and predication. The expected output from the process is to find information or insight about cyber threat that helps to prevent cyberattacks from threat such as new virus or system compromise that could lead to data theft. The output can include the profile of potential actors or hackers, motivation and motivators of the potential and actual attacks, techniques and strategies for attacks, and the indicators of attacks and breaches.

.

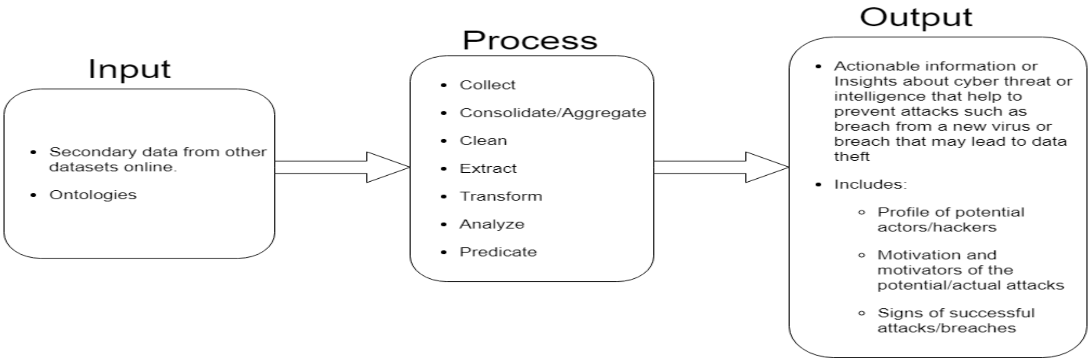
Input
Secondary data from other datasets online.
Ontologies
Process
Collect
Consolidate/Aggregate
Clean
Extract
Transform
Analyze
Predicate
Output
Actionable information or Insights about cyber threat or intelligence that help to prevent attacks such as breach from a new virus or breach that may lead to data theft
Includes:
Profile of potential actors/hackers
Motivation and motivators of the potential/actual attacks
Signs of successful attacks/breaches

Figure 1: IPO Chart

This research uses the CRISP data mining approach which is an iterative process with multiple stages as presented in Figure 2.

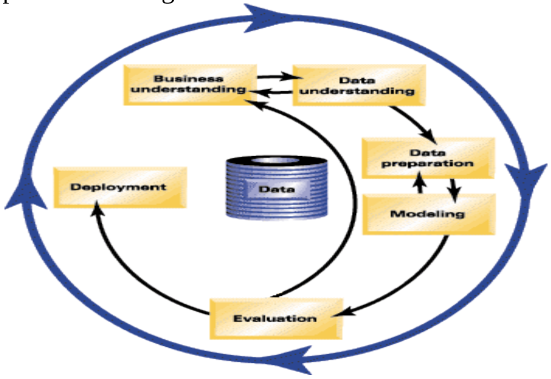

Figure 2: CRISP DM Flowchart (IBM Knowledge center)

3.2 Dataset

The dataset used in the research was provided by University of Arizona's Artificial Intelligence Lab [11]. The dataset contained forum posts from the darknet that had attachments. The data was collected using offline explorer and java-based crawlers and parsers. Offline explorer is a program used to download and collect bulk web content from websites like Facebook, Twitter, Instagram and other social networking sites [12]. The data also comes from a variety of forums and different languages. The datasets were SQL files and opened with MySQL. They were then exported into an HTML file and converted into a CSV file as well as ARFF file format needed by WEKA, the machine learning software.

