International Journal of Scientific & Engineering Research Volume 11, Issue 7, July-2020
ISSN 2229-5518

MACHINE LEARNING BASED MALICIOUS WEBSITE DETECTION.

Jino S Ganesh, Niranjan Swarup.V, Madhan Kumar.R, Harinisree.A, P.G Scholars,
Guided by, Dr. Giri Raj.M, Associate professor,
School of Mechanical Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India

Abstract: In the present generation of the digital world era there are a lot of opportunities to steal data. The data can be any type and any form. Some are mention like Top classified data, industrial data, finical data, intellectual property data, personal data, IOT & IIOT, and big data. threats and different viruses where used against people to steal data or amounts of money through the internet by using different types of malware. In the present world, the whole life of people is depending on the mobiles or pc or laptops with the internet. So, Data thieves use the internet as their opportunity to steal the people's data by creating viruses and using them as a URL site to do their work. So, with the help of some Machine learning algorithm using python, we try to identify the malicious websites which are insecure for the people to use.

Keywords: Malware, Machine learning, python, URL site.

Introduction

There are a lot of good things on the internet which we use in our day to day life. As we can say every human mostly start their day with the internet. As I said there are a lot of good things on the internet like so there is an equal amount of worst inside the internet. Now because of the recent growth of the internet, the frauds start to commit their crime through the internet by stealing the data of the people by sending the people a harmful virus in sort of message or through website, advertisements, third-party applications, and so on. They commit these crimes to steal the personal data from their computer or to hack their computer and take control of the system or to steal the money from them. The malicious URL involves spam and phishing which are types of frauds used by criminals who try to steal data from trusted organizations or people. This malicious URL consists of will contain malware or the trojan which will take whole control of your system when u just access the URL which results in the loss of information which will the user's personal information of accessing his webcam which will be accessed without his knowledge.

In this paper, we are going to see the malicious website in which the websites which are harmful to open as they contain the virus in them to hack the system of the people who opens the URL. With the help of machine learning, we try to find out that malicious websites which are very harmful to browse in our system. As we use a certain algorithm to train them to find the harmful viruses which attack the user and prevent them by alerting the user not to access the site which he is trying to browse.

Literature review

Implementing phishing detection system using seven different machine learning algorithms such as decision tree, Adaboost, K-Star, kNN (n = 3), Random Forest and SMO. Naive Bayes, and various number/type features based on NLP Functions, word vectors, mixed functions.

Build effective functions to improve the accuracy of the detection system Lists are an important task. All the algorithms have been tested and result is compared. In this they have implemented a new proposal of combining NLP (natural language processing-based features) and word vector as a hybrid model. This has advantages like language independence, real time execution, detection of new websites, independent from third party, use of feature-rich classifiers' etc. [14]. Two parallel modules that extract functional representations of URLs. The first is the character-level CNN module. The other is an attention-based hierarchical RNN module is proposed to find phishing URL detection. They observed that Random forest classifier Bag-of words trained SVM classifiers that reach the highest AUC get the highest accuracy. In this paper, an effective deep learning-based algorithm have shown the effectiveness of the phishing URL detection approach. Ablation experiment Character level spatial feature representation extracted word-level temporal feature representation From character-level CNNs and attention-based URLs Hierarchical RNN modules improve performance, The generalization ability of this approach is high [12]. This article solves the problem of identifying phishing URLs under weak supervision, which requires A smaller amount of labeled data to start the learning process. In an active learning framework, follow effective Human-computer collaboration method, the existing model

is Fine-tune gradually by exploring training without annotations sample. The automation solution is designed to select downward And identify the relevant unannotated samples needed Human intelligence to obtain reliable annotations. in a In a batch learning environment, priority activities are recommended Learning can further accelerate the learning process by automatically screening a set of samples useful to humans annotation. Effective feature weighting process for evaluation the relative importance of element size improvement on behalf of the task is effective. Due to manual intervention Only a small number of unannotated samples are needed, our active learning framework greatly reduces the burden Security analyst while ensuring faster Converge to the best solution. [13].

They developed a potential URL attack detection system Based on PU learning. Compared with the method based on supervised learning, this method requires only a few malicious URLs, and Use unmarked URLs, this is suitable for the actual situation That was encountered. The developed system mainly includes three parts: First, the feature extraction process is performed to transmit the original URL. Enter the digital feature vector; second, the two-stage strategy and Use a cost-sensitive strategy to train the classification model; in the end, each new URL will first be converted to Numerical feature vectors, and then input into the learned model, URLs with high scores will be considered potentially malicious The possibility of URL is very high. Empirical results show that the developed system can effectively Discover potential URL attacks. The system can be deployed as a help to existing systems or used to help Network security engineers can effectively discover potential attacks mode.[14]. In this article, the author understands that through certain technologies, hackers can obtain our personal and private Information, they can bypass our several methods to detect them and protect themselves from it as well. One of these serious treatments, and this article focuses on malicious URLs, Hackers have multiple techniques and algorithms to obfuscate their URLs to bypass defensive measures. It's a promising option to surpass them in this competition to protect our data, they chose This article will introduce machine algorithms for detecting URLs or classifying them as benign or malicious. Although they are a good way to improve security, they are expensive and difficult to adapt to certain environments.[15]

URL (Uniform Resource Locator).

