Learning-based models to detect runtime phishing activities using URLs

Surya Srikar Sirigineedi
Florida International University
Miami, FL, USA
ssiri005@fiu.edu

Jayesh Soni
Florida International University
Miami, FL, USA
jsoni@fiu.edu

Himanshu Upadhyay
Florida International University
Miami, FL, USA
upadhyay@fiu.edu

ABSTRACT

Phishing websites are fraudulent sites that impersonate a trusted party to gain access to sensitive information of an individual person or organization. Traditionally, phishing website detection is done through the usage of blacklist databases. However, due to the current, rapid development of global networking and communication technologies, there are numerous websites and it has become difficult to classify based on traditional methods since new websites are created every second. In this paper, we are proposing a real-time, anti-phishing system. In the first step, we extract the lexical and host-based properties of a website. In the second step, we combine URL (Uniform Resource Locator) features, NLP and host-based properties to train the machine learning and deep learning models. Our detection model is able to detect phishing URLs with a detection rate of 94.89%.

CCS Concepts

Security and privacy→Phishing

Keywords

Phishing URLs; Machine learning; NLP; Deep Learning

1. INTRODUCTION

The web has become a platform for supporting a wide range of cyber-crimes such as financial fraud (e.g., via phishing), as a way to inject malware (e.g., via drive-by download) and cybercrime (e.g., via stealing identities). Cyber-crimes have been steadily on the rise over the past two decades due to the rapid development of the internet. According to the U.K. Department for Digital, Culture, Media & Sport 2019, around 32 percent of organizations had cybersecurity breaches in 2018. When compared with 2017 (43 percent) it went down by 11 percent. They are still a major threat to organizations as cyber-attacks cost millions of dollars to organizations, and sometimes the reputation of the company may be at risk. Common types of cyber-attacks are phishing (80%), viruses, spyware, malware, ransomware attacks (27%) and impersonations (28%) [1].

Phishing is a type of social engineering attack used to steal a user's personal information involving the manipulation of people who have less knowledge about these types of attacks by impersonating a trusted third party website. In these kinds of attacks, the attackers bait the end-user by sending emails which seem like they are sent by legitimate banks in order to disclose the victim's bank details for financial gain, and redirecting to impersonated websites to fill their personal details for identity theft. Most of the time, the attacker tries to gather victim sensitive data like their social security number, credit card number, password, bank account number, etc.

If we are able to classify the URL before a user opens it in the browser, we can avoid most of these attacks that cause problems. Currently, it is difficult to achieve this by using the traditional blacklist approach, since most of the phishing URLs are too new or never properly validated. Therefore, by using our model running on the client-side, we can classify the URL as malicious or benign.

In this paper, we've focused on analyzing the information of a URL by extracting significant features from it and training a prediction model on training data of both the legitimate and fraudulent URLs. For this experiment, we have used 36,400 legitimate and 37,175 phishing URLs to perform our analysis.

The rest of the paper is as follows. Section 2 describes the related work. Section 3 provides a feature extraction technique. Section 4 gives a high-level overview of different learning models. Section 5 describes the performance metrics used to evaluate the trained model. Section 6 provides a description of the Analytics platform. Experimental results are discussed in Section 7 and finally, we conclude in Section 8.

2. RELATED WORK

URL is the global address of a resource on the World Wide Web. A URL has three main components: a) Protocol: Whenever there is a need to exchange data among the computer on network or internet, they need to follow the set of rules. E.g., TCP/IP, HTTP, HTTPS, FTP. b) Hostname: A hostname is a domain name that has at least one associated IP address. Every hostname starts with Subdomain and Second Level Domain (SLD) which refers to the organization name. Ends with Top Level Domain (TLD) like gov, edu, org, net, and int. as shown in Figure 1.

