ZKG INTERNATIONAL ISSN: 2366-1313

PHISHING WEBSITE DETECTION USING MACHINE LEARNING

[1] Mrs.K.Swapna, [2] Sanuthi Sutharapu, [3] Sai Anuhya Ramadugu, [4] Shrey Tiwari

[1]Associate Professor, Dept. Of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

Kanagiri.Swapna@tkrec.ac.in

[2, 3, 4,] BTech Student, Dept. Of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad

sanuthisutharapu08@gmail.com, ramadugusaianuhya@gmail.com, shreytiwari1111@gmail.com

Abstract: *A specific kind of internet security issue that targets human weaknesses instead than technological flaws is phishing websites. It can be characterized as the practice of luring internet users in order to get private data, including passwords and usernames. We present an intelligent algorithm in this study that may recognize phishing websites. The system operates as an extension to a web browser, providing the user with more functionality by notifying them instantly when it comes across a phishing website. The system is built using supervised learning, a type of machine learning. Because of the Random Forest technique's strong classification performance, we have chosen it. By analysis the features of phishing websites and selecting a most effective combination of them to train the classifier, our goal is to develop a superior performance classifier. Consequently, we end our work with a combination of 26 features and an accuracy of 98.8%.*

I INTRODUCTION

Today's world has made technology a vital part of the twenty-first century. One of these technologies that is increasing quickly each year and has a significant impact on people's lives is the internet. It has developed into a useful and practical instrument for simplifying public transactions, including online banking and e-commerce. Because of this, individuals now believe that providing the Internet accessibility to their private data is handy. Consequently, there is now a serious security issue due to the security thieves who have begun to target this information. One of these issues is considered to be phishing websites. They are trying to deceive the user into providing them with their personal information through a scam known as social engineering.

ISSN: 2366-1313

According to the Anti-Phishing Working Group (APWG) [1] and Kaspersky Lab [2], which recorded phishing attacks, the number of phishing attacks is on increase and presents a security risk to user information. It has increased by 47.48% from all of the phishing attacks that have been detected during 2016.

Many studies have attempted to address the problem of phishing in recent times. Some researchers checked the URL with lists of dangerous websites which were already in existence or that they had developed themselves, while other researchers used the URL in the other way, comparing it with a white list of websites that were lawful. The second approach makes use of heuristics to determine whether a website is phishing by comparing its signature to a database of previous assaults. Researchers additionally utilized Alexa to measure website traffic as an additional tool to recognize fraudulent websites.

In addition, methods based on machine learning are being employed by other researchers. Computer science, a subfield of artificial intelligence (AI), is an examination of machine learning, which is the performance of tasks with the capacity to learn or behave in a smart way. Supervised learning and unsupervised learning are its two distinct forms of learning.

II LITERATURE SURVEY

Title: Content-Based Methodology

Writer: Zhang.

Description: Using the well-known TF-IDF algorithm, the design and assessment of CANTINA, an individual content-based technique for identifying phishing websites, were presented. It evaluates a page's textual content on its own merits. They completed trials using a few straightforward criteria that can be utilized to reduce false positives. Thus, employing heuristics, it can capture around 90% of phishing sites with just 1% false positives, while a pure TF-IDF technique can catch about 97% of phishing sites with only 6% erroneous positives.

Written by Rao and Ali

Developed a desktop program that uses an individual heuristic based on URLs and page content to identify phishing websites. They detected phishing websites using the Phish Shield program utilizing copyright, null footer links, zero links in the body of the HTML, links with maximum frequency domains, and white lists. It had an FP of 0.035% with an accuracy of 96.57%.

ISSN: 2366-1313

Title: Approaching URLs a fresh strategy

Written by Nguyen et al.

A technique of identifying phishing sites has been suggested, that includes collecting multiple components from the URL and calculating a metric for each element. Subsequently, the page rating and the attained metrics will be combined to determine if the websites are phishing websites. The methodology can identify more than 97% of phishing websites, based on the data. A technique to forecast phishing URLs was published by Jeeva and Rajsingh [8] using association rule mining to generate rules. They selected known information from frequently seen item set attributes that were extracted out of the dataset using the a priori approach.

Writers: Rajsingh and Jeeva

Description: Besides to using a priori, which only mark rules via the confidence technique, this algorithm also uses predictive apriority, which works on hidden data to assess the accuracy of association rules. It incorporates both confidence and support techniques that are measured in its accuracy. They thereby displayed key elements of the URL that assist in identifying if it is authentic or phishing.

Title: Artificial Intelligence

In order to increase blocking efficiency, Sanglerdsinlapachai and Rungsawang [9] established new features to the heuristic features of CANTINA. Their features were able to increase detection accuracy by 15% and 20% in terms of f-measure and error rate, respectively. A layered solution called CATINA+ has been proposed by Xiang et al. [10] is advancement over Zhang's work [3]. With the addition of more features and the implementation of machine learning techniques, CATINA produced a 92% true positive rate and a 0.4% false positive rate.