URL i.e., Uniform Resource Locator which is termed as web addresses to access a website. It is a web source that identifies its location on the computer network and helps in accessing the web page. The URL of the web page will generally indicate the HTTP protocols, the hostnames, and a file name generally we can say as a syntax form.

URL format:

protocol://hostname/other_information.

The protocol i.e., HTTP will specify how the information from the link is transferred. Mostly we use Http protocol which is hypertext transfer protocol. There are other protocols other than Http which are FTP, telnet, newsgroups, and the Gopher. Other than the protocol the URL uses the domain name which tells the browser where and how to retrieve the data of the source.

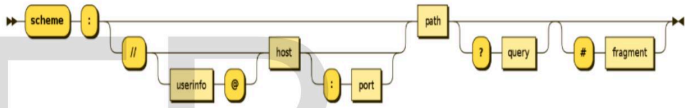


Fig.1. URL syntax diagram

There are two kinds of URLs which are an absolute URL and a relative URL. The Absolute URL is a kind of URL that contains all the information necessary to locate a resource. A relative URL locates a resource using an absolute URL as a starting point.in effect, the complete URL of the target is specified by linking the two URLs which are absolute and relative URL. The above format is the absolute URL format and the relative URL consists of only the path and optionally the resource but no scheme or server. The scheme specifies how the resource will be accessed and the server the name of the computer where the resource is located. The path specifies the sequence of directories which lead to the target and the resource which is the target and name of the file. This may be a simple file containing a single binary stream of bytes or structured document which contains more storages and binary of streams.

Malicious URL

The malicious URL looks like a normal URL. A simple URL will cause heavy damage to your digital device. These Malicious links are considered as one of the biggest threats to the present digital world. Even they can do these attacks by

sending these types of viruses or any links through email too[5].

It is said that the Malicious URL link was created with a purpose for promoting the scam, attacks, and frauds. If we click the URL which is an affected one the malware or trojan will automatically download and will take control over the device. The most common malicious URL scam involves Spam and Phishing. Which are the types of frauds used by criminals who try to hack and steal the personal information of the victim?

We will get these types of URLs by receiving in the email by an unknown or stranger or through message. These malicious URLs are considered as the launchpads for the attacks. By accessing these URLs which will help the frauds in stealing the personal information or will automatically install malware or viruses or trojan without the user's knowledge. There are considered as threats to cybersecurity threats[8].

Since the usage of Social websites or applications like Facebook, Instagram, Twitter, Telegram or WhatsApp and so on were used widely by almost 90% of the people from the world wide so the Data thieves now just using these platforms to do their crimes by creating a spam website or advertisements or pictures while we select those spams the virus will be automatically downloaded in our devices without our knowledge and it will do their work for which they are created.

How phishing or malicious compromise our PC:

We think that we can avoid the malware from the websites infecting our pc by avoiding selecting the URL or the ads or by refusing to download the software from the malicious websites. But there are still many more ways for us to be infected by

- Drive-by-download: It automatically without our knowledge when visiting some malicious sites and installs in our device and activates without our knowledge.
- Java Script Infection: Since most of the websites are designed by Js. The hackers create a JavaScript malware that infects our device by downloading the .js infected file which our browser executes automatically and starts its process by downloading the malware.
- Malvertising: Ad networks are hijacked by hackers and infect those ads and spread it far and wide.
- URL Injection: Injecting the malicious URL to a normal site when browsing those sites will automatically redirect to the malicious site.
- Malicious Redirect.
- Browser Hijackers.

Phishing website:

It is a type of threat that mostly will happen with the knowledge of the user. The above threats have happened without the knowledge of the human, unlike this Phishing website in which we give our personal information to the fake websites like fake eBay site or any fake online delivery site. Since these phishing sites are very tricky and look like the clones of the original site in which the users can easily fall.

Detection of Malicious URL

General ways to detect Malicious URL:

The Malicious sites can be detected by some general patterns which can be a noticeable pattern when it comes to user infections. Mostly the sites like gambling, gaming, porn, and video streaming are the frequent targets. When we go through these sites when we select a link without our permission 2-3 windows popup in the browser or else it will ask us to download new software or browser extension. They run on only two things like traffic or ad-click. The malicious hacker will make full use of the weak plugins to infect the ads or popups which infect the user, end-user[10-15].

In this paper, we try to identify malicious websites with the help of Machine Learning. We try to train different types of the algorithm to identify the type of website which we are trying to searching is malicious URL or safe URL. Machine Learning is one of the present technology in which if we train them to do one thing and get the results which will do it automatically for the next processes.