ICCDA 2020, March 9–12, 2020, Silicon Valley, CA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7644-0/20/03…$15.00
https://doi.org/10.1145/3388142.3388170

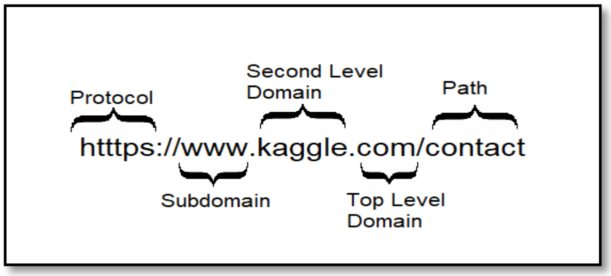


Figure 1. Example of a URL

Le, Markopoulou et al. [2] uses attributes of URLs such as directory, filename etc to identify phishing websites. For offline classification, the author uses support vector machines whereas confidence weighted and online perceptron were used for online classification. Results show that the use of adaptive regularization increases the detection rate. Traditional based and learning-based models are two different types of algorithms that are used for phishing detection. The traditional based algorithm is able to classify the known attacks whereas learning-based models learn the behavior or patterns of the URLs and thus able to detect even unknown attacks with higher detection rates [3][4]. Jain and Gupta [5] extracted 19 client-side features to discriminate between legitimate and phishing websites. They train their machine learning model on the PhishTank and Openfish datasets. Their proposed approach gives a higher true positive rate. Feng et al. [6] use the Monte Carlo algorithm for detection. With 30 different features extracted from URLs, they achieved a significant accuracy rate. Buber et al. [7] proposed a phishing detection system with 209-*word vector* features and 17 *NLP based* features and compares three different machine learning algorithms by increasing the number of NLP vectors. In Rao & Pais [8], authors use a hybrid method which comprises of machine learning approaches and image checking. Features used were hyperlink-based, third-party based, and URL based features. In recent years, deep learning has shown higher performance by automatically extracting features[9][10]. Mohammad, Thabtah, and McCluskey [11] uses adaptive self-structuring neural networks for detecting phishing URLs. The model needs much more time since it is dependent on third-party services. However, it produces high accuracy rates even when trained using a small dataset.

In our approach, we extract NLP based, Lexical based and Host-based features. Among them, host-based and NLP based features have a significant impact on the detection rate. We applied several machine learning and deep learning algorithms.

3. FEATURE EXTRACTION

We categorize the features that we extract from URLs into lexical, NLP and host-based.

Lexical features: In most of the cases, a URL can be classified as a phishing URL based on the lexical features. These features include the length of the domain and the subdomain. Usually, an attacker prefers to have a long domain to make the URL unsuspected by the victim. We can also include special characters like '/', '?', '.', '=', '-' and '_' being used in URL which are real numbers.

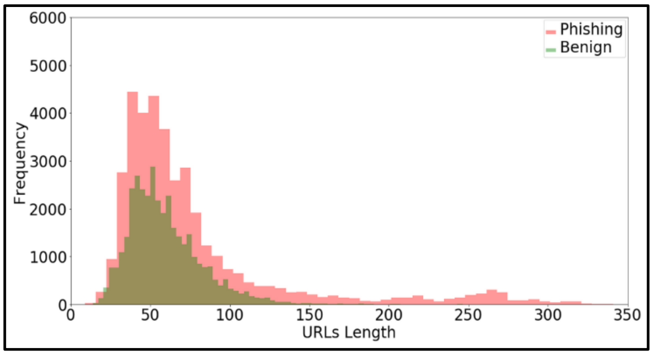


Figure 2. Distribution of URLs Length

NLP features: NLP has been used in classifying the spam emails based on semantic analysis of the message in email. We are using NLP to look for random characters, combined word usage, typosquatting, cybersquatting, etc. The main aim of this part is to detect the words, which are similar to known brand names, random words, and keywords [12].

Host-based Features: The properties here describe more of the domain properties. We borrowed this idea from [13] and used the WHOIS domain properties such as:

1. Creation date, the date when the domain name actually created.
2. Expiration date, the date when the domain is getting expired
3. Registrar, the registrar name.

The hostname properties have a significant role in classifying URLs. Most of the phishing URLs have a short-lived lifespan. A large number of phishing websites contain IP addresses in their hostname. Therefore, having these kinds of features will help us in identifying the website as phishing or benign.

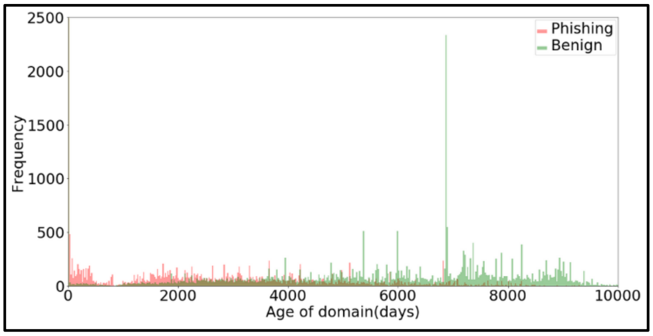


Figure 3. Distribution of Age of Domain in Days

From Fig. 3, we notice that the histogram distribution of the phishing URLs has a short timespan and benign one has a longer timespan.

