IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 6.252

Volume 2, Issue 1, June 2022

Phishing Attack Detection using Machine Learning

Prof. Ritesh Shrivasta, Gaurav Kumar Rajbhar, Aniket Krishna Suryavanshi, Mohammad Yasir khan, Preshit Ravi Lute, Gokul Dinesh Panse, Mohammad Hasan Abbas

Department of Computer Science
Anjuman College of Engineering and Technology, Nagpur, Maharashtra, India

Abstract: *Phishing is a fraudulent attempt to extract sensitive information from individuals or organisations, such as usernames, passwords, and credit card information, by impersonating a trustworthy organisation in a digital communication. Phishing attacks pose significant risks to users' privacy and security. The goal of this research is to provide an overview of various phishing attacks and techniques for protecting information. It also discusses MachineLearning-based categorization for phishing website data in the Machine Learning Object storage database. As we move closer to a better future to better technological advances each year, the danger of credit card information being compromised grows. Credit card fraud has risen dramatically in recent years. This includes details hacking, phishing, and other totally incorrect and illegal means to steal credit card data. In this construction and operation, we will use Machine Learning to implement the phishers URL phishing detection and prevention technique, which will provide real significance of the checked URL and fetched Email.*

Keywords: Phishing, Personal information, Machine Learning, Malicious links, Phishing Domain Characteristics, Algorithm, Machine Learning, SVM, Security

I. INTRODUCTION

Phishing is a broad term used to describe a group of people who scam people by sharing personal information such as customer name, password, lending card number, and so on, and who manipulate data for dissemination purposes. The first contact is made to a large group of people all at once, so anybody can be a victim. They will touch their victims via URLs, social networks, emails, and phone calls. The only goal of these people's attack is to send a counterfeit correspondence that appears to have emerged from the actual organisation, in the hope that a huge crowd will follow this same links provided by these contacts and reveal their personal details to the phishers. Phishing Automated detection methods are used to defraud outsiders of billions of dollars, and phishing technology exploits human behaviour as well as the rise of the internet to defraud millions of people worldwide[1]. By hiding behind a legitimate entity, social networks are used for deceptive, cultivated, and perceptive data from internet users.

The primary goal of online fraud technology is to fraudulently carry out financial on behest of web users [2]. According the anti-phishing workgroup (APWG), an NGO society (a non-profit group), the global phishers survey 2016 has already shown all phishing scams from 2012 to 2016. (Figure 2) [3]. The anti-phishing group of experts (APWG) also reported 180,768 phishers incidents detected in the first quarter of 2019 (January, February, and March) [4]. Various methodologies are currently being used to detect phishing sites and emails. SajidYousufBhat et al. [5] propose a method for "Spammer classification using ensemble techniques over structural social media network features." [5] determines whether a URL on a social media network with society features is spam or legitimate.Mouad Zouina et al. [6] propose "A Novel lightweightURL phishers detection using SVM as well as similarities index." Six features are used in [6] to detect phishing from URLs. The use of SVM and the similarity index is intended to improve the overall acknowledgement of the phishers detection system.

II. LITERATURE REVIEW

Several techniques for detecting phishing attacks have been published in the literature. This review presents an overview of investigative techniques for phishing attacks. In general, phishing detection methods are divided into two categories: user education and software-based anti-phishing techniques. There are three types of software-based techniques: ranking, heuristic-based, and visual similarity-based. List-based anti-phishing techniques keep a black-list, white-list, or a

DOI: 10.48175/IJARSCT-4578

combination of the two. A black-list of dubious domain names and IP addresses is maintained in the black-list-based anti-phishing approach. Although black-lists are frequently updated, most black-list-based approaches are inadequate to deal with zero-hour malicious emails. After 12 hours, 47 percent to 83 percent of phishing domains are updated in the blacklist. Google Private Browsing mode API, Domain controller black-lists, and predictive dark are some of the approaches that use black-lists. However, maintaining a blacklist necessitates a significant investment in assets for reporting and verifying suspicious links. Because vast numbers of phishing web sites are formed every day, it is difficult to keep the black-list up to date. Some anti-phishing solutions recommended by the literary works to protect users from phishing scams are listed below:

Google offers a safe browsing service that allows applications to verify URLs against a list of suspect domains that is updated regularly by Google. It is an exploratory API that works with Browsers Such as Google Chrome Firefox and is extremely simple to use.