Machine Learning

Machine Learning is work on their own according to their experiences. It is a type of computer algorithm which works automatically by their experiences and their mistakes. Machine Learning is the subsets of AI (Artificial Intelligence)[8-10]. There are lots of algorithms in machine learning to train the data. This machine learning is widely used in almost all the fields to get a result from the computer. To use this Machine Learning process in any of the fields to get the

results independently we need to feed the required algorithm and data to the system in advance and should define the analysis rules for the recognition of the pattern. After this process, the system performs the following tasks

- Finding the data, retrieving and summarizing the data's
- Making the prediction based on the analysis data
- Calculating the probabilities of the specific results
- Adapting to certain development autonomously
- Optimizing the process based on the recognized pattern.

Thus, by doing these processes It will be automatically trained to do their work. Thus, by these types of processes, we will try to find the malicious websites by training them with certain algorithms.

In this process, we train the machine learning algorithms using python to find the malicious websites which can be used in the browser which we use. When we type the web address in the address bar of the browser which we use the algorithms which we trained will automatically do their work by checking the web address and if it's a safe site it will process by browsing the website and if it is a malicious URL then it will display a dialog box in which it says a message that it is an unprotected malicious website do u still want to process further to continue means select yes[14-18].

The Machine Learning types which we used for the identification of malicious URL are explained below there are many machine learning algorithms in which we chose a particular type of algorithm to do the process.

Logistic Regression

The probability of a certain class or event existing such as pass/fail cases. This can be extended to model several classes of events such as determining whether an image contains a bike, bus, car. Each particular are detected in the image which would be assigned a probability between 1 and 0, with a sum of one. Logistic regression is another technique borrowed by machine learning from the field of statistics[1].

Logistic Function

The logistic function also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

Equation for Logistic Regression:

1/(1+e^-value)

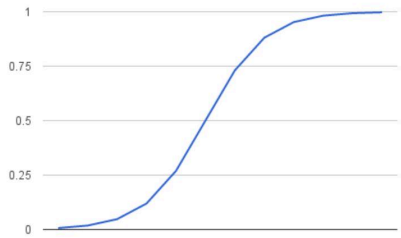


Fig.2.Logistic Regression

The representation used for logistic regression:

Logistic regression uses an equation as the representation, very much like linear regression. Input values (X) are combined linearly using weights or coefficient values to predict an output value (Y), A key difference from linear regression is that the output value is being modeled is a binary value (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

Y=e^(b0+b1*x) / (1+e^(b0+b1*x))

Malware Detection Logistic regression

Traditional virus scanners use file signatures to detect malware. when a file is scanned, its signature is compared to a database of unknown malicious file signatures. If there is a match, the file is flagged as being malicious. However, this approach may fail to detect newly created malware, it depends upon having an up-to-date database of signatures. By using machine learning, malicious applications can be detected without the need for a database of signatures[13].

Classification using Logistic Regression

Logistic regression is a method of performing regression on a database that has categorial target values. The logistic function is used to transform linear combinations of the explanatory

variables into possibilities[1-3]. The definition of the logistic function is as follows

$$\sigma(t) = \frac{1}{1+e^{-t}}$$

Equation 1: Logistic function

This function used to transform the typical linear regression formula

$$f(x) = \beta_0 + \beta_1 x$$

Equation 2: Linear Regression function

In this formula p(x) represents the probability that an input sample belongs to the target 1. That is the probability that an application is malicious given that it is making the observed system calls.

$$p(x) = \frac{1}{1+e^{-\beta_0-\beta_1 x}}$$

Equation 3: Logistic Regression

DECISION TREE

The decision tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a decision tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data. In decision trees, for predicting a class label for a record we start from the root of the tree[3]. We compare the values of the root attribute with the record's attribute. Based on the comparison, we follow the branch corresponding to that value and jump to the next node.

Types of decision trees

Types of decision trees are based on the type of target variable we have. It can be of two types:

1.Categorical Variable Decision Tree

2.Continuous Variable Decision Tree

Random forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Since the random forest combines multiple trees to predict the class of the dataset, some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier.

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

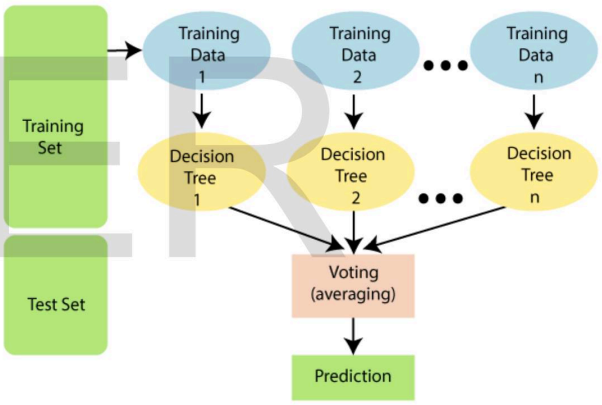


Fig.3.Random forest

MULTILAYER PERCEPTION

In the Multilayer perceptron, there can more than one linear layer. If we take the simple example of the three-layer network, the first layer will be the input layer and the last will be the output layer and the middle layer will be called *the* hidden layer. We feed our input data into the input layer and take the output from the output layer. We can increase the number of the hidden layer as much as we want, to make the model more complex according to our task[4-9].

