Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 7.056

IJCSMC, Vol. 10, Issue. 7, July 2021, pg.56 – 60

Examination of Assorted Social Engineering Attack by Different Types of Machine Learning Algorithms

Amal Alhamad[1]; Dalal Aldablan[2]; Raghad Albahlal[3]

College of Science, Department of Computer Science and Information in Majmaah University, Al Zulfi 15941, Saudi Arabia

371200075@s.mu.edu.sa; 371200844@s.mu.edu.sa; 371200332@s.mu.edu.sa

DOI: 10.47760/ijcsmc.2021.v10i07.008

Abstract: The most powerful attack on the systems is Social Engineering Attack because of this attack deals with Psychology so that there is no hardware or software can prevent it or even can defend it and hence people need to be trained to defend against it.[1]

Social engineering is mostly done by phone or email.

In this research, which is based on previous research we have conducted, the aim of it was of it was to highlight the different social engineering attacks and how they can prevent in social network because social engineering is one of the biggest problems in social network, a concern the privacy and security.

This project is using a set of data then analysis it uses the Weka tool, to defend against these attacks we have evaluated three decision tree algorithms, RandomForest, REPTree and RandomTree.

It was also related to an J48 algorithm, On the contrary, here contains a complete overview of social engineering attacks, also more than one algorithm was searched.

Keywords: **Social Engineering, Social engineering attacks, RandomForest,, RandomTree, REPTree, Uniform Resource Locator (URL).**

I. Introduction

At present, technology occupies great importance concerning facilitating communication between people, as technology has had a negative and positive impact, making it affect the whole lifestyle, On his bad side the emergence of social engineering.

In our daily life we spend most of our time to looking inside or work on our mobiles (computers). Thus, we share information and data with people who do not necessarily know them, or that we have already met them.[1]

a true definition of social engineering is the act of manipulating a person to take an action that may or may not be in the “target’s” best interest. This may include obtaining information, gain- Ing access, or getting the target to take certain action.[3]

Today some of social networks like Facebook and Twitter are become the largest and most important sources of information, data exchange, and online services by virtue of its rapid growth. The social networks give full support to find new friends in addition to the exchange of data. And thus a new source of information is added to our knowledge. Clearly, most social network sites are critical with respect to user’s security and privacy due to the large amount of information available on them, as well as their very large user base.[2]

This paper is an extension of a previous paper in which we talked about social engineering and its impact on society, we have performed the detecting of URLs as benign or phishing. And how our modifying of the feature's effect on it.[4]

II. THEORETICAL CONSIDERATION

Social engineering attacks:

Attacks are divided into two types:

1- Human-Based Social Engineering Attack

In these attacks, the attacker collects sensitive information by deceiving and manipulating the victim to gain trust.

2- Social computer engineering attack

This is the type of social engineering that usually relies on technology and a computing system. Hence, we will discuss those attacks that will prove that social engineering can be a technical attack. See how these attacks rely on technical computing. [5]

1- Phishing

It is a fake email attack. It is used by a cybercriminal to obtain confidential information of the victim such as passwords and credit card numbers by opening links to hostile websites or clicking on attachments that contain malware. In the phishing technique, these messages are sent to all users.

2- Spear Phishing

Attacks focus on specific people or companies. The spear phishing technique needs more attempts on the part of the attacker and phishing may take a long time.

3- Vishing (voice or VoIP phishing)

It works like phishing but is often carried out using voice technology. A phishing attack can be performed by voice email, VoIP (Voice over IP), landline or cell phone.

4- POP-UP Window Attack

This type of attack sends the attacker an attractive popup or contains a prize or a winning opportunity for the victim to use any site, then he enters his username and password, from here the attack takes place and the attacker seizes the information.

5- Baiting

It is a bait that the attacker places on the victim in such a way that he is bound to steal all his data or impose malware on his system. The target person clicks on the bait out of their curiosity and then puts it in a work or home computer to develop it in the automatic installation of the malware. This is a type of brutal attack.

6- Pretexting

The attacker begins by pretending to need important information from the victim to carry out the assessment task. attacker Obtains data through fraud. And the attacker masquerades as a co-worker, the police, and the tax officials who have the power to find out things.

7- Scareware

This includes victims who panicked by malware. Scareware has also been mentioned as fraud tools. This type is spread through spam emails that spread fraudulent malware threats. Users are told that malware has corrupted their system, in order to convince them to interfere with programs that only benefit the attacker or the malware itself. [6]

In our data preparation, we have labelled the phishing URLs 1 and the benign URLs 0, in the training phase, we us decision tree algorithms RandomForest, RandomTree and REPTree. This phase shows three train model, so we have evaluated three decision tree learning model in our dataset, in the testing phase we have used unknown URLs has tasted using the train model as phishing or benign.

Feature extraction:

URLs feature:

The 49 are from the literature [7], these features are shown in Table1.

Sr.No	Feature name	Type
Features used in the literature		
1.	NumDots	numeric
2.	SubdomainLevel	numeric
3.	PathLevel	numeric
4.	UrlLength	numeric
5.	NumDash	numeric
6.	NumDashInHostname	numeric
7.	AtSymbol	numeric
8.	TildeSymbol	numeric