The dataset was cleaned for duplicates. This was done through the duplicate function of excel. The attachments data set were checked for duplicates based on post ID. This cut down the attachments data set from a 14865 to 6069 cleaned, extracted and transformed data. The time frame for the dataset was 2003 to 2016. Examples of forum posts are presented in Table 1. The data structure or attributes of a forum post are:

- postID - Unique number for each post
- forumName - Name of the forum collected from
- authorName - 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 designed to exploit
- assetType - Type of asset, default value of attachment
- assetName - Name of asset, default value of null
- attachmentURL - The link to the attachment
- Contentwithhtmltag - Text written by the author in the forum post with html tags

postID	flatContent	exploitType
7283	Keywords: Mail notification send e-mail Author:Zombie Web:http://freenet.am/~zombie ICQ:51962597Description: This little program sends mail to people using a specified SMTP server.	Network
1282	Hello like i promised i changed the code in order to be able to use the progress bar on file upload. I think id did not work because the OnFileRead procedure was called every time the socked read something so if the file didn t get through all at once then it wasn t stored correctly (the file received was different in size from the original sent file). In order to correct this i eliminated the OnFileReceived procedure and created another one called StartWaiting. This new procedure is called when a new connection is established and it has an infinite loop that reads everything received on the socket and handles the request. The only request handling i changed in this part was the . I created a new procedure that receives the file sent. All those changes were made on the server part. In the client i added a #10 character on each command sent in order to be able to use ServerSocket1.Socket.Connections[0].receiveText on the server. Example: ClientSocket1.Socket.SendText(+#10); I also changed the procedure used to send the file. I really hope you keep using the changes i made since i think they ll make it easier for you to add more functionalities in the future and is less dependable from sockets events handled by delphi. I ll upload the source with the changes but i really recommend you take those changes and implement them on your own source code in case i left some flags for debugging by mistake. (in the source i disabled the register part to avoid infecting myself) Let me know if it helps Unknown.	System

Table 1: Examples of Forum Posts

3.3 Analysis

The research uses Watson Analytics for descriptive and visual analytics. Furthermore, the research uses several machine learning algorithms for CTI discovery. The WEKA software tool is used for machine learning. We use an unsupervised filter to take each string in flatContent and turn it into a word vector. The feature used for classification was the flatContent that contains the textual content of the posts in the forum, and we used exploit type as the class label. Naïve Bayes, Random Three and Random Forest machine learning algorithms are used. The 10-fold cross validation technique is used to evaluate machine learning models. Accuracy, Matthews Correlation Coefficient (MCC) and ROC (*receiver operating characteristic*) area are the performance measures. The MCC is a measure of how random the classification is, -1 would indicate a completely random selection and 1 would indicates a completely correct classifier.

4 Results and Discussion

This section presents the results obtained from and discussion on descriptive and visual analytics and predicative analytics which are classification models.

4.1 Descriptive Analytics Results

Figure 3 shows the top 10 authors who frequently posted on all three forums of the data set: Opensc, Ashiyane and Futs4you. Half of the top authors on the forums are from the forum Ashiyane. Through this graph we see some key and influential members of these hacker communities. The number of posts that the top authors had ranged from 49 to 123 postings.

Key members of forums can also be seen through analysis of this data set. This information shows that there are authors that have more influence and presence in a hacker forum. For CTI, knowing and monitoring the key members of a hacker forum can be helpful in keeping the security in cyberspace. These key members have some leadership or influence in the forums. The knowledge of what these members are up to can keep security professionals aware of any emerging problem or threats.

Figure 4 shows the number of posts per exploit type. The number of posts per exploit types ranges from 50 to 3944 with mobile being the lowest and system being the highest. System made up about 67% of exploits in the dataset, in perspective, the next highest exploit type is network with only 15%. Web and mobile were the least of the exploit types.

There is an abundant amount of system exploit posts throughout all 3 forums. A big proportion of the dataset contains system exploit posts. Hackers seem to create more tools for exploiting systems – mainly operating systems such as Windows, Unix/Linux, Mac, ios and android. With hackers seemingly giving a bigger importance to system exploits, focusing more on the system may keep defense measures proactive.

Looking further into the system exploit tools shows the most common tools in the dataset that shared on the forums in the darknet. Crypter, which is a software that creates malware that can bypass security controls and get installed in to systems, was the most common tool and buffer and keylogger tied as the 4th most common. The other top two most common tools are password cracker and RATs (Remote Administration Tools).

The line graph in Figure 5 shows the number of posts by exploit type from year 2003 to 2016. Posts on mobile, database and web exploits stayed consistently low. System exploits posts overwhelmingly increased from 2005 to 2009. The system posts then stayed consistent for two years until a steep drop from 2011 to 2015. There was an increase of network, website and database posts from 2009 to 2010.

