Detection of Phishing Attacks with Machine Learning Techniques in Cognitive Security Architecture

Iván Ortiz-Garcés*, Roberto O. Andrade†, and María Cazares‡

1. Facultad de Ingenierías y Ciencias Aplicadas, Universidad de Las Américas,Quito, Ecuador,
2. Facultad de Ingeniería en Sistemas, Escuela Politécnica Nacional, Quito, Ecuador,
3. IDEIAGEOCA Research Group, Universidad Politécnica Salesiana,Quito, Ecuador
Email: *ivan.ortiz@udla.edu.ec, †roberto.andrade@epn.edu.ec, ‡mcazares@ups.edu.ec

***Abstract*—The number of phishing attacks has increased in Latin America, exceeding the operational skills of cybersecurity analysts. The cognitive security application proposes the use of bigdata, machine learning, and data analytics to improve response times in attack detection. This paper presents an investigation about the analysis of anomalous behavior related with phishing web attacks and how machine learning techniques can be an option to face the problem. This analysis is made with the use of an contaminated data sets, and python tools for developing machine learning for detect phishing attacks through of the analysis of URLs to determinate if are good or bad URLs in base of specific characteristics of the URLs, with the goal of provide realtime information for take proactive decisions that minimize the impact of an attack.**

Keywords- Artificial Intelligence, Artificial Neural Networks, Phishing, Cybersecurity, Machine Learning

I. INTRODUCTION

Nowadays data is useful to help people to take decisions for developing tools, buy products, or make electronic transfers; but there is a potentially problem in this context, the security of that the data has integrity, and the data source is no fake. In 2014 yahoo was attacked, hackers used phishing to stole information, over more of 300 millions accounts were affected [1].

Phishing attacks try to get information about someone or something, and is necessary find tools that help to people and specially security analyst by handle this types of attacks. Artificial Intelligence- AI is one of these possible solutions, it can help to detect anomalous behavior, but even better AI can offer new possibilities to protect sensible information, and it is capable to detect anomalous behavior quickly; this is why is so important in new cibersecurity approaches [2].

Cognitive security as the ability to generate cognition for efficient decision making in real time by the human or a computer system based on the perception of cybersecurity that the computer system generates from its environment (situational-awareness) and the knowledge about itself (self-awareness or insights), through the analysis of any type of information (structured or unstructured) using artificial intelligence techniques (data mining, machine learning, natural language processing, and human-computer interaction) and data analysis (bigdata, processes stochastics, game theory) emulating the cognitive process for decision making by security analysis [3].

Our proposal for the automation of incident response related with phishing attacks is based on the importance of establish situation awareness to make the right decision based on an understanding of the aspects of the attack.

The remainder of this work is organized as follows. Sections II shows an theoretical background related with the use of machine learning in cibersecurity. Section III describes the methodology used for detect phishing attacks using machine learning techniques. Section IV presents the results of analyze different parameters of a dataset that contains phishing websites logs. Finally, Section V presents the conclusions with the respective contributions of this work

II. THEORETICAL BACKGROUND

A. Artificial Intelligence

There is a lot of vision about what is AI, some author define AI as:

- ≪The new and exciting effort of make computers learn... machines with brains, in the most literal wide sense≫[4].
- ≪Computational Intelligence is the study of design smart agents≫[5].
- ≪AI... is linked with intelligence behaviors on devices≫[6].

Today A.I becomes popular in the analyze of large amount of data, increasing speed, and capability of process a variety of data. A.I is capable to do task like recognition of patterns more faster than humans beings, which is better for medium, and large companies. A.I can introduce the intelligence of a computer system to the cognitive process develop every day for the human, this increase the success in strategic and operative tasks of the organization. A.I is used to that an system make some cognitive function like solve problems, or even think. AI has been divided in two sections: Weak Systems A.I and Strong Systems A.I.
Weak A.I, is well known as narrow, is a system of Artificial Intelligence design and trained to do a particular task. Personal Assistance like Siri from Apple, Cortana from Windows, Alexa from Amazon are an example of weak A.I.