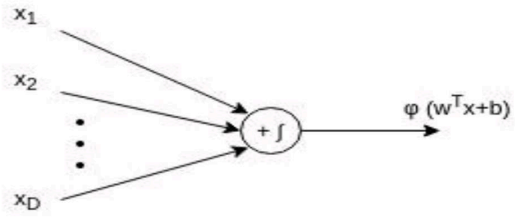


Fig.4. a. Multilayer perceptron

Equation for multilayer Perceptron:

weight = weight + learning rate * (expected - predicted) * x

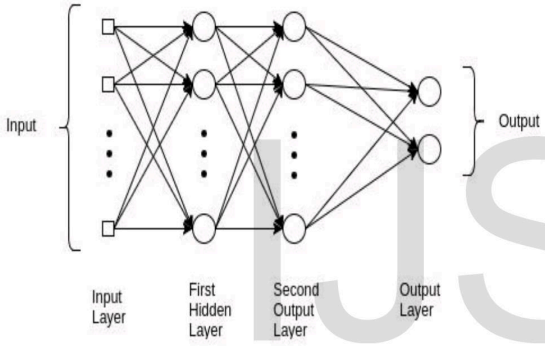


Fig.4. b. Multilayer perceptron

Proposed work

In this paper, as we mentioned in Fig.5 which represents the working process of our system i.e., the methodology of our work. Where algorithm goal is to find out the current URL website which is browsing is good to go for browse or bad. By using the machine learning algorithm, we are going to classify which type of URL is a suspicious URL. For that detection process, We have a data set about URL from GitHub which we use for the processing and extract the feature and through the data set we used for the processing and a huge dataset which is classified over, The data set is split for training and testing data are in ratio 75/25. The data set is trained and after the train and accuracy obtain later that passing the test dataset and predict the result for the test dataset and final we use the new data for the prediction of unknown data and the result is obtained. Classification error which is reduced by the change in this iteration and result to classify the given URL whether it is good to browse or malicious one to work needs to be precautioned.

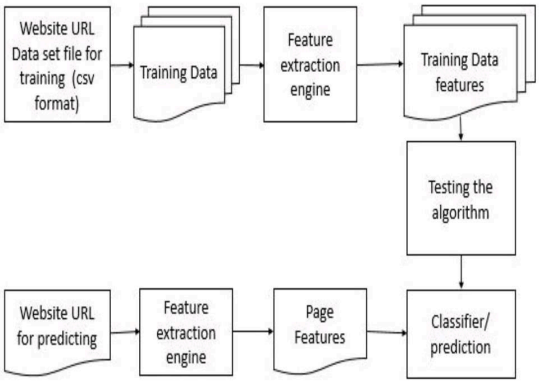


Fig.5. Detection of the malicious URL Flow Diagram

We ran our experiment on the machine with intel core i7-9750 CPU @ 2.60 GHz to 4.7 GHz processor with 8 GB memory. Memory lag or exhaustion was not an issue used typically 650MB to 1250MB megabytes of RAM. We implemented a Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Multilayer Perceptron package from Scikit Learn Package used for machine learning algorithm mathematics package in PYTHON and each has tokenized dataset and train and test dataset is common and compares each algorithm gives a slight difference in result accuracy.

Fig.6. Logistic regression

Fig.7. Logistic Regression training process

The fig.6 and fig 7 shows the working progress i.e., the training progress of the web URL which we used for detecting the URL the safe one or

not. By training these data it will automatically detect the website when we just type in the browser where we browse the URL.

The Fig.8 shows the training process of the decision tree algorithm in which the URL is safe to browse or not.

Fig.8.Decision Tree training process

The Fig. 9 explains about the URL which we are training is good to browse or malware website and is not a protected site to browse in the device which we use with the help of Random Forest algorithm.

Fig.9.Random forest training process

The Fig.10 explains about the working iterations of machine learning algorithm. In that image shows the process of multilayer perceptron working process.

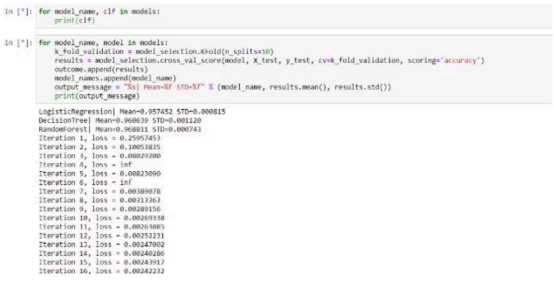


Fig.10. Working iteration of Multilayer Preceptron

Result:

The experimental result and comparison detail of the proposed algorithms model classification and used feature extract are trained and test for the data and the prediction of the new URL website are shown in the given table.1 below in which it explains the types of algorithm which we used and the out come of the training process and the mean values and the average value and the accurarcy of the result about the website which is save ie this much % is safe for browsing the URL.