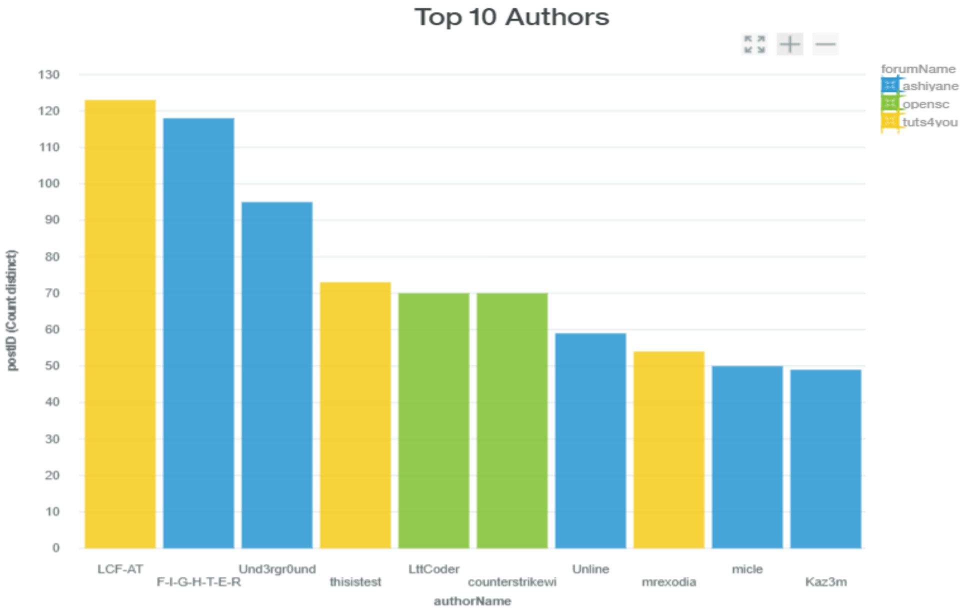

Figure 3. Top 10 Authors along with associated Forums

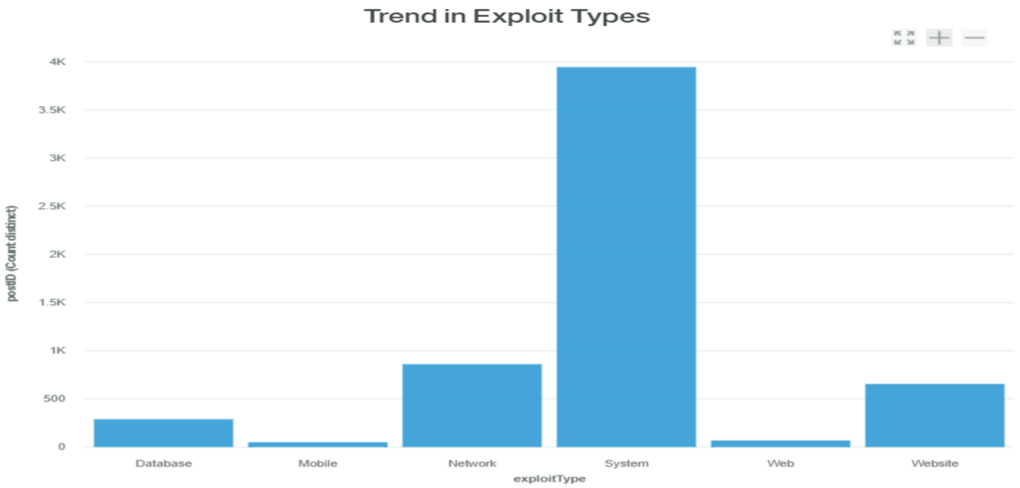

Figure 4. Number of Posts per Exploit Type

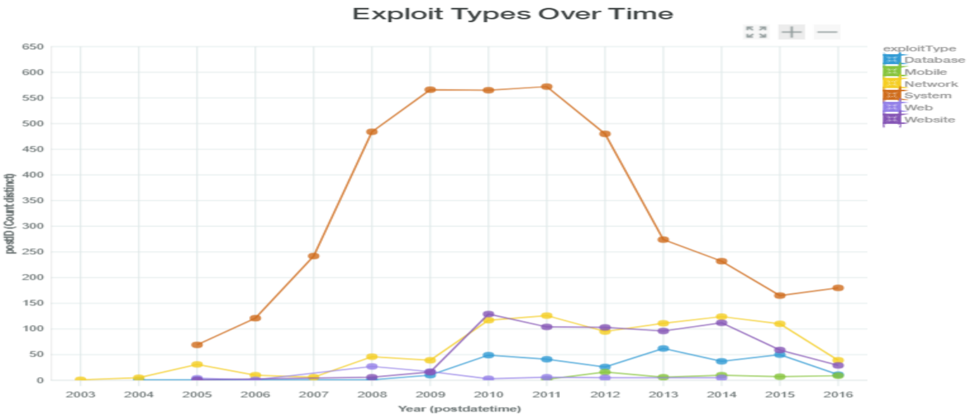

Figure *5*. Exploit Types Over Time by Year

4.2 Machine Learning Results

The classification problem is to identify the type of exploit based upon content of a forum's post. An exploit is a software or codes that finds and takes advantages of security weakness to breach or attack systems, networks and hardware. The six class labels for the class attribute exploit type are database, mobile, network, system, web and Website. The summary of the results for the three classification models presented in Table 2 and are:

Model 1: Naive Bayes - A naive Bayes model works on and applies knowledge of probability. For a Naive Bayes model, it is assumed that the features are independent of each other which makes the model run relatively fast. The correctly classified instances, the accuracy, for this model is 52.81% and the exploit type with the highest accuracy (66%) was the web exploits.

Model 2: Random Tree: - A Random Tree model randomizes the feature at each node of the tree and does not prune the tree and is mainly used for classification. The Random Tree model produced accurate model with about 67.47% accuracy. The model is a decision tree with each node being random and used it to classify each forum post. Most accurately predicated exploit type was System, with 84.26% accuracy followed by Website with 43% and Network with 34.46%.

Model 3: Random Forest - A Random Forest model is built out of a large amount of decision trees and used as a method of classification. The forest notation comes from the idea of multiple trees. Its accuracy for the model was the highest with 78.27% overall accuracy. When broken down by class, System exploits had 97.87% accuracy followed by Website at 56.04%. This model also had a relatively high MCC rating for System, Network, Database, and Website.

