IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 4, May 2023

Phishing Detection with Machine Learning Technology

Aishwarya Shahane[1], Shivam Shete[2], Manasi Nemane[3], Pravin Katkade[4], Dr. H. B. Jadhav[5]
Students, Computer Engineering[1,2,3,4]
Professor, Computer Engineering[5]
Adsul Technical Campus, Ahmednagar, Maharashtra, India

Abstract: *Today everyone is highly addicted to Internet. Everyone has done online shopping and online activities like online bank, online booking, online recharge and others on the Internet. Phishing is a type of web threat and phishing is Illegally on the original website Information such as login name, password and credit card information. This paper proposed an efficient phishing detection based on machine learning technique. Overall, the experimental results show that the proposed when integrated with the Support vector machine classifier, has the best performance in accurate discrimination 95.66% of phishing and suitable websites are used by only 22.5% innovative functionality. The proposed technique shows optimistic results when compared to a number of benchmarks "University of California Irvine (UCI)" phishing data files archives. Therefore, the proposed technique is preferred and used for machine learning-based phishing detection*

Keywords: phishing , web , machine learning , director component analysis, vector machine support.

I. INTRODUCTION

The web has become a platform for a wide variety of criminals Companies such as spam, financial fraud, proliferation of operators malware. There is a valid business reason for this plan It's different, but the common thread is a user need No need to visit their site. This visit links from emails, web inquiries, or other pages on the Site; However, the customer must tap to move. For example,Showing and retrieving the ideal URL (Uniform Resource Locator) Important data to overcome this, the security community must We responded by creating a packaged blacklist service

Provide toolbars, devices, search engines, alerts, or Alerts with accurate feedback. The site is very new Many harmful sites are uncategorized because they are uncategorized or miscategorized Blacklisted. Phishing is a type of cyber attack that exploits a site. Consistent shopper insights such as store card numbers, accounts and logins Qualifications are just the tip of the iceberg. In June 2018, "APWG (Anti-Phishing Working Group)" 51,401 suggest phishing sites [1]. According to another RSA According to reports, phishing incidents cost around $9 billion worldwide. According to 2016 statistics [2], traditional anti-phishing Options and efforts were fruitless. The most widely used anti-phishing solutions About existing common blacklist warning system A web browser such as Chrome, Internet Explorer, Mozilla Firefox. Blacklist interrogation tools include: A database of possible phishing URLs, which lead to Discover newly launched phishing websites [3, 4].Reliability of machine learning based phishing detection tools It is efficient in terms of accuracy. Most anti-phishing researchers are looking for new optimizations. Feature suggestions or classification algorithms. Appropriate functional analysis and development of selection techniques A design is not important [5, 6]. In [5], these 12 features. The sites are legitimate, phishable and effective The positive rate is 97% and the false positive rate is 4%. From Features include meta tagging, web page content,URL, Link, TF-IDF, Moore.

In general, there are two main techniques for feature selection Usage: filter size and packaging. On the other hand, measuring filters are indicators. Calculated from statistics, A useful theory that can reflect everyone's values Character functions other than calling exact classifiers. Rapper This technique is repeated with each run Generate a subset of elements and evaluate them Classification. When evaluating a set of features, Different iterations of packaging engineering scale exponentially It becomes practical computing for real-time applications [7, 8]. [9] describes an anti-phishing technique that removes 19 Ability to identify phishing sites on the buyer's side From verified sites using

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 4, May 2023

machine learning. They used 2141 [10] and [11] Phishing pages and famous Alexa websites, some online cash portals and some good banks website.