Clients can use the Safe BrowsingLookupAPI to send suspicious URLs to the Safe Browsing service, which will determine whether the URL is valid or malicious. The client API sends URLs via GET or POST requests, which are checked against Google's malware and phishing lists. The following are some of the Private Browsing mode Lookup API's shortcomings: (i) No hashing is conducted before sending a URL, and (ii) the lookup server's response time is not limited.

A target list is a list of URLs that are suspicious or prohibited and must be blocked or rejected access to the network or system. This method is extremely simple to put into action. Its sole purpose is to deny any suspicious URLs network access. However, this method is insufficient to detect the large bulk of phishing incidents because new threats, such as zero-day attacks, emerge on a daily basis. This method is incapable of perceiving or preventing any new type of attack. It necessitates maintaining a detailed list of malicious sites and their reports, which consume a significant amount of system resources. Phishers may create URLs specifically to avoid detection besides tools that employ a blacklist system. Finally, this method fails to detect some kinds of attacks that are directed at a profitable organisation.

III. DRAW BACKS

Aburrous et al. proposed a smart system for detecting phishing websites in banking. They created a method that incorporates fuzzification with such machine learning algorithms to detect and classify phishing websites using 10-fold cross-validation. This model had a grouping accuracy of 86.38 percent. This model, however, has a high proportion of false positives. Basnet et al. proposed a heuristic-based strategy to group phishing URLs using only URL data. To detect phishing URLs, the authors used a binary classification approach that divided URLs into the phishing URLs as well as legitimate URLs. The results of experiments demonstrate that the suggested approach outperforms related work in detecting phishing URLs.

However, this method has only been evaluated on a data set of less than 300 rows. It might not work well on a huge database. Jain and Richariya devised a new technique for identifying spam scams that makes use of link-based features. To detect phishing attacks, a technology demonstrator web browser was used to process evey incoming email. The prototype as well as their algorithm work together to keep the system user set informed of potential attacks and prevent people from clicking on malicious URLs.

IV. PROBLEM STATEMENT

In the SLR, 55 primary study articles were chosen, and 51 types of attacks/threats were recognised in these articles provide a complete explanation of the airstrikes identified by various researchers. The most severe dangers to the online banking system, according to the majority of research studies (16.98 percent), are trojans (all types) and malware (14.55 percent), related to social designing, pharming, phishing, weak passwords, port scanners, computer bugs, message sniffers, denial - of - service, as well as automated reply. Trojans. It has become one of the world's fastest-growing cybercriminal techniques, involving the theft of private details from unsuspecting users.

V. AIM AND OBJECTIVE

Because phishing attacks exploit user weaknesses, they are difficult to prevent, but it is critical to improve phishing detection techniques. Phishing is a scam framework that makes use of a mix of social engineering and advancement to

DOI: 10.48175/IJARSCT-4578

personal and sensitive data, such as passwords and open-end credit unfussy elements by impersonating the features of a trustworthy person or firm in electronic medium.

- To create an effective sensing tool for tracking and detecting malicious web pages.
- To identify phishing websites, a combined approach of building resource description structure models and classifying websites using deep - learning and group learning methods is used.

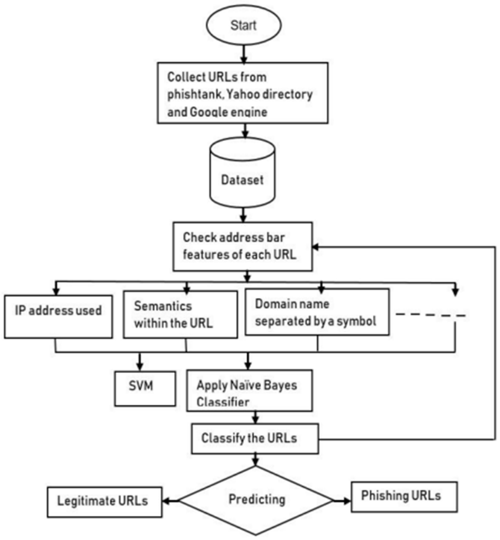
VI. Proposed system

A single layer convolutional neural network is used to implement a dynamic technique for identifying phishing methods. In this paper, the values of six heuristics have been calculated using same algorithm and in the first step of the method. In the execution, a set of data of URLs is used, with a mix of 13phishing and non-phishing URLs. The dataset was obtained from the UCI Repository. Machine Learning models with the lowest RMSE and output layers achieve high accuracy. The learning proportion is often used as a result parameter. When the accuracy of both methods was compared, the greatest accuracy was obtained in ANN PSO.