978-1-7281-5584-5/19/$31.00 ©2019 IEEE
DOI 10.1109/CSCI49370.2019.00071

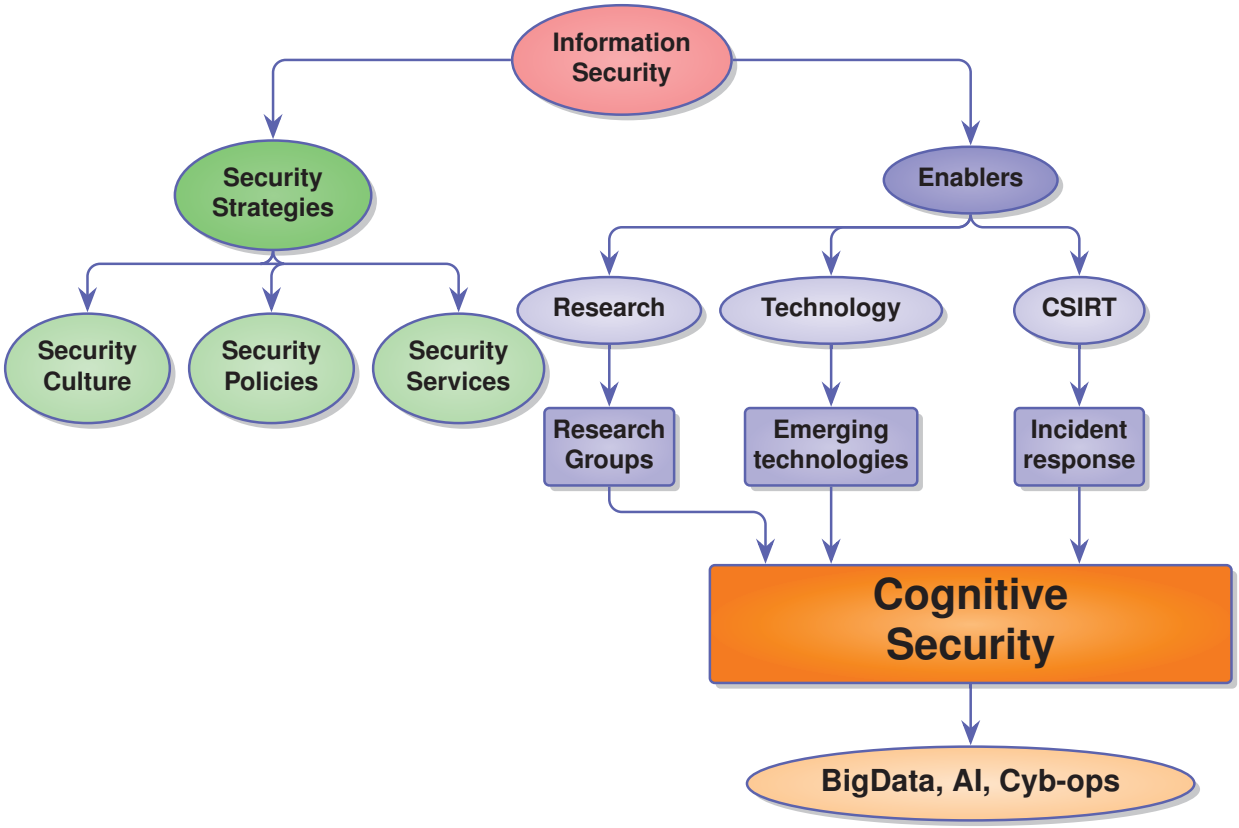

Figure 1. Cognitive Security Organization Model Proposal.

Instead Strong A.I, well known as True Intelligence or Artificial General Intelligence, is a system of AI able to make unknown task, Turing test was developed by Alan Turing in 1950, was to determine if a computer can think like a human.

B. Phishing

Phishing is a method used to stole people information.But phishing is not just to send an email and wait for someone who picks up the email; Phishing is a type of computer attack that communicates socially engineered messages to humans via electronic communication channels in order to persuade them to perform certain actions for the attacker's benefit.

C. Blacklist

Blacklist is a complete list of malicious file types and file extensions, including shortcut files, PowerShell and JScript files. Blacklist is one of the first options used for detecting phishing attacks, especially is used to separate fake URLs from real URLs; if URL is into Blacklist then is rejected otherwise it will be verify by another methods to determine if it is a secure URL.

D. Cybersecurity

According to Gomez [7], cybersecurity is defined like: "Any measure that prevents the execution of non authorized operations about a system o network, whose effects con lead damage over information, to compromise its confidentiality, authenticity or integrity, decrease performance devices or block the access of authorized-users to the system".