9.	NumUnderscore	numeric
10.	NumPercent	numeric
11.	NumQueryComponents	numeric
12.	NumAmpersand	numeric
13.	NumHash	numeric
14.	Num Chars	numeric
15.	No Https	numeric
16.	RandomString	numeric
17.	IpAddress	numeric
18.	DomainInSubdomains	numeric
19.	DomainInPaths	numeric
20.	HttpsInHostname	numeric
21.	HostnameLength	numeric
22.	PathLength	numeric
23.	QueryLength	numeric
24.	DoubleSlashInPath	numeric
25.	NumSensitiveWords	numeric
26.	EmbeddedBrandName	numeric
27.	PctExtHyperlinks	numeric
28.	PctExtResourceUrls	numeric
29.	ExtFavicon	numeric
30.	InsecureForms	numeric
31.	RelativeFormAction	numeric
32.	ExtFormAction	numeric
33.	AbnormalFormAction	numeric
34.	PctNullSelfRedirectHyperlinks	numeric
35.	FrequentDomainNameMismatch	numeric
36.	FakeLinkInStatusBar	numeric
37.	RightClickDisabled	numeric
38.	PopUpWindow	numeric
39.	SubmitInfoToEmail	numeric
40.	IframeOrFrame	numeric
41.	MissingTitle	numeric
42.	ImagesOnlyInForm	numeric
43.	SubdomainLevelRT	numeric
44.	UrlLengthRT	numeric
45.	PctExtResourceUrlsRT	numeric
46.	AbnormalExtFormActionR	numeric
47.	ExtMetaScriptLinkRT	numeric
48.	PctExtNullSelfRedirectHyperlinksRT	numeric
49.	{0,1}	nominal

Table1.URLs Feature

Decision tree algorithms:

Machine learning contain a several classifications techniques one of the most widely of them is a decision tree algorithm, in our research we use three types of it.

RandomForest algorithm:

Random forest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. It is an effective classifier in prediction. Random forest generally exhibits a substantial performance improvement over the single tree classifier such as CART and C4.5 [8]

RandomTree algorithm:

Random Tree is a supervised Classifier; it is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. In standard tree each node is split using the best split among all variables. In a random forest, each node is split using the best among the subset of predicators randomly chosen at that node.[10]

REPTree algorithm:

RepTree uses the regression tree logic and creates multiple trees in different iterations. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances.[9]

Counting Gain:

This process uses the "Entropy" which is a measure of the data disorder. The Entropy of is calculated by
And Gain is "[11].

$$\boldsymbol{Entropy}(\vec{y}) = -\sum_{j=1}^{n} \frac{|y_i|}{|\vec{y}|} \log\left(\frac{|y_i|}{|\vec{y}|}\right)$$

$$\boldsymbol{Entropy}(j|\vec{y}) = \frac{|y_j|}{|\vec{y}|} \log\left(\frac{|y_j|}{|\vec{y}|}\right)$$

$$\boldsymbol{Gain}(\vec{y}, j) = \boldsymbol{Entropy}(\vec{y} - \boldsymbol{Entropy}(j|\vec{y}))\ |$$

Tool:

We have used Weka in our algorithms to gain the final decision on, weather the URL is benign or phishing.

III.Experimental Consideration

Data source and dataset:

The dataset is from Mendeley Data contains 48 features extracted from 5000 phishing webpages and 5000 legitimate webpages. An improved feature extraction technique is. And this is the features that in the set.

Task	Benign	Phishing	Total
Training	5000	5000	10000

Table2.The Dataset.

Evaluation results:

We have evaluated the decision tree algorithms classifiers in our dataset, so we modified some features to get butter rustles and increase the accuracy.

Modify of features:

Algorithm	Sr.No	Feature name	Type
RandomForest algorithm	14.	Num Chars	numeric
RandomTree algorithm	36.	FakeLinkInStatusBar	numeric
REPTree algorithm	1.	NumDots	numeric

Table3. Features Deleted.

Accuracy:

Classifier	Accuracy without modifying	Accuracy with modifying	Change %
RandomForest	98.37%	98.48%	0.11% ↑
Random	95.98%	96.08%	0.1%
REPTree	96.67%	96.76%	0.12%

Table 4. Accuracy.

Conclusions

In this paper, we discussed the issue of social engineering and its impact on society also we have performed the detecting of URLs as benign or phishing. We have evaluated RandomForest, Random, REPTree decision tree algorithms in term of detecting accuracy, and how our modifying of the features effect on it. We have a prepared dataset 10000 URLs, among wish 4958 are benign and 4942 are phishing. Also, we talk about social engineering and the effects of it in society.

ACKNOWLEDGEMENT

We thank Allah for everything that made us capable of accomplishing this research.

REFERENCES

[1] KOYUN,A &AlJanabi,E(2017), Social Engineering Attacks , Journal of Multidisciplinary Engineering Science and Technology (JMEST).
[2] D. Irani, M. Balduzzi, D. Balzarotti, E. Kirda, and C. Pu. Reverse social engineering attacks in online social networks. Detection of Intrusions and Malware, and Vulnerability Assessment, 2011.
[3] Hadnagy,D.(2011).Social Engineering: The Art of Human Hacking .Canada:Wiley Publishing, Inc.
[4] Al-dablan,D,A , Al-hamad,A,H , Al-Bahlal,R,F & Badaw,M,A(2020) ,An Analysis of Various Social Engineering Attack in Social Network using Machine Learning Algorithm ,IJCSNS International Journal of Computer Science and Network Security,10.22937/IJCSNS.2020.20.10.7.
[5] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.401.7920&rep=rep1&type=pdf
[6] M. NazreenBanu et al, A Comprehensive Study of Phishing 1.Attacks‖ , IJCSIT, Vol. 4, 2013, pp.783-786
[7] Phishing Datset,data.mendeley.com
[8] Breiman,L.Random Forests. Machine Learning 45, 5–32 (2001).https://doi.org/10.1023/A:1010933404324.
[9] Kalmegh,S(2015), Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News, IJISET.
[10] N. Landwehr, M. Hall, and E. Frank. Logistic model trees, 2003.
[11] Mazraeh, S., Modhej, A., Neysi,(2016), S.H.N.: Intrusion detection in computer networks using combination of machine learning tech- niques. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) 16(8), 122 (2016).