Fig. 4, shows the plot between registrar and number of websites registered under that registrar. There are very few registrars like MarkMonitor Inc which host few number of phishing websites.

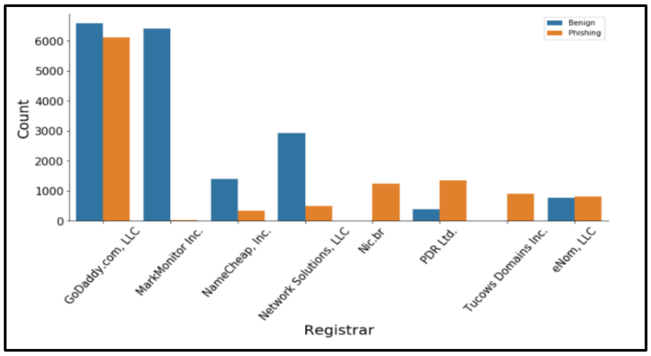


Figure 4. Registrar Domain and Count

4. LEARNING MODELS

We applied different machine learning and deep learning classification algorithms available in scikit-learn and Tensor flow by using Jupiter notebook, Python programing language, and Analytics Platform.

4.1 Machine Learning Algorithms

The six machine learning algorithms that were considered for model building are:

1. K-Nearest Neighbors: KNN is a classification algorithm that simply stores all the available class data points and classifies the new data points based on the distance similarity [14].
2. Logistic Regression: Logistic Regression uses the sigmoid function to classify data and works well when one dependent binary variable and one or more nominal, ordinal independent variables are present in the dataset.
3. Support Vector Machines: Support Vector Machines (SVM) can be used for classification and regression analysis [15]. SVM looks at extreme points in the given dataset and draws a hyperplane that segregates the given number of classes.
4. Gradient Boosting Classifier: Gradient Boosting is a machine learning technique for regression and classification problems, use with the ensemble method, which groups the weak machine learning models like decision tree into a prediction model which typically produces good results [16]. In Gradient Boosting the performance of the model is improved by iteratively identifying the points that are being misclassified by large residuals computed in the previous iterations.
5. Ada-boost classifier: Ada-boost is another type of classifier that combines the weak classifier models into a strong model. In Ada-boost, the performance of the model is improved by iteratively up-weighting the points that are being misclassified before.
6. Random Forest Classifier: Random forests or random decision forests are an ensemble learning method for classification and regression, which are built using multiple decision trees [17][18].

4.2 Deep Learning Algorithms

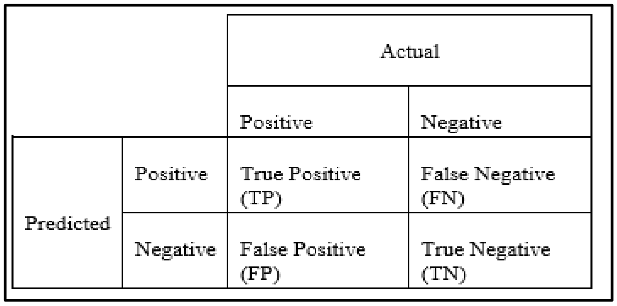
A neural network is a series of neural layers which learn the relationship between features and target label by mimicking the human brain. It has following hyperparameters:

1) Batch size: It is used to specifies the number of data points to be used at a time to train the model before updating the model weights.
2) Epoch: It is the number of times that the neural network will get trained on the entire dataset.
3) Learning rate: It controls how much to change the model in response to the estimated error, each time the model weights are updated.

5. PERFORMANCE METRICS

In our experiments, we use different metrics to evaluate the performance: Confusion Matrix, Accuracy, F1- Score and Receiver Operating Characteristic (ROC).

The confusion matrix is used to describe the performance of an algorithm.