- TP (True Positive): phishing URLs detected in number.
- FN (False Negative): Incorrect URLs.
- TN (True Negative): correct Legitimate URLs being classified.
- FP (False Positive): Incorrect Phishing URLs which are classified.

Extraction of Features from Data Sets as well as URLs A huge number of data sets (36,874), debated in sub-1, were gathered and analyzed to make them appropriate for the requirement of this study. Many stages were involved in the processing, including web page feature extraction, data standardisation, and attribute weighting. These steps are critical in order for the classifiers to comprehend the data sets as well as appropriately categorise them into there own classes. To learn about new phishing trends, the classifier is given training to new phishing web pages. The results of this phase are fed into the next section of the suitable classifiers. We develop a hybrid machine learning techniques for effectively classifying phishing URLs based on the evidence provided for each URL. Phishing URLs are treated as a binary classification problem, with benign URLs falling into the negative category and phishing URLs falling into the positive category. To create our data sets, we gathered phishing as well as benign URLs from PhishTank, Yahoodirectory, and the Google engine. Following that, we extract many characteristics that have proven effective in predicting phishing Websites by classifying the sets of data into their respective classes using various publicly available resources. We use SVM as well as DecisionTree algorithms to build models from training data that include feature extractions as well as class labels.

VII. FLOW CHART


Start
Collect URLs from phishtank, Yahoo directory and Google engine
Dataset
Check address bar features of each URL
IP address used
Semantics within the URL
Domain name separated by a symbol
SVM
Apply Naïve Bayes Classifier
Classify the URLs
Legitimate URLs
Predicting
Phishing URLs

VIII. EXPERIMENTAL RESULTS

The first section focuses on data collection, data set processing, and URL feature extraction. We look at various heuristic features in URL structure, such as a generic social manipulation feature, a lexical highlight in the URL, numerous alphabets, as well as phishing target brand name. The feature vector is built with 13 major characteristics to prototype our classifiers. The second section evaluates our approach by classifying a data set that used a hybrid of classifiers. We carried out various experiments. The experiment results show that the proposed scheme achieves an average accuracy of 97.8 percent.

IX. OUTPUT

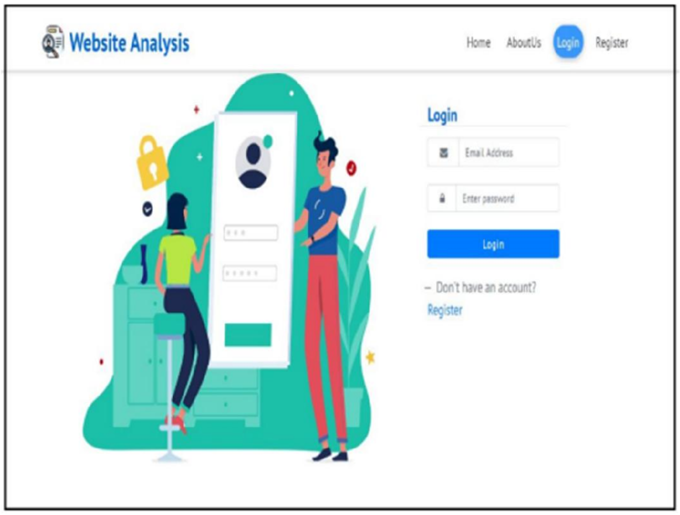
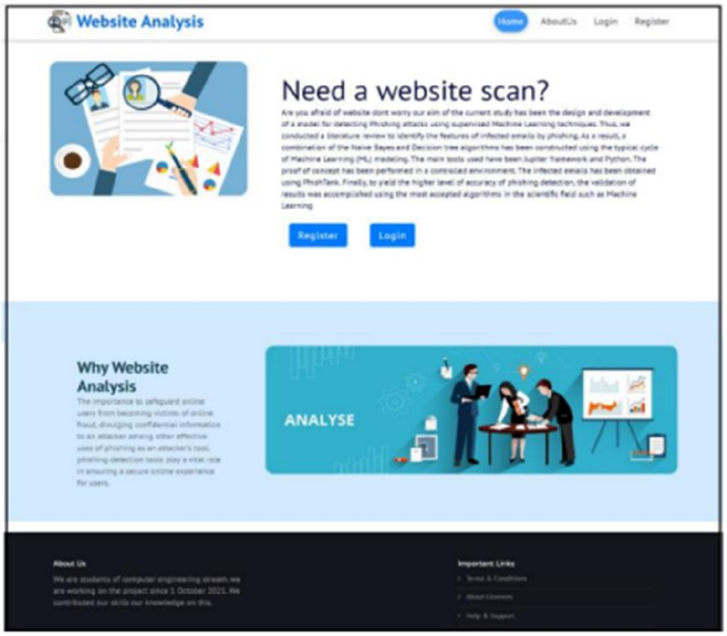


