International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Phishing Website Detection

Sachin Nandrajog[1], Niraj Sawant[2], Dhruv Somvanshi[3], Dr. Narendra Shekokar[4], Mrs. Neha Ram[5]

[1, 2, 3] *Student, Department of Information Technology, D. J. Sanghvi College of Engineering, Mumbai, India.*
[4]*Group Guide, Department of Computer Engineering, D. J. Sanghvi College of Engineering, Mumbai, India.*
[5]*Group Guide, Department of Information Technology, D. J. Sanghvi College of Engineering, Mumbai, India.*

Abstract - *A phishing attempt is a fraudulent email or website that tricks you into giving out personal information like passwords, usernames, and credit/debit card details. The phishing site will appear just as legitimate sites do but instead directs the user to a page where they are asked for their sensitive info. There's no way of telling which pages on the web aren't legit until it's too late so make sure your browser has anti-phishing features. Our method predicts the phishing attacks by identifying certain features and also gives a maximum accuracy. This method is used to detect phishing site URLs that have these specific characteristics, which we identified in order to keep up with our times. Our system accurately predicts URL based phishing attacks while also being able to protect the user's identity. There are many tools and algorithms that can help in making decisions, predictions, etc., one of which is machine learning. The algorithm allows for better decision-making with less risk than other alternatives such as prediction markets or expert systems because it creates a model from historical data without bias on what will happen next - this makes outcomes more accurate overall. Our system is designed to detect phishing attacks by using different machine learning algorithms. One of the methods being used, a hybrid algorithm approach combining multiple other ML algorithms increases accuracy and detection rates while still maintaining its simplicity as compared with more complex models.*

Key Words: Phishing, Machine Learning, Algorithms, URL, Legitimate.

1. INTRODUCTION

Don't fall for phishing scams! Phishers use a technique called 'phishing' to make it appear as if they're coming from the original source. They imitate all of the characteristics and features of an email, so that you think this message has come from your bank or another trusted company but in reality, these emails are trying to get malicious links clicked on by tricking users into visiting their website. Remember - never click on any suspicious looking link found in emails unless you know who sent them. Phishing websites are a popular way for hackers to steal personal information. These phishers create convincing copies of real company's website and use alarming messages or validation requests in order to get people's private data, which they can then misuse. Phishing is often done with the goal of obtaining sensitive information such as account numbers, credit card and banking details. Phishers make their websites look legitimate to ensure that unsuspecting users will choose it over other options like Google or Yahoo! The user's input closely resembles what they see on a standard website window so there are no clues indicating this site could be dangerous. Users don't realize these sites aren't actually from well-known companies until after submitting personal data about themselves which then goes into the hands of criminals who use them for malicious purposes including identity theft, phishing scams and more. [8]. Phishing is a type of cybercrime, where attackers try to gain access to your valuable data using email phishing. Phishers often use less cost and effort than other types of hackers because they are after the more precious information that you have on these emails. And this leads down various paths - from malware infections all the way up to identity theft or loss of personal data. The data which attackers want are password, OTP, credit/ debit card numbers CVV. These criminals also gather information that can give them access to social media accounts and email addresses. [3]

Phishing is a big problem for businesses. One way to detect phishers' tricks and traps before we go in too deep, so that our company doesn't get any of the information they're trying to steal from us, is through using software or approaches with algorithms. These are used at academic as well as commercial levels - let's take an example:

- While malignant URLs have many features, which separate them from regular URLs like lengthiness of name and confusing web address (URL), where on one hand it can be very long & difficult to read but also look legitimate while being quite hard for humans not trained in this area to spot differences & tell if something looks off; on the other end there might be shorter names. They use IP address rather than domain name. They also use a shorter domain name which is not close to original name.

We can use the labelled data in the training phase which there are samples of a legitimate domain and phishing area to be able to improve detection accuracy. We should only use websites whose classes are known, which means that we label as phishing those pages whom we know were detected as phishing and legitimate URLs will be classified as such. There are many types of algorithms, each has their own way of working. The existing systems uses one machine learning algorithm to detect illegitimate website. The current algorithm is not as accurate as we would hope given how important this topic is to us all. This approach isn't

improving prediction accurateness enough. Individual algorithms are not as accurate in detecting phishing websites so a new approach is required which is by using Hybrid algorithm.