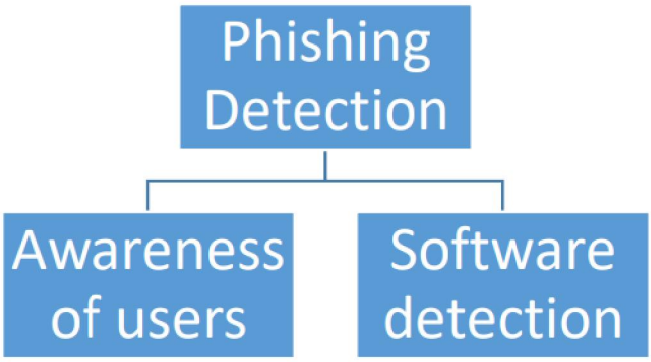

Fig. 1. Phishing detection methods

In [12], the use of probability minimization is proposed Standard and Monte Carlo algorithms with new neurals Network-based classification techniques for phishing detection A net score of 30 points was used for four classifications. The main area, especially the base of the rod, the base of the anomaly, HTML and JavaScript. In some studies, such as [11], the authors Hybrid technology for photo inspection Laptop mastery approach. The most essential defect Photo/visual phishing detection is required in advance Basic photo information (web history) database or network web page [13]. However, the proposed method This dependency is gone. I used natural Language Processing (NLP) is available in the literature. Hybrids [14] extracts features for URLs, text, and web. Content creation and use of extreme machine learning (ELM) Technology. The first step in this method is to write the text content of the classifier to determine the content of the label text In this case, use OCR software to retrieve the text. From the hybrid image based on the second stage. Combine text with other function classifiers. In [15], I proposed an approach to the construction of probability theory. neural network (PNN). Advantages of PNN high-speed train time, Paralysis to outliers, and generalizations are the best. However, since PNNs can significantly increase the data, High space and time as a result, author group K-medoids Use PNN to minimize training items. in [16] Anti-phishing technique in Iran's electronic banking system. From The author identifies 28 features used by attackers Fraud of Iranian banking sites. In the banking system of Iran The detection accuracy was 88%. This method is specific Developed to discover Iranian banking sites, it can only be filtered Phishing and legitimate websites of all kinds. Machine learning-based techniques usually rely on it Doubtful site performance and A certain set of functions [17]. Leading to, Accuracy of the system is a set of functions and The accuracy with which defenders select features [18,19]. In the present study [20], NLP was implemented in phishing. Run email detection to identify malicious targets Semantic analysis of email content (plain text). In this paper, we use a machine learning algorithm for detection Web page URL to identify phishing web pages. When Implement a machine learning algorithm that also It is important to extract features from the dataset. as a result, Google collects a large number of legitimate and malicious web pages A URL of an existing dataset. The effectiveness of the offer Systems are measured by functions defined by words take it.

II. PROPOSED TECHNIQUE

Figure 2 shows the proposal Phishing site detection technology using machines Learn the techniques.

a. Step 1

Automatic collection of web pages using GNU Wget, In addition to the entire HTML document, the Python script Also related resources (such as images, CSS, JavaScript) to enable the browser to render the entire web. Downloaded page also screenshots of all web pages It is stored for further inspection and filtering. Download data sets that are further processed and loaded Remove phishing and legitimate data collection or web pages. Registration is at this time A proposed technique for integrating basic functions. The website is stacked in two separate classes. Between January and May 2015 and May and June 2017. Specifically, we have 5,000 phishing web pages and Pages are especially stable based on URLs. Fish The repository is completely based on Alexa and Common Crawl URLs Archive.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 4, May 2023

B. Step 2

Vocabulary, host, word used to extract the feature vector From the input URL. Vocabulary features are textual features URL includes hostname size, URL length, tokens, etc. such as URLs. Easy math, security, accuracy Overclassification of Machine Learning Vocabulary Property. The vocabulary function is the URL text of the property Not the page itself, but itself. These properties include length hostname, full URL length, Points to URLs, hostnames (separated by ``), and binaries URL path ('/', '=', '?' ' '-'East' _ '). This is also called "pocket". Host-based features can explain "where the malicious sites are" Below are the hosts. The attribute is identified by the hostname as part of the URL. The words in the vector are especially favorable for execution An important specific attribute is the URL of the web page. That It mainly consists of text with many words. With the exception of that Text changes in this guide, automatic vectoring Excellent process each URL is converted into a carrier specific URL Words that use the Weka function called "StringtoWordVector".Once you have the corresponding vector, you can simply use it with the machine learning algorithm of your choice.