Algorithm	Mean	Standard	Accuracy
Logistic Regression	0.9372	0.0017	94.5%
Decision Tree	0.9505	0.0018	95.2%
Random Forest	0.9601	0.0023	96.8%
Multilayer perceptron	0.8625	0.0047	83.5%

Table 1. Training results of algorithm

The table explains the resulting outcome of the training algorithm about how efficient it is to detect the malicious URL website and algorithm with good accuracy gives us better classification results of any website either it is Good or Bad. This is easy to understand by any user and it makes them alert that they are using a malicious website and which can save them from an online scam, online attack, hacking, phishing, credentials detail steal. This all can be done with machine learning and we can train it easy and get better accuracy results for our application.

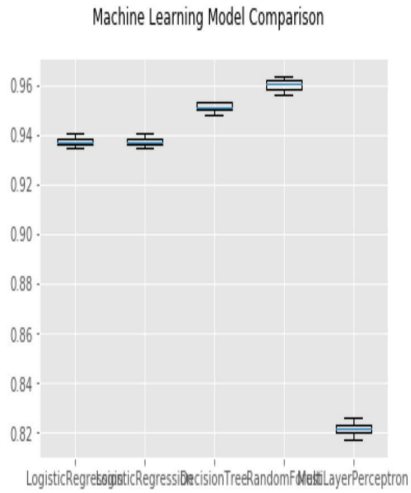


Fig.11. Accurarcy in Machine learning

Conclusion and Future Scope

We can clearly see that some-techniques and method of hacking which is know to be social attacking using phishing malware website, etc, a hack can obtain our data, credentials information, corporate

information which can be either public or private data, where they make changes using our profile for their benefits, blackmail, damage reputation of the organisation all are done with a single click on the malicious website you giving you data in wrong hands. This paper mainly focuses on malicious websites and URLs where hackers have several techniques and algorithms to bypass al your firewall and system defense protection.

In the race of hackers, we need to protect our data for that we have implemented a machine-learning algorithm to detect and inform the user about malicious websites. In our system by using four different machine learning algorithms, like logistic regression, decision tree, random forest, multilayer perceptron neural networks. To increase the accuracy of the detection system, the construction of good efficient feature for malicious detection.

In future we try to implement these algorithms in the browser in which when we copy an URL and paste in in the browser the machine learning automatically with the help of these algorithm will help the user by showing a message and sking permission to proceed or deny the process by selection the option which will be given below. If the URL is a safe one it automatically will do the process ie., it starts to browse the page automatically without any intreption.

Reference:

[1] A. I. Schein and L. H. Ungar, *Active learning for logistic regression: An evaluation*, vol. 68, no. 3. 2007.

[2] M. R. Segal, "Machine learning benchmarks and random forest regression. *Center for Bioinformatics and Molecular Biostatistics, UC San Francisco, USA.*," 2004.

[3] F. Livingston, "Implementation of Breiman's Random Forest Machine Learning Algorithm," *Mach. Learn. J. Pap.*, pp. 1–13, 2005.

[4] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences," *Atmos. Environ.*, vol. 32, no. 14–15, pp. 2627–2636, 1998.

[5] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data Soc.*, vol. 3, no. 1, pp. 1–12, 2016.

[6] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Inform.*, vol. 35, no. 5–6, pp. 352–359, 2002.

[7] D. Landgrebe, "Iiiii!Iiiiiiiiiii!Iiiii Iii Iiiiiiiiiiiiiiil," 2017.

[8] X. Yan, Y. Xu, B. Cui, S. Zhang, T. Guo, and C. Li, "Learning URL Embedding for Malicious Website Detection," *IEEE Trans. Ind. Informatics*, vol. 3203, no. c, pp. 1–1, 2020.

[9] M. Alazab and S. Fellow, "Malicious URL Detection using Deep Learning," pp. 1–9, 2020.

[10] Y. L. Zhang *et al.*, "Poster: A PU learning-based system for potential malicious URL detection," *Proc. ACM Conf. Comput. Commun. Secur.*, pp. 2599–2601, 2017.

[11] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *J. Inf. Secur. Appl.*, vol. 50, p. 102419, 2020.

[12] R. Vinayakumar, K. P. Soman, P. Poornachandran, V. S. Mohan, and A. D. Kumar, "ScaleNet: Scalable and hybrid framework for cyber threat situational awareness based on DNS, URL, and email data analysis," *J. Cyber Secur. Mobil.*, vol. 8, no. 2, pp. 189–240, 2018.

[13] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Syst. Appl.*, vol. 117, pp. 345–357, 2019.

[14] Y. Huang, Q. Yang, J. Qin, and W. Wen, "Phishing URL detection via CNN and attention-based hierarchical RNN," *Proc. - 2019 18th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. IEEE Int. Conf. Big Data Sci. Eng. Trust. 2019*, pp. 112–119, 2019.

[15] A. Fathima and K. Vaidehi, *Advances in Decision Sciences, Image Processing, Security and Computer Vision*, vol. 4, no. Vm. Springer International Publishing, 2020.