E. Machine Learning

Machine Learning is described by Jones [8], how the capability of introduce intelligence to an computer system to make specific actions. Machine learning is capable of use two methods: unsupervised and supervised methods. If the case an unsupervised method, this no implies the absence of training, in this method the computer systems learns from the identifying of data by itself.
Machine Learning is focused on methods and algorithms that allow a computer to learn. Given information, typically in data sets, machine learning can transform data into information through learning relationship between data. Usually algorithms like decision trees, nearest neighbor learning or Markov models are use to explore possible solutions. Algorithms of machine learning is used everyday, some examples of machine learning each day are: detecting spam, facial recognition, suggestion on Netflix and so on. Machine learning algorithms are involved in to many devices or activities that we do everyday.

III. METHODOLOGY

A. Background of Phishing Attacks

In may of 2017 when a bunch of hackers developed ”WannaCry” a ransomware attack, according to Cisco Annual Cybersecurity Report, there was a confusion about how this ransomware worm was infiltrating the systems, after that some private and public companies reported that this leak was through a phishing campaign with malicious mails and attached files [9].

According to Cisco Report, Telecom services were the main target, the first campaign was:

- Almost 60000 URLs containing subdomains like aaaainformaion[dot]org.
- Random strings with 50-62 letters.
- Hackers used inexpensive privacy to register all the domains observed in campaign.

In the same report, Cisco indicates that the second campaign was:

- The used name of real tax agency in United Kingdom to disguise their actions.
- Hackers employed 12 top-level domains, eleven of those domains were with 6 random letters.

Every day people use technology, because of their studies, jobs or business, but with this necessity of technology usually comes security problems and people doesn't know how to handle, detect or even if they are victims for this issues.

B. Increase of A.I techniques developed

According with World Intellectual Property Organization - WIPO [10], in its report of January 31st of 2019, conclude that:

- Since A.I started in 1950s, innovators and researchers have filed applications for nearly 340,000 AI-related inventions, and published over 1.6 million scientific publications.
- Companies represent 26 out of the top 30 A.I patent applicants, with universities or public research organizations accounting for the remaining four.
- Machine Learning, especially neural networks is the domain technique, in 2013 were 9,567 new patent applications, in 2016 were 20,195 applications, this means an 111 percent overall increase.

Due to this new trends we are looking for possible solutions to detect anomalous behaviour using AI-based techniques; machine learning algorithms are become popular, people try to implement programs where computers can take decisions instead humans, it means this will be an automatic process and due to big process capacity of devices, information, decisions and incident response will be more faster that years ago.

C. Applications with Machine Learning

Develop machine learning algorithms some times could be difficult, especially if we don't know how to star or what kind of programming language is better to develop machine learning applications. There exist many options, and to choose the right one, is necessary to ask some questions, like:

- What type of program will you develop?.
- Are you developing something with a big amount of data to be process?.

We decided use how programming language python for our machine learning solution. We decided python because is easy to use and it has some libraries used to develop machine learning algorithms. Anaconda was used to develop the machine learning algorithm, this give an graphic programming interface for python denominated Jupiter, and was necessary for our project import pandas library for data frame, numpy to do scientific operations; we use some packages that are made for machine learning like TfidfVectorzer from sklearn.feature_extraction.text wich is used to Convert a collection of raw documents to a matrix of TF-IDF features. Libraries are so important here, without them is impossible to make analysis, or worse develop machine learning programs. A function is created and it is called makeTokens(), is into this function that we will develop all actions required to convert URL string into a token.

Machine learning has different ways, so here we will discuss about how useful are Logistic Regression and Neural Networks for detecting anomalous behaviour.

1) Logistic Regression: Is one of the techniques of machine learning, is used in classification process, that is the reason for used this technique in this work, due of the amount of data that we need to process. In phishing data set, the analysis come from a matrix [11055] [32]. All this data is process into three options, legitimate, suspicious and phishing.

2) Artificial Neural Networks: Artificial Neural Networks -ANNs are based on the human brain process, its objective is simulate human behavior to take decision, learn, adapt or even abstract information. All this is get by training, but the way these networks learn is using simple processing units called ”Artificial Neurons”. ANNs are interesting, because of they are able to learn even if the information given is incomplete. This change the way of computing systems works, systems could fail if a part of it fails, but with ANNs if a neuron fails, this output will be overwritten by correct output of its neighbor neurons. ANNs can be used when there is little knowledge of the relationships between attributes and classes, are suitable for continuous value inputs and outputs, unlike most algorithms, are successful in a wide variety of real world problems, including recognition of manuscript characters, pathology's and medicine. In addition, parallelization techniques can be used to accelerate the computational process, several techniques have been recently developed for the extraction of rules from trained ANNs. These factors contribute to the usefulness of ANNs for numerical classification and prediction in data mining [**?**].