		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 5. Confusion Matrix

Where TP represents the number of legitimate URLs misclassified into legitimate ones, and FN represents the number of phishing URL's misclassified to legitimate URLs, and FP represents the number of legitimate URLs that are misclassified into phishing ones, and TN represents the number of legitimate URLs correctly classified as legitimate ones.

FPR is the percentage of legitimate URLs that are misclassified as phishing URLs.

$$FPR = FP / (TN + FP) \quad (1)$$

TPR is the percentage of phishing URLs that are correctly classified as phishing URLs.

$$TPR = TP / (FN + TP) \quad (2)$$

ROC: It is a plot of the true positive rate versus false-positive rate for several diverse candidate threshold values between 0.0 and 1.0.

Accuracy refers to the percentage of URLs that are correctly classified, including the phishing and legitimate URLs to the total number of URLs used for the experiment.

$$Accuracy = (TP + FN) / (FN + TP + FP + TN) \quad (3)$$

Precision refers to the percentage of URLs that are correctly classified as phishing URLs to the sum of correctly classified phishing URLs and the number of legitimate URLs that are misclassified into phishing ones.

$$Precision = TP / (TP + FP) \quad (4)$$

Recall refers to the percentage of URLs that are correctly classified phishing URLs to the sum of correctly classified phishing URLs and the number of phishing URLs that are misclassified into legitimate ones.

$$\text{Recall} = TP / (TP + FN) \quad (5)$$

F1 score formula is represented below:

$$\text{F1 score} = 2. (\text{Precision. recall}) / (\text{Precision} + \text{recall}) \quad (6)$$

6. ANALYTICS PLATFORM

Analytics Platform is an in-house client application framework which is used to preprocess data and further train the machine learning and deep learning algorithms. The frontend of this framework is built on Windows Presentation Foundation and Microsoft SQL Server integrated with Machine learning server is used for backend services to train the model.

7. EXPERIMENTAL RESULTS

7.1 Dataset

We used the URL dataset available on [19]. In total, we have a dataset of 73,575 URLs of which 36,400 legitimate and 37,175 phishing URLs. For the experiment, we used 51,502 URLs samples for training the model and the remaining 22,073 URLs samples for testing the model.

Table I. Data Set Description

	Phishing	**Benin**
Train	26107	25395
Test	11068	11005

7.2 Result Analysis of Machine Learning Model

We trained the different machine learning algorithms like K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Gradient Boosting Classifier (GBC), Ada Boost Classifier (ABC), Random Forest classifier (RFC). The performance of different algorithms is recorded in Table II.

Table II. Performance Of The Classifiers

Algorithm	**Accuracy**	**Precision**	**F1**	**ROC**
KNN	92.5882	92.5882	92.5882	92.5896
LR	89.1949	89.1949	89.1949	89.1885
SVM	87.7860	87.7860	87.7860	87.7743
GBC	92.0944	92.0944	92.0944	92.0891
ABC	***94.7221***	***94.7221***	***94.7221***	***94.7201***
RFC	88.6332	88.6332	88.6332	88.6215

From Table II, we observe that the ensemble model, Gradient Boosting Classifier, Ada Boost Classifier gives better results when compared with other classification algorithms like K-Nearest Neighbor (KNN), Logistic Regression (LR), and Support Vector Machine (SVM).

7.3 Result Analysis of Deep Learning Model

We used a fully connected dense network with 4 dense layers. The model was trained with 100 epochs and a 0.0001 learning rate. We tested the same deep learning model architecture with different batch sizes: 32, 64, 128, 256 and 512.

Our trained model gives an accuracy of 96.6% with a batch size of 128 as shown in figure 6.

The performance (Accuracy and Loss) of a trained neural network w.r.t batch size is shown in Table III.

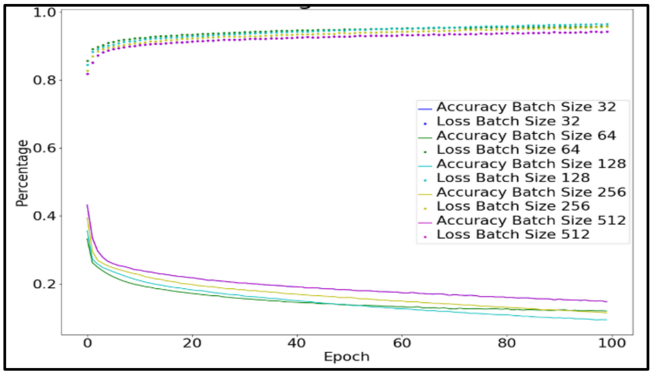


Figure 6. Training Performance of the Model

Table III. Accuracy and Loss w.r.t Batch Size

Batch Size	**Accuracy**	**Loss**
32	94.89	0.1547
64	95.91	0.1195
128	96.6	0.0940
256	95.75	0.1142
512	94.34	0.1472