While the Naive Bayes model is not extremely accurate, it does offer a starting point. The model could be used to automatically analyze forum posts for exploit type without needing to open/download the possibly harmful attached file. It also provides a technique that will cut down on time needed to specify the type of threat within the forum post. While System is the most accurately identified exploit type, it is also the exploit type with the most instances in the dataset.

The use of a decision tree saw a significant increase in accuracy. The problem with decision trees is that they can easily overfit data to the training data. This means that it knows the training data very well but when exposed to new data it tends to fail. This model is probably running into this issue with such many of exploits being miscategorized as system exploits.

While the random tree and random forest models had the highest accuracy, the issue of overfitting came up. Upon further analysis of the dataset, it was made clear that there are not enough data points for each exploit type. The machine will overly classify exploit types as system exploits due to their magnitude in the attachments dataset. Overall, the highest performing model was random forest.

	Naive Bayes			Random Tree			Random Forest		
Exploit Type	Accuracy	MCC	ROC Area	Accuracy	MCC	ROC Area	Accuracy	MCC	ROC Area
System	0.564	0.441	0.79	0.843	0.466	0.718	0.979	0.565	0.916
Web	0.658	0.145	0.851	0.055	0.061	0.524	0.00	0.001	0.883
Network	0.281	0.239	0.627	0.345	0.248	0.607	0.368	0.482	0.869
Database	0.612	0.407	0.894	0.293	0.248	0.626	0.319	0.529	0.969
Website	0.61	0.453	0.901	0.430	0.375	0.683	0.564	0.616	0.958
Mobile	0.62	0.202	0.911	0.120	0.165	0.558	0.02	0.141	0.953
Avg	**0.528**	**0.403**	**0.783**	**0.675**	**0.403**	**0.688**	**0.783**	**0.545**	**0.916**