[16] S. Das Bhattacharjee, A. Talukder, E. Al-Shaer, and P. Doshi, "Prioritized active learning for malicious URL detection using weighted text-based features," *2017 IEEE Int. Conf. Intell. Secur. Informatics Secur. Big Data, ISI 2017*, pp. 107–112, 2017.

[17] D. K. Karnase, "A Review on Malicious URL Detection using Machine Learning Systems," pp. 214–219.

[18] M. Ferreira, "Malicious URL Detection using Machine Learning Algorithms," pp. 114–122, 2019.

[19] Detecting Malicious URLs with Machine Learning, https://ritcsec.wordpress.com/2017/12/07/detecting malicious-

urls-with-machine-learning/, last accessed 11/12/2018
[20] Y.-T. Hou, Y. Chang, T. Chen, C.-S. Laih, and C.-M. Chen, "Malicious web content detection by machine
learning," Expert Systems with Applications, vol. 37:1, pp. 55–60. (2010)

International Journal of Scientific & Engineering Research Volume 11, Issue 7, July-2020
ISSN 2229-5518

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

MACHINE LEARNING BASED MALICIOUS WEBSITE DETECTION

Guided by:
Dr. Giri Raja. M

Presented by-
Jino S Ganesh - 19MMS0009
Niranjan Swarup.V - 19MMS0005
Madhan Kumar.R - 19MMS0011
Harinisree.A - 19MMS0008

Abstract

In the present generation of the digital world era there are a lot of opportunities to steal data

The data can be any type and any form

Data thieves use the internet as their opportunity to steal people's data by creating virus and use them as a URL site to fetch information

In this paper, with the help of some Machine learning algorithm, we try to identify the malicious websites

Introduction

The Digital theft through internet have been the present generation frauds or crimes which are going around.

The websites which we use in our day to day life like all social websites are mainly used by the hackers to send the viruses

In this paper we explained about the hackers using the URL to do their things

With the help of the URL they will try sending the viruses to the user who browse these URL.

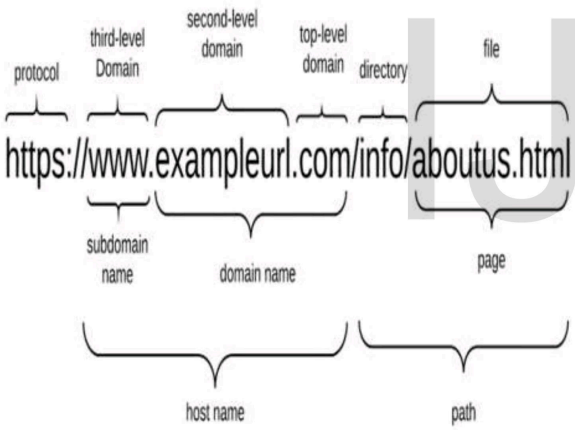
We try to detect these types of URL with the help of Machin e Learning algorithm.

International Journal of Scientific & Engineering Research Volume 11, Issue 7, July-2020
ISSN 2229-5518

Literature review

Title of the paper	Authors	Description
Machine learning based phishing detection from URLs	Ozgur Koray Sahingoz a, Ebubekir Buber, Banu Diri	Implementing phishing detection system using seven different machine learning algorithms such as decision tree, Adaboost, K-Star, kNN (n = 3), Random Forest and SMO.
Phishing URL Detection via CNN and Attention-Based Hierarchical RNN	Yongjie Huang, Qiping Yang, Jinghui Qin, Wushao Wen	Two parallel modules that extract functional representations of URLs. The first is the character-level CNN module. The other is an attention-based hierarchical RNN module is proposed to find phishing URL detection.
Prioritized Active Learning for Malicious URL Detection using Weighted Text-Based Features	Sreyasee Das Bhattacharjee, Ashit Talukder, Ehab Al-Shaer, Pratik Doshi	This article solves the problem of identifying phishing URLs under weak supervision, which requires A smaller amount of labeled data to start the learning process.
A PU Learning based System for Potential Malicious URL Detection	Ya-Lin Zhang, Longfei Li, Jun Zhou, Xiaolong Li, Yujiang Liu	They developed a potential URL attack detection system Based on PU learning. Compared with the method based on supervised learning, this method requires only a few malicious URLs, and Use unmarked URLs, this is suitable for the actual situation That was encountered

URL (Uniform Resource Locator) :



- A URL is termed as web address that is specifies its location on a computer network and a mechanism of retrieving it.
- A URL is the fundamental network identification for any resource connected to the web.
- You can find it in the address bar of the web browser. Alternatively, you can find the URL for a link by right-clicking it and copying the link.