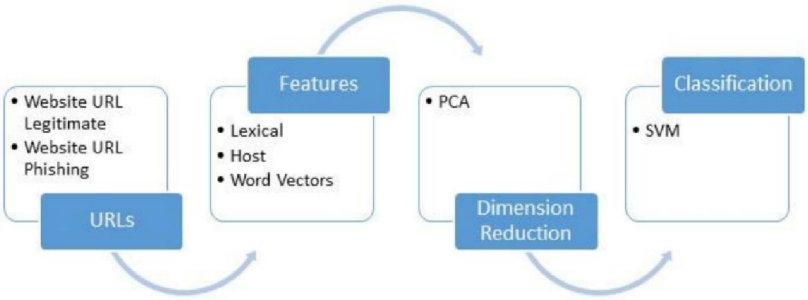

Fig. 2. Proposed technique

C. Step 3

To avoid high dimensions, main components Analysis (PCA) [21] is used for features. This is the purpose of PCA Reduce large variable sets to smaller variable sets. Information is retained. A famous statistical method We try to describe the true covariance using the form A small range of components. These components are linear A mixture of real variables is often allowed Higher interpretation and understanding of various sources Prescription

D. Classification

This step uses a classifier to get the final result. From A classifier is just a trained machine learning algorithm Predicting results and performing classification. Because it doesn't exist A single complete and accurate classification. Classifier It was chosen mainly because it is used for purposes like Google. Issues such as spam detection, phishing and phishing emails Malicious websites and URLs. The system simply tries to use This system is for prediction and final classification activities. Support vector machines are used for classification. Support vector machines (SVM) are often used. SVM phishing attack detection classifier works with example Create a map of training and set changes Change the feature set that creates the feature room and save the instance URLs from two classes with great features The space changes.

III. RESULTS AND DISCUSSION

The proposed method is based on machine learning Compared to conventional technology in our tests, Equivalent classification methods are used to teach split test. Each partition leaves 70% of the statistics used for training For testing purposes, the accuracy is calculated using Equation 1.

Accuracy = (TP+TN)/(TP+FP+FN+TN)

TP stands for positive, TN stands for true Mean negative, false negative FN and false positive FP Ability according to this formula, Accuracy and results of machine learning algorithms from The overall performance of the proposed method is Other previous techniques See Table 1 for complete details. Implementation of the proposed technique and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 4, May 2023

other techniques Technology. Therefore, machine learning is proposed Techniques are very effective in reducing factors. It has been said that It provides the overall performance of the proposed technique Improves the accuracy of classification algorithms. Again, Promising results show that the proposed technique Effective and flexible for different data sets. TP stands for positive, TN stands for true Mean negative, false negative FN and false positive FP Ability according to this formula, Accuracy and results of machine learning algorithms from The overall performance of the proposed method is Other previous techniques See Table 1 for complete details. Implementation of the proposed technique and other techniques Technology. Therefore, machine learning is proposed Techniques are very effective in reducing factors. It has been said that It provides the overall performance of the proposed technique Improves the accuracy of classification algorithms. Again, Promising results show that the proposed technique Effective and flexible for different data sets.