Table 2: Performance of Models

5 Conclusion and Future Research

The focus of this research was to discover cyber threat intelligence from hacker forum posts in the darknet. Using descriptive analytics, the research has helped to find trends in exploit type and key actors in darknet forums. Furthermore, the deeper analysis on system exploits showed that Crypter, password cracker and RATs (Remote Administration Tools), buffer overflow exploit tools, and keylogger tools are most common shared tools by hackers. The knowledge of what tools are most common will help security professionals know which tools to look out for and optimize accordingly. The small number but increasing web and mobile exploits also shows that hackers may still be developing tools in these fields. A close eye on these new types of exploits may be useful in security.

The second phase of research aimed at building morels for predicting exploit type, given text in the post associated with the attachment. The Random Forest classifier provided a higher accuracy than the Random Tree and Naïve Bayes classifiers. Therefore, this research has shown that relevant, timely, and actionable cyber threat intelligence can be extracted from these hacker forums as well as exploit types classifiers can be build using machine learning algorithms.

Future research will focus on specifics of each system exploits to allow for entities to make use of the intelligence. Entities will be able to defend against specific exploit types, in turn making their systems more resilient to these threats. Future research will also include to find more examples of the other exploit types and retrain the classifiers. In addition, other machine learning algorithms such as support vector machine (SVM) and artificial neural network (ANN) can be considered. With our current training dataset, the model is unable to classify pieces of text as not a threat. With further research, the model will learn from pieces of texts that are not threats. This will allow the model to classify both threats and not threats.

References

[1] Benjamin, V., Li, W., Holt, T., & Chen, H. (2015). Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*. doi:10.1109/isi.2015.7165944

[2] Bissell, K., Lasalle, R. M., Van Den Dool, F., & Kennedy-White, J. (2018). Executive Summary Gaining Ground on the Cyber Attacker. *Accenture Security*. Retrieved July 16, 2018.

[3] Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., & Shakarian, P. (2016). Darknet and deepnet mining for proactive cybersecurity threat intelligence. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*. doi:10.1109/isi.2016.7745435

[4] Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. Retrieved July 15, 2018, from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#34f1986760ba

[5] Mavroeidis, V., & Bromander, S. (2017). Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence. *2017 European Intelligence and Security Informatics Conference (EISIC)*. doi:10.1109/eisic.2017.20

[6] Biswas, B., Mukhopadhyay, A., & Gupta, G. (2018). "Leadership in Action: How Top Hackers Behave" A Big-Data Approach with Text-Mining and Sentiment Analysis. *Proceedings of the 51st Hawaii International Conference on System Sciences*. doi:10.24251/hicss.2018.221

[7] Deliu, I., Leichter, C., & Franke, K. (2017). Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks. *2017 IEEE International Conference on Big Data (Big Data),*3648-3656. doi:10.1109/bigdata.2017.8258359

[8] Hearst, M. A. (1998). Support vector machines. *IEEE Intelligent SystemS,*18-28. Retrieved June 25, 2018.

[9] Samtani, S., Chinn, R., & Chen, H. (2015). Exploring hacker assets in underground forums. 2015 IEEE International Conference on Intelligence and Security Informatics (ISI). doi:10.1109/isi.2015.7165935

[10] Johnson, J., & Karpathy, A. (n.d.). Convolutional Neural Networks (CNNs / ConvNets). Retrieved June 25, 2018, from http://cs231n.github.io/convolutional-networks/

[11] Hacker Assets Portal. (n.d.). Retrieved June 26, 2018, from https://www.azsecure-data.org/hacker-assets-portal.html

[12] Systems, M. (n.d.). MetaProducts Systems. Retrieved June 26, 2018, from https://metaproducts.com/products/offline-explorer.

Acknowledgement: This research was supported by National Science Foundation, Award Number:1757924.