2. LITERATURE SURVEY

In this section, we will learn about the different classifiers used to predict phishing with the help of machine learning algorithms. We will also discuss our proposed approach to detect a phishing website - so you don't get scammed. In the next section we will explain our proposed system.

2.1 Machine Learning Algorithms and Methods to Detect the Phishing Website

If you have been made a victim of phishing, it might be difficult for some to identify the malicious websites. Machine learning has helped in many areas such as email and website detection but there is still room for improvement with other methods like malware or SMS text messages. Our focus is on detecting website phishing (URL). Using Hybrid Algorithm, we combine different algorithms to give better prediction and accuracy. There are different applications for data analysis in general, we cannot say which algorithm is better or worse. Tree classifiers have been around for a long time, but they don't work well with the vast amount of data we generate today. Here are some solutions that could make this old classification method more effective. [5]

Naive Bayes Classifier: This classification technique is also called a Generative Learning Model. The outcome of the classifier here relies on Bayes' theorem, which assumes that there are independent predictors for each instance and features within these instances do not affect one another's likelihood to be true or false. In simple words, this means that regardless of whether certain traits exist in an object/individual- it will still have the same probability as any other trait existing without taking into account all related properties at once because they're assumed to be unrelated. When each feature in a dataset interacts with another or is dependent on other features, it is considered as independent contributions to the probability of an event happening. Naïve Bayes can classify large datasets and is easy to use.[14]

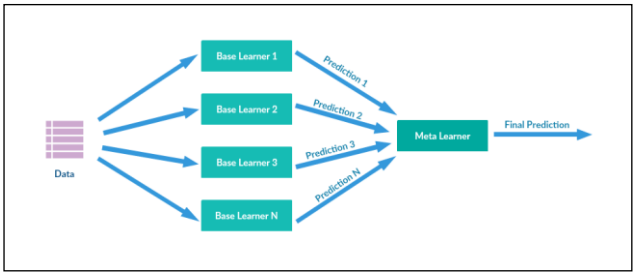
Random Forest: Random forest operates similarly to the combined learning method of classification, like regression and other tasks. It is done by building a group of decision trees at training data level during output for individual classes or prediction in regression mode. Random forest accuracy for decision trees practice of overfitting the training data set. [8][14]

Support vector machine (SVM): SVM comes under the category of supervised learning and is easy to use. It can be used for both classification and regression applications, but it's more famous as a tool in classifications. . In this algorithm each point which is a data item is plotted in a dimensional space, this space is also known as n dimensional plane, where the 'n' represents the number of features of the data. The classification is done based on the differentiation in the classes, these classes are data set points present in different planes.

XGBoost: XGBoost has been recently discovered by researchers and it is proving to be very useful for machine learning classification. Its fast execution of boosted decision trees makes its performance better than most other algorithms on the market today. The output model improves the performance of models and also speeds up computation. [21]

Once the model is trained it's important to evaluate and validate your classifier so you can be sure that it will work. We've seen all advantages and disadvantages of each type. Hence we recommend to use more than one algorithm that is, if you can get two good ones. You should try combining them so they complement each other's strengths and weaknesses. This will make your predictions stronger because the models are checking for different things when making a prediction. If possible, this combination could go beyond just using Naive Bayes or Random Forest (both excellent algorithms), by adding another algorithm into the mix who has complementary qualities with both aforementioned algorithms- take care not to overwhelm yourself though; start small. After the processing is done by the algorithms the result is generated and the website classifies as phishing website or a legitimate website.

Fig -1: Stacking Technique Design



Stacking: Stacking is a technique that combines multiple classification models and provides a better output. This means that there are multiple different learners and these learners gives us an output. The learners at stage 1 are called Base Learners. The output of Base Learners (Intermediate prediction) is then given to a Meta Learner. Meta learner then identifies the mistakes of the base learner and then finally makes a prediction that is our final output. In our case, the base learners are Naïve Bayes and SVM and the Meta learner is the Random Forest algorithm.