Malicious URL:

- Malicious URL will look like a normal URL
- It’s a type of URL when browsed will automatically redirect to two or three links with out our permission.
- Those links are the URL which contains Viruses created by data frauds.
- This URL is used by the Data frauds to steal the personal data's or moneys or will automatically download the viruses to the user's device.
- The virus will take control over the device and will automatically hack the device by staling the data's or hacking the camera and giving the live feeds to the hacker with out the knowledge.
- The users will go through the URL through any gaming websites, video streaming website, porn sites or even might be received to the user by an unknown mail or message through the social website.

Phishing website:

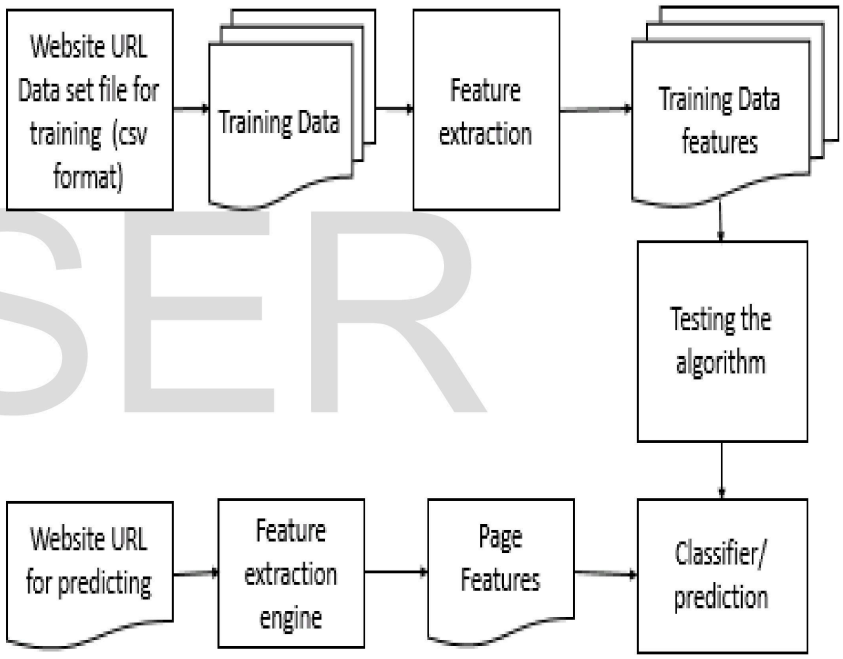
- Phishing website will be displayed like the clone of the Famous websites.
- The websites that are cloned are mostly the websites like online purchase websites like amazon , eBay , flip kart, ekart etc.
- They will steal our personal information's or bank details which the user provides to them

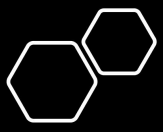
Machine Learning in Cyber Security:

- With machine learning, cyber security systems can analyse features and learn from them to help prevent similar attacks and respond to changing behaviour.
- It can help in preventing threats and responding to active attacks in real time. Here we could detect a malicious URL from a non-malicious URL using some machine learning algorithms
- **Analysis:**
- Logistic Regression
- Decision Tree
- Random Forests

Methodology

- Upload data set to algorithm
- Take the training data set and split it for 75% for training the algorithm and 25% for testing the algorithm
- Train the data set using machine learning algorithm
- After the training now test with 25% dataset
- Now algorithm ready to predict new data or unknown data can be predict the result





TOOLS

- Python 3.0 version
- Jupiter notebook IDLE
- SCIKIT LEARN
- PANDAS
- NUMPY

CODE (logistic Regression)

- Logistic regression
- Uses log function fit the data between value of 0 to 1
- Where the value of the output is easy for classify either yes or no , good or bad , on or off

```
[21]: # Using Default Tokenizer
      vectorizer = TfidfVectorizer()
      # Store vectors into X variable as Our XFeatures
      X = vectorizer.fit_transform(url_list)
      X_train, X_test, y_train, y_test = train_test_split
      # Model Building

      logit = LogisticRegression(max_iter=5000)  #using
      logit.fit(X_train, y_train)

[21]: LogisticRegression(C=1.0, class_weight=None, dual=F
                         intercept_scaling=1, l1_ratio=Nc
                         multi_class='auto', n_jobs=None,
                         random_state=None, solver='lbfgs
                         warm_start=False)

[22]: # Accuracy of Our Model with our Custom Token
      print("Accuracy ",logit.score(X_test, y_test))

      Accuracy  0.9586791797672106
```

Code (Decision tree)

- Decision tree
- Supervised learning algorithm
- Used for solving Regression and classification problem
- The decision tree train model used to predict the class of the model
- Classification is labeled

```
In [23]: classifier = DecisionTreeClassifier(random_state=5)
         classifier.fit(X_train, y_train)
         print("Accuracy ",classifier.score(X_test, y_test))

         Accuracy  0.9623084862985354
```

```
In [24]: X_predict = ["https://www.youtube.com//",
         "google.com/search=faizanahmad",
         "pakistanifacebookforever.com/getpassword.php/",
         "www.radsport-voggel.de/wp-admin/includes/log.exe",
         "ahrenhei.without-transfer.ru/nethost.exe ",
         "www.itidea.it/centroesteticosothys/img/_notes/gum.exe"]
         X_predict = vectorizer.transform(X_predict)
         New_predict = classifier.predict(X_predict)
         print(New_predict)

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