D. Data-sets

In this work we use the dataset share through Kaggle with information about phishing attacks Kummar [12]. We can observe different characteristics of URLs, like length, having sub-domain, and so on. It help to make an idea of what are the

most common characteristics in bad URLs, this was proposed by the work of Kummar. We develop a program in python who is capable to learn how detect bad URLs through machine learning algorithm, the respective analysis and afterwards give information about what URLs are bad.

IV. Results

This dataset is a matrix [420464 rows] [2 columns], the second column is titled label, as we can see all those are bad URLs. After develop the machine learning application we can see the accuracy of algorithm to find or determinate if a URL is bad or not. The result of algorithm is show in the Figure 2; exist a variable which is put as a parameter called "test_size", if value of this variable is equal or greater than 1 , the accuracy will be 1.0 but if the value of this variable change to less than 1 greater , then the accuracy will be less. In this case was used 0.8 as the value of this variable.

```
logic = LogisticRegression()
logic.fit(X_train, y_train)

LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)

print("Accuracy in this Logistic Regression is: ",logic.score(X_test, y_test))

Accuracy in this Logistic Regression is:  0.9339927223431201

xPredict = ["https://www.wipo.int/pressroom/en/articles/2019/article_0001.html",
"https://www.cisco.com/c/dam/m/hu_hu/campaigns/security-hub/pdf/acr-2018.pdf",
"https://code.likeagirl.io/an%C3%A1lisis-y-visualizaci%C3%B3n-de-datos-con-pandas-matplotlib-85ee4d7b4cad",
"www.radsport-voggel.de/wp-admin/includes/log.exe",
"ahrenhei.without-transfer.ru/nethost.exe ",
"www.itidea.it/centroesteticosothys/img/_notes/gum.exe"]

xPredict = vectorizer.transform(xPredict)
newPredict = logic.predict(xPredict)
print(newPredict)

['good' 'good' 'bad' 'bad' 'bad' 'bad']
```

Figure 2. Accuracy and Test of the algorithm

In the array of URLs, three of those are good URLs and the rest are bad, here is interesting that the algorithm made a mistake, because one of the three good URLs is marked as bad URL, that's because of the algorithm could need more specifications about bad URLs. So here is necessary to add information and make run the algorithm, it is probably that the accuracy will be greater.

Due to large amount of data, it was necessary to select specific elements of the data set, then it was better to select URL cases and analyze them (see Figure 3). This data set has to many rows and columns, and each column has different information. Here is important to know that each column is independent of other columns.

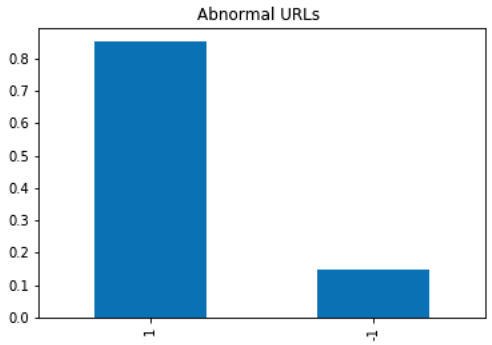

Figure 3. Abnormal URL Analysis

Information reveals something interesting, the URL length is an parameter to determinate if an URL could be phishing. In this case more than 80% of URLs were phishing, almost 19% was legitimate and only 1% was suspicious. Organizations like Anti-Phishing-Working-Group - apwg, made an analysis of shortening URLs, so they talk about shortening URLs and if this is a good option or no, and they conclude that in shortening URLs cases seen, phishing attacks are most active 4 hours before the reported date. As we can see length of URLs are so important (see Figure 4).

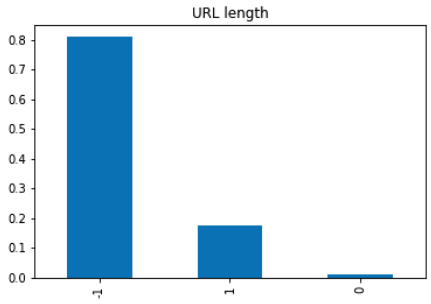

Figure 4. URL Length

This case is interesting, because if the attacker have a sub-domain, this data set say that is almost 31% possible that these URL is phishing, 36% is surely legitimate and 33% possible that is suspicious. As we can see these percentages are close from each other, this means is having sub-domain many times could be difficult to determinate if this URLs is reliable or not (see Figure 5).