2.2 Machine Learning Algorithms and Methods to Detect the Phishing Website

	Decision Trees	Neural Networks	Naïve Bayes	kNN	SVM
Accuracy in general	**	***	*	**	****
Speed of learning with respect to number of attributes and the number of instances	***	*	****	****	*
Speed of classification	****	****	****	*	****
Tolerance to missing values	***	*	****	*	**
Tolerance to irrelevant attributes	***	*	**	**	****
Tolerance to redundant attributes	**	**	*	**	***
Tolerance to highly interdependent attributes (e.g. parity problems)	**	***	*	*	***
Dealing with discrete/binary/continuous attributes	****	***(not discrete)	***(not continuous)	***(not directly discrete)	**(not discrete)
Tolerance to noise	**	**	***	*	**
Dealing with danger of overfitting	**	*	***	***	**
Attempts for incremental learning	**	***	****	****	**
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*
Model parameter handling	***	*	****	***	*

Fig -2: Algorithms Comparison

2.3 Proposed System

We feed our dataset that contains phishing and legitimate URLs to the system which then pre-processes the data so that it can be analyzed. The features have around 30 characteristics of phish websites, each with their own set attributes and values that are defined for them. We extract the features from the URL and identify input range values. These values identify the risk level assigned to each phishing website, providing an accurate representation of how dangerous it maybe. The output for each input is a range of numbers between 0-100. If the attribute has been successfully phished or not, this will be represented with binary code on whether it was present as no 1 and no 0 which indicates its presence or absence respectively.

After the features are extracted and the values are assigned to it, we apply appropriate machine learning algorithm to it. The algorithms will be explained in more detail later on but are already discussed at length before now. We use a hybrid classification, combining two classifiers, Naive Bayes and Random forest to predict an accurate detection for phishing URLs. This method of testing data is called a hybrid approach, in this process we recommend using the combination of two classifiers. We will then test our output and evaluate how accurate it was against previous systems.

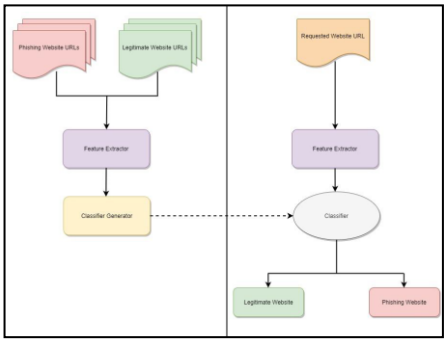


Fig -3: Proposed Architecture

The training phase is extremely important because it provides a chance to not only label data samples, but also practice classifying them. Labelled data can be used in this stage of the process and will help with classification by utilizing its sample phish area and legitimate area. We should use samples whose classes are known to us, which means we will only give our labelling of phishing and legitimate URL during this process. The dataset to be used must actually consist of these features. There are many different types of algorithms, each having its own way of working. So, it's important to understand the requirements that every individual algorithm comes with. Since each algorithm has some disadvantages, it is not recommended to use individual algorithm in order to classify phishing websites because this would be more time consuming as well as costly. [10]

3. CONCLUSION

We have learned that phishing attacks are very dangerous and it is important for us to detect them and be safe. As if the user's personal information can be leaked through these websites, we need more of an effort put into this problem in order to stop anything from happening further down the line. One easy solution would involve any machine learning algorithm with general classifiers as they could help monitor those sites much better than before-hand; whereas before not many people were really looking at what was going on there anyways so now you're able find out everything about someone just by clicking around online. There exist classifiers which gives us good prediction rate of phishing attacks. However, after our survey and research we've found that it would be better to use hybrid approach for predicting these types of websites in order to further improve accuracy rates on their predictions. After our survey, we have seen that the existing systems gives less accuracy because of using single machine learning algorithms. So, we propose a new method that uses URL based features and we generate classifiers through several machine learning algorithms. We have found that our system provides us with 70.18% accuracy for Naïve Bayes algorithm, 89.44% accuracy for Random Forest algorithm and finally 91.43% of accuracy when using Hybrid algorithm. Hence, we can conclude by saying that Hybrid algorithm is recommended while detecting phishing attacks because of its high accuracy. Our

approach is also much secure as it detects previous and new phishing websites.