Fig Login Module



Need a website scan?

Are you afraid of website dont worry our aim of the current study has been the design and development of a model for detecting Phishing attacks using supervised Machine Learning techniques. Thus, we conducted a literature review to identify the features of infected emails by phishing. As a result, a combination of the Naive Bayes and Decision tree algorithms has been constructed using the typical cycle of Machine Learning (ML) modeling. The main tools used have been Jupiter framework and Python. The proof of concept has been performed in a controlled environment. The infected emails has been obtained using PhishTank. Finally, to yield the higher level of accuracy of phishing detection, the validation of results was accomplished using the most accepted algorithms in the scientific field such as Machine Learning

Why Website Analysis

The importance to safeguard online users from becoming victims of online fraud, divulging confidential information to an attacker among other effective uses of phishing as an attacker's tool, phishing detection tools play a vital role in ensuring a secure online experience for users.

About Us

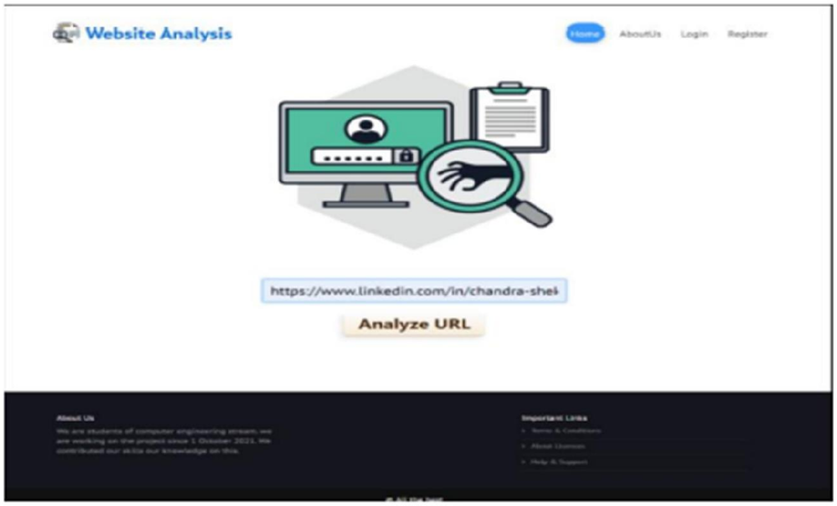
We are students of computer engineering stream we are working on the project since 1 October 2021. We contributed our skills our knowledge on this.

Important Links

- Terms & Conditions
- About Licenses
- Help & Support

fig home page

DOI: 10.48175/IJARSCT-4578



https://www.linkedin.com/in/chandra-shel

About Us

We are students of computer engineering stream, we are working on the project since 1 October 2021. We contributed our skills our knowledge on this.

Important Links

- Terms & Conditions
- About Domain
- Help & Support

© All the best

FIG URL Analyzing Page

X. CONCLUSION

The developed system can raise public awareness and security regarding email phishing attacks. Nowadays, the web is now one of the most popular and widely used platforms for phishing attacks. As a result, the developed system can protect its consumers from such attacks by determining which emails are secure and which are not. As a result, the implemented system serves as just an anti-phishing system. It will use a Deep Learning Algorithm for detecting whether an email has been phished as soon as possible, providing high accuracy while also trying to protect the end user from becoming a target of email phishing. Using different approaches together will improve the system's accuracy, resulting in an effective protection system. The disadvantage of this system is that it detects some minor false negative results. These disadvantages can be overcome by adding much richer features to nourish the machine learning algorithm, resulting in much higher accuracy.