```
In [25]: # https://db.aa419.org/fakebankslist.php
         X_predict1 = ["https://www.youtube.com/",
         "https://vit.ac.in/",
         "http://www.arielmarine.net ",
         "www.silkroadmeds-onlinepharmacy.com" ]
         X_predict1 = vectorizer.transform(X_predict1)
         New_predict1 = classifier.predict(X_predict1)
         print(New_predict1)

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

Code(Random forest)

- Random forest is a supervised learning algorithm which is used for both classification as well as regression.
- It is unexcelled in accuracy among current algorithms.
- It runs efficiently on large data bases.
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification

```
In [26]: clf = RandomForestClassifier(n_estimators=20, random_state=5)
         clf.fit(X_train, y_train)
         print("Accuracy",clf.score(X_test, y_test))

         Accuracy 0.9699761692978742

In [27]: X_predict = ["https://www.youtube.com//",
         "google.com/search=faizanahmad",
         "pakistanifacebookforever.com/getpassword.php/",
         "www.radsport-voggel.de/wp-admin/includes/log.exe",
         "ahrenhei.without-transfer.ru/nethost.exe ",
         "www.itidea.it/centroesteticosothys/img/_notes/gum.exe"]
         X_predict = vectorizer.transform(X_predict)
         New_predict = clf.predict(X_predict)
         print(New_predict)

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

In [28]: # https://db.aa419.org/fakebankslist.php
         X_predict1 = ["https://www.youtube.com/",
         "https://vit.ac.in/",
         "http://www.arielmarine.net ",
         "www.silkroadmeds-onlinepharmacy.com" ]
         X_predict1 = vectorizer.transform(X_predict1)
         New_predict1 = clf.predict(X_predict1)
         print(New_predict1)

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

Code (Multilayer perceptron)

- A multilayer perceptron is a class of feedforward artificial neural network its term MLP
- MLPClassifier trains iteratively since at each time step the partial derivatives of the loss function with respect to the model parameters are computed to update the parameters.
- It can also have a regularization term added to the loss function that shrinks model parameters to prevent overfitting.
- This implementation works with data represented as dense NumPy arrays or sparse SciPy arrays of floating-point values.

```
) = MLPClassifier(hidden_layer_sizes=(20, 3), max_iter=50, alpha
).fit(X_train, y_train.values.ravel())
int("Accuracy",mlp.score(X_test, y_test))
```

```
predict = ["https://www.youtube.com//",
oogle.com/search=faizanahmad",
akistanifacebookforever.com/getpassword.php/",
ww.radsport-voggel.de/wp-admin/includes/log.exe",
renhei.without-transfer.ru/nethost.exe ",
ww.itidea.it/centroesteticosothys/img/_notes/gum.exe"]
predict = vectorizer.transform(X_predict)
_predict = mlp.predict(X_predict)
int(New_predict)
```

```
ttps://db.aa419.org/fakebankslist.php
predict1 = ["https://www.youtube.com/",
ttps://vit.ac.in/",
ttp://www.arielmarine.net ",
ww.silkroadmeds-onlinepharmacy.com" ]
predict1 = vectorizer.transform(X_predict1)
_predict1 = mlp.predict(X_predict1)
int(New_predict1)
```

Result

- Here the result of the data analysis using different algorithm
- Where our first result logistic regression gives result with accuracy of 93.5%
- Decision tree algorithm produce and output accuracy of 95%
- Random forest accuracy output is produced by the algorithm 97%
- Where least accuracy in the multilayer perceptron which have only 82.5% of accuracy

Result output

- Predict outcome of the unknow variable data
- The data is given from any website
- Which is given to Predict variable then the outcome is predict by the algorithm

```
7]: X_predict = ["https://www.youtube.com//",
"google.com/search=faizanahmad",
"pakistanifacebookforever.com/getpassword.php/",
"www.radsport-voggel.de/wp-admin/includes/log.exe",
"ahrenhei.without-transfer.ru/nethost.exe ",
"www.itidea.it/centroesteticosothys/img/_notes/gum.ex
X_predict = vectorizer.transform(X_predict)
New_predict = clf.predict(X_predict)
print(New_predict)
```

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

```
8]: # https://db.aa419.org/fakebankslist.php
X_predict1 = ["https://www.youtube.com/",
"https://vit.ac.in/",
"http://www.arielmarine.net ",
"www.silkroadmeds-onlinepharmacy.com" ]
X_predict1 = vectorizer.transform(X_predict1)
New_predict1 = clf.predict(X_predict1)
print(New_predict1)
```

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

Result Table

Algorithm	Mean	Standard	Accuracy
Logistic Regression	0.9372	0.0017	94.5%
Decision Tree	0.9505	0.0018	95.2%
Random Forest	0.9601	0.0023	96.8%
Multilayer perceptron	0.8625	0.0047	83.5%