In the Figure 6 shows that google index, is almost 90% more accurate to detect that an URL is legitimate and only over 10% is phishing. This is because google works with its bots and they check the web page and those bots conclude if the

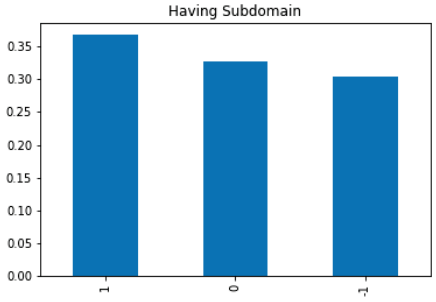

Figure 5. Having Subdomain

URL of that web is secure or not.

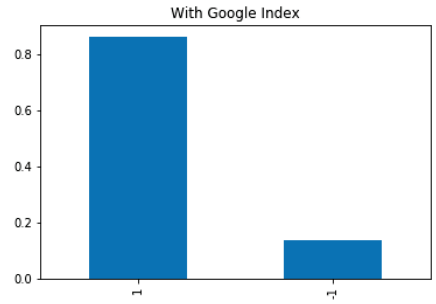

Figure 6. Google Index

Google is capable to tell us if something is legitimate or phishing.

- Google Index and Abnormal URLs are good way to determinate if URLs are reliable or not.
- Having sub-domains and URL length are unreliable therefore this options will show problems for community.

V. Conclusions

Anomalous behaviour is a common problem as we saw. In this document we go through these common problems and we analyzed them, it is thanks these analyzes and results that we can say that:

- Artificial Intelligence is a good tool to face this anomalous behavior, because it is faster, is efficient and modern technology let us develop better applications.
- Some phishing techniques like shortening URLs could be faced with tools like this machine learning application that is capable to determinate if a URL are good or bad, next step after that is add this URL to blacklist.
- Although this machine learning couldn't be right always we can check those URLs into a web checker of shortening URLs. Which bring us to next recommendation.
- Is recommendable not to open a shortened URL before check it, because we already know that many of those URLs could be phishing attacks, malware and so on.
- By knowing this weaknesses, as a community it is possible, to try to develop new tools that help us to improve this weak techniques.

References

[1] BBC News. (2019). Students blamed for college cyber-attacks. [online] Available at: https://www.bbc.com/news/education-45496714 [Accessed 27 Jul. 2019].

[2] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: A feature-rich machine learning framework for detecting phishing web sites," ACM Transactions on Information and System Security. vol. 14, no. 2 , pp. 1–28, Sept. 2011.

[3] Andrade, R. O., Fuertes, W., Cadena, S., Cadena, A., Tello-Oquendo, L., Córdova, D., ... and Cazares, M. F. (2019). Information Security Management in University Campus Using Cognitive Security. International Journal of Computer Science and Security (IJCSS), 13(4), 124.

[4] Vellino, Andre. (1986). Artificial intelligence: The very idea: J. Haugeland, (MIT Press, Cambridge, MA, 1985); 287 pp.. Artificial Intelligence. 29. 349–353.

[5] Poole, David and Mackworth, Alan and Goebel, Randy. (1998). Computational Intelligence: A Logical Approach.

[6] Nilsson, J. (1998). Real-Time Control Systems with Delays, Ph.D. Thesis. Dept. Automatic Control, Lund Institute of Technology, Lund Sweden.

[7] Gómez Á. (2010). Seguridad Informática básico. 1st ed. España: StarBook—, pp.13, 14.

[8] T. Jones, Los lenguajes de la Ai, [Online], Available: https://www.ibm.com/developerworks/ssa/library/cc-languages-artificial-intelligence/index.html

[9] Cisco, Cisco 2018 Annual Cybersecurity Report, Pg 8, 19, [Online] Available in: https://www.cisco.com/c/dam/m/hu_hu/campaigns/security-hub/pdf/acr-2018.pdf

[10] WIPO, (2019, 01, 31). WIPO's First "Technology Trends" Study Probes Artificial Intelligence: IBM and Microsoft are Leaders Amid Recent Global Upsurge in AI Inventive Activity [Online]. Available: https://www.wipo.int/pressroom/en/articles/2019/article_0001.html

[11] Bigus, J.P. (1996) Data Mining with Neural Network: Solving Business Problems from Applications Development to Decision Support. McGraw-Hill, New York

[12] A. Kummar, Phishing website dataset, kaggle, [Online] Avialable: https://www.kaggle.com/akashkr/phishing-website-dataset