REFERENCES

[1] Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.
[2] Rao, R. S., & Pais, A. R. (2019). Jail-Phish: An improved search engine based phishing detection system. Computers & Security, 83, 246-267.
[3] Ding, Y., Luktarhan, N., Li, K., & Slamu, W. (2019). A keyword-based combination approach for detecting phishing webpages. Computers & security, 84, 256-275.
[4] Marchal, S., Saari, K., Singh, N., & Asokan, N. (2016, June). Know your phish: Novel techniques for detecting phishing sites and their targets. In 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS) (pp. 323-333). IEEE.
[5] Shekokar, N. M., Shah, C., Mahajan, M., & Rachh, S. (2015). An ideal approach for detection and prevention of phishing attacks. Procedia Computer Science, 49, 82-91.
[6] Rathod, J., & Nandy, D. Anti-Phishing Technique to Detect URL Obfuscation.
[7] Hodžić, A., Kevrić, J., & Karadag, A. (2016). Comparison of machine learning techniques in phishing website classification. In International Conference on Economic and Social Studies (ICESoS'16) (pp. 249-256).
[8] Pujara, P., & Chaudhari, M. B. (2018). Phishing Website Detection using Machine Learning: A Review.
[9] Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017, May). Malicious web content detection using machine leaning. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1432-1436). IEEE.
[10] Lakshmi, V. S., & Vijaya, M. S. (2012). Efficient prediction of phishing websites using supervised learning algorithms. Procedia Engineering, 30, 798-805.
[11] Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In Cyber Security (pp. 467-474). Springer, Singapore.
[12] Kazemian, H. B., & Ahmed, S. (2015). Comparisons of machine learning techniques for detecting malicious webpages. Expert Systems with Applications, 42(3), 1166-1177.
[13] Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2019). Phishing page detection via learning classifiers from page layout feature. EURASIP Journal on Wireless Communications and Networking, 2019(1), 43.
[14] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2012, December). An assessment of features related to phishing websites using an automated technique. In 2012 International Conference for Internet Technology and Secured Transactions (pp. 492-497). IEEE.
[15]https://www.researchgate.net/publication/226420039-Detection-ofPhishing-Attacks-A-Machine-Learning-Approach
[16]https://www.proofpoint.com/us/threat-reference/phishing
[17]https://towardsdatascience.com/phishing-domain-detection-with-ml5be9c99293e5
[18]https://en.wikipedia.org/wiki/Phishing
[19]https://www.techrepublic.com/article/how-to-tackle-phishing-withmachine-learning/
[20]https://www.irjet.net/archives/V5/i3/IRJET-V5I3580.pdf
[21]https://www.hackerearth.com/practice/machine-learning/machine-learningalgorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/
[22]
https://www.datacamp.com/community/tutorials/svm-classification-scikitlearn-python
[23] He, M., Horng, S. J., Fan, P., Khan, M. K., Run, R. S., Lai, J. L., ... & Sutanto, A. (2011). An efficient phishing webpage detector. Expert systems with applications, 38(10), 12018-12027.
[24] Le, A., Markopoulou, A., & Faloutsos, M. (2011, April). Phishdef: Url names say it all. In 2011 Proceedings IEEE INFOCOM (pp. 191-195). IEEE.
[25] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345-357.
[26] Tewari, A., Jain, A. K., & Gupta, B. B. (2016). Recent survey of various defense mechanisms against phishing attacks. Journal of Information Privacy and Security, 12(1), 3-13.
[27] Jain, A. K., & Gupta, B. B. (2016, March). Comparative analysis of features based machine learning approaches for phishing detection. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 2125-2130). IEEE.
[28] Yuan, H., Chen, X., Li, Y., Yang, Z., & Liu, W. (2018, August). Detecting Phishing Websites and Targets Based on URLs and Webpage Links. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 3669-3674). IEEE.
[29] Nguyen, L. A. T., To, B. L., Nguyen, H. K., & Nguyen, M. H. (2013, October). Detecting phishing web sites: A heuristic URL-based approach. In 2013 International Conference on Advanced Technologies for Communications (ATC 2013) (pp. 597-602). IEEE.