XI. FUTURE SCOPE

The widely publicised Gmail phishing scam that happened earlier this year is each example of the new threat that affected a large number of users. Users have been sent an e - mail that appeared legitimate and guided them to a real Google page in this case. While most phishing scams direct users to malicious domains, this particular move simply tricked them into granting broad authorizations to a malicious application. Hackers could then see the victims' contacts, read one‘s emails, learn about their locations, and view files created in G Suite. The Gmail fake email attack demonstrates how sophisticated these methods have become – it was hard to detect and prevent. A key takeaway is that the invasion was able to overcome the psychological trust barrier. Users were duped into offering permissions to a third-party app even though they trusted it; they thought the app was a Google-approved service. A minor change in how the application web address was disguised successfully persuaded users that application was reliable.

REFERENCES

[1]. R. B. Basnet, A. H. Sung, "Mining web to detect phishing URLs", Proceedings of the InternationalConference on Machine Learning and Applications, vol. 1, pp. 568-573, Dec 2018.

[2]. Abdelhamid N., Thabtah F., Ayesh A. (2019) Phishing detection based associative classificationdatamining.Expert systems with Applications Journal. 41 (2019) 5948-5959.

[3]. Mohammad, R. M., Thabtah, F. & McCluskey, L. (2019) Predicting Phishing Websites usingNeuralNetwork trained with Back Propagation. Las Vigas, World Congress in Computer Science, ComputerEngineering, and Applied Computing, pp. 682-686.

[4]. Aburrous M.., Hossain M., Dahal K.P. and Thabtah F. (2020) Experimental Case Studies for InvestigatingE-

Banking Phishing Techniques and Attack Strategies. Journal of Cognitive Computation, Springer Verlag, 2(3):242-253.

[5]. Mohammad R., Thabtah F., McCluskey L., (2014B) Intelligent Rule based Phishing Websites Classification.Journal of Information Security (2), 1-17. ISSN 17518709. IET.

[6]. Jain, Ankit Kumar, and B. B. Gupta. "Comparative analysis of features based machine learning approaches for phishing detection." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. IEEE, 2016, pp. 2125-2130.

[7]. R.Aravindhan, Dr.R.Shanmugalakshmi, Certain Investigation on Web Application Security: PhishingDetection and Phishing Target Discovery, January 2019.

[8]. L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, "A novel approach for phishingdetectionusing URL-based heuristic," 2014 Int. Conf. Comput. Manag. Telecommun. ComManTel 2014, pp. 298–303,2014.

[9]. Naghmeh Moradpoor, Employing Machine Learning Techniques for Detection and Classification Of PhishingEmails, July 2017.

[10]. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. (2009) The WEKADataMiningSoftware: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[11]. X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, "Boosting the Phishing Detection Performance bySemantic Analysis," 2017.

[12]. K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectorsandtechnical approaches," Expert Systems with Applications, vol. 106, pp. 1–20, 2018.

[13]. M. Priya, L. Sandhya, and C. Thomas, "A static approach to detect driveby-download attacks onwebpages,"Proc. of International Conference on Control Communication and Computing (ICCC'13), pp. 298–303, 2013.

[14]. H. Yuan, X. Chen, Y. Li, Z. Yang, and W. Liu, "Detecting phishing websites and targets basedonurlsandwebpage links," Proc. 24th International Conference on Pattern Recognition, pp. 3669–3674, 2018.

[15]. J. Feng, L. Zou, O. Ye, and J. Han, "Web2vec: Phishing webpage detection methodbasedonmultidimensional features driven by deep learning," IEEE Access, vol. 8, pp. 221214–221224, 2020.

[16]. D. K. McGrath, A. Kalafut, and M. Gupta, "Phishing infrastructure flfluxes all the way," IEEESecurityPrivacy, vol. 7, no. 5, pp. 21–28, 2009

DOI: 10.48175/IJARSCT-4578