TABLE I. SUMMARY OF RESULTS

Classifier	Feature set	No of Features	Accuracy
FACA [2]	Full	30	90.44
Random Forest [22]	Full	30	94.27
Random Forest [22]	HEFS	5	93.22
SVM	**Proposed**	**5**	**95.66**

IV. CONCLUSION AND FUTURE WORK

This white paper describes machine-based phishing detection. Learning technology is also a machine classifier A learning algorithm identifies legitimate phishing websites. From The proposed method using SVM with an accuracy of 95.66% A very low false positive rate is the proposed technique Identify and mitigate new temporary phishing sites. caused by a phishing attack. Offer performance Machine learning based methods are more effective Old phishing detection technology It will be useful to investigate its impact in the future. Feature selection using different classification algorithms.

REFERENCES

[1] Higashino, M., et al. An Anti-phishing Training System for Security Awareness and Education Considering Prevention of Information Leakage. in 2019 5th International Conference on Information Management (ICIM). 2019.
[2] H. Bleau, Global Fraud and Cybercrime Forecast,. 2017.
[3] Michel Lange, V., et al., Planning and production of grammatical and lexical verbs in multi-word messages. PloS one, 2017. 12(11): p. e0186685-e0186685.
[4] Rahman, S.S.M.M., et al. Performance Assessment of Multiple Machine Learning Classifiers for Detecting the Phishing URLs. 2020. Singapore: Springer Singapore.
[5] He, M., et al., An efficient phishing webpage detector. Expert Systems with Applications, 2011. 38(10): p. 12018-12027.
[6] Mohammad, R.M., F. Thabtah, and L. McCluskey. An assessment of features related to phishing websites using an automated technique. In 2012 International Conference for Internet Technology and Secured Transactions. 2012.
[7] Abdelhamid, N., A. Ayesh, and F. Thabtah, Phishing detection based Associative Classification data mining. Expert Systems with Applications, 2014. 41(13): p. 5948-5959.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 4, May 2023

[8] Toolan, F. and J. Carthy. Feature selection for Spam and Phishing detection. in 2010 eCrime Researchers Summit. 2010.
[9] Jain, A.K. and B.B. Gupta, Towards detection of phishing websites on client-side using machine learning based approach. Telecommunication Systems, 2018. 68(4): p. 687-700.
[10] 1PhishTank, Phishing dataset. 2018, Verified phishing URL.
[11] 1Openfish, Phishing dataset. 2018.
[12] 1Chiew, K.L., et al., Utilisation of website logo for phishing detection. Computers & Security, 2015. 54: p. 16-26.
[13] Benavides, E., et al. Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review. 2020. Singapore: Springer Singapore.
[14] Zhang, W., et al., Two-stage ELM for phishing Web pages detection using hybrid features. World Wide Web, 2017. 20(4): p. 797-813.
[15] El-Alfy, E.-S.M., Detection of phishing websites based on probabilistic neural networks and K-medoids clustering. The Computer Journal, 2017. 60(12): p. 1745-1759.
[16] Montazer, G.A. and S. ArabYarmohammadi, Detection of phishing attacks in Iranian e-banking using a fuzzy–rough hybrid system.Applied Soft Computing, 2015. 35: p. 482-492.
[17] Wang, Y.-G., G. Zhu, and Y.-Q. Shi, Transportation spherical watermarking. IEEE Transactions on Image Processing, 2018. 27(4): p. 2063-2077.
[18] De Maio, C., et al., Time-aware adaptive tweets ranking through deep learning. Future Generation Computer Systems, 2019. 93: p. 924-932.
[19] De Maio, C., et al., Social media marketing through time aware collaborative filtering. Concurrency and Computation: Practice and Experience, 2018. 30(1): p. e4098.
[20] Peng, T., I. Harris, and Y. Sawa. Detecting Phishing Attacks Using Natural Language Processing and Machine Learning. in 2018 IEEE 12th International Conference on Semantic Computing (ICSC). 2018.
[21] Abdi, H. and L.J. Williams, Principal component analysis. WIREs Computational Statistics, 2010. 2(4): p. 433-459.
[22] Sahingoz, O.K., et al., Machine learning based phishing detection from URLs. Expert Systems with Applications, 2019. 117: p. 345-357.