8. CONCLUSION

In this paper, we have implemented a phishing detection system based on URLs. The efficient feature list is crucial for any detection system to be accurate. We extracted three types of features for phishing detection: host-based, NLP based and lexical based features. We used several learning algorithms such as KNN, LR, SVM, GBC, ABC, RFC, and NNs. From the experiment, we found that the Neural Network gives a higher detection rate with an accuracy of 94.89 and binary cross-entropy loss of 15.47 compared to other learning-based models.

9. ACKNOWLEDGMENT

This research was supported by the Department of Defense (DoD) - Test Resource Management Center (TRMC), USA.

10. REFERENCES

[1] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/813599/Cyber_Security_Breaches_Survey_2019_-_Main_Report.pdf

[2] Le, A., Markopoulou, A.,&Faloutsos, M.(2011).Phishdef: URL names say it all. In 2011 Proceedings IEEE INFO COM, 2011(pp.191–195)

[3] Soni, J., Prabakar, N. (2018) "Effective Machine Learning Approach to Detect Groups of Fake Reviewers", Proceedings

of the 14th International Conference on Data Science (ICDATA'18), Las Vegas, NV, 2018.

[4] Soni, J., Prabakar, N. and Upadhyay, H. (2019) "Feature Extraction through Deepwalk on Weighted Graph", Proceedings of the 15th International Conference on Data Science (ICDATA'19), Las Vegas, NV, 2019.

[5] Jain,A.K.,&Gupta,B.B.(2018).Towards detection of phishing web sites on the client–side using a machine learning-based approach. Telecommunication Systems, 68(4),687—700.

[6] Feng, F., Zhou, Q. , Shen, Z. , Yang, X. , Han, L., & Wang, J. (2018). The application of a novel neural network in the detection of phishing websites. *Journal of Ambient Intelligence and Humanized Computing*.

[7] Buber,E.,Diri,B.,&Sahingoz,O.K.(2017a).Detecting phishing attacks from URL by using NLP techniques. In 2017 International conference on computer science and Engineering (UBMK) (pp.337–342).28

[8] Rao, R. S. &, & Pais, A. R. (2018). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications*.

[9] Jayesh Soni, Nagarajan Prabakar, Himanshu Upadhyay (2019) "Deep Learning approach to detect malicious attacks at system level". In WiSec'19: Proceedings of 12th ACM Conference on Security & Privacy in Wireless and Mobile Networks, May 15-17, 2019, Miami, FL, USA.

[10] Soni, J., Prabakar, N. and Kim, J-H. (2017) "Prediction of Component Failures of Telepresence Robot with Temporal Data ". 30th Florida Conference on Recent Advances in Robotics.

[11] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting phishing web- sites based on self-structuring neural network. *Neural Computing and Applications, 25* (2), 443–458.

[12] O.K. Sahingoz, E. Buber, O. Demir, B. Diri, Machine learning-based phishing detection from URLs, Expert Syst. Appl. 117 (2019) 345–357

[13] Garera S., Provos N., Chew M., Rubin A. D., "A Framework for Detection and measurement of phishing attacks", In Proceedings of the ACM Workshop on Rapid Malcode (WORM), Alexandria, VA.

[14] Altman, N. S. *(1992).* "An introduction to kernel and nearest-neighbor nonparametric regression" (PDF). *The American Statistician. 46 (3): 175–185.* DOI*:*10.1080/00031305.1992.10475879. HDL*:*1813/31637

[15] https://en.wikipedia.org/wiki/Support-vector_machine

[16] https://en.wikipedia.org/wiki/Gradient_boosting

[17] Ho, Tin Kam (1995). Random Decision Forests (PDF). *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from* the original (PDF) *on 17 April 2016*. Retrieved 5 June 2016.

[18] Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF)*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844.* doi*:*10.1109/34.709601.

[19] https://github.com/ebubekirbbr/pdd