Muhammad et al.

Using neural network architecture, an anti-phishing technological advances was created that can predict phishing attempts ahead of time (NN). The program implements optimal generalized performance utilizing optimal NN structure on a dataset containing 600 authentic URLs and 800 phishing URLs. Testing accuracy utilizing 17 characteristics includes IP address, lengthy URL, URL with '@', abuse of HTTPs, sub domain in URL, & request URL, is 92.48% effective when NN epochs were 500.

Title: Kannan and Pradeepthi

Description: provided an overview of studies done on classification methods for phishing URL identification. Lexical

ISSN: 2366-1313

features, URL-based features, network-based features, and domain-based features are the four categories into they've split the 4500 URLs used as the dataset. NB, multi-layer perception, J48 tree, Logistic Model Tree (LMT), RF, random tree, C4.5, ID 3, C-RT, and K-nearest neighbour (KNN) was among the machine learning approaches they worked with. They came to their conclusion that the best classifiers for the objective of classifying phishing URLs are the ones that are based on trees.

Marshal and associates.

Has launched Phish Storm, an approach that uses lexical analysis of the URL to detect phishing URLs. The method makes use of twelve elements, including the amount of relevant and associated phrases identified in search engine inquiries, Alexa Rank, popularity of the registered domain, and statistics based on these words in URL. Using supervised classification, these attributes were applied to a dataset of 96,018 genuine and phishing URLs.

III SYSTEM ANALYSIS

Existing System

Numerous researches have attempted to address the issue of phishing in recent times. Some researchers checked the URL with lists of dangerous websites that were already in existence or that they had created, while other researchers used the URL in the other way, comparing it with a white list of websites that were lawful. The latter method makes use of heuristics to determine whether a website is phishing by comparing its signature to a database of previous assaults. Researchers have also used Alexa to measure website traffic as an additional tool for identifying fraudulent websites.

3.1.1 Drawbacks

- They are more inclined to give in if they lose private information such as a credit card number, login, or password.
- Is able to propagate malware.
- Personal data loss, including cookies, search history, and other information

3.2 Proposed System

Determining if a given URL is a phishing website or not is the goal of this research project. The results of the experiment indicate that random forest-based classifiers are the most effective, with a remarkable 97.47% classification accuracy for the phishing site dataset. We may use this model in the future to additional Phishing datasets that are greater in size than the ones we currently have, and we will then evaluate how well those

ISSN: 2366-1313

classification algorithms perform in terms of classification accuracy.

3.2.1 Advantages

- Utilize Reinforcement Learning (RL) to model the identification of phishing websites. An agent uses the provided input URL to learn the value function in order to complete the classification job.
- Use a deep neural network to execute Reinforcement Learning to map the step-by-step decision-making process for classification.
- Evaluate and contrast the deep reinforcement learning-based phishing URL classifier's performance with that of the various ones already in use.

IV SYSTEM DESIGN

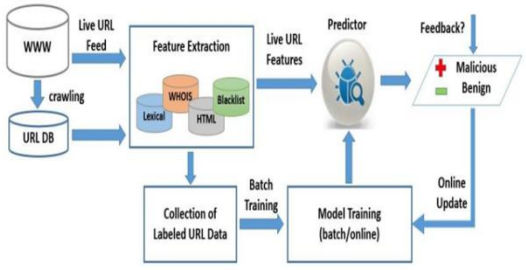


Fig-3 System Architecture

4.2 System Study

4.2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The

ISSN: 2366-1313

developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4.3 UML DIAGRAMS

UML stands for Unified Modelling Language. An industry-standard general-purpose modelling language used in object-oriented software engineering is called UML. The Object Management Group developed and oversees the standard. The intention is for UML to spread as a standard language for modelling object-oriented software. The two main parts of UML as it exists now are a notation and a meta-model.

V IMPLEMENTATION

MODULES:

1. URL-Based Features
2. Domain-Based Features
3. Page-Based Features
4. Content-Based Features
5. Protocol Based Features
6. publisher
7. Admin
8. Users

Publishers:

Publishers can provide their servicers by registering their domain URL's with complete description about their service and they can edit or delete their services.

Admin:

In this module admin can find malicious service URL's and he can delete their services and admin is having capable of deleting the users who are registering malicious services

User:

In this module user can search for the any kind of popular services and user find the service URL's which are useful to him.

Domain-Based Features:

ISSN: 2366-1313

Host-based features explain "where" phishing sites are hosted, "who" they are managed by, and "how" they are administered. We use these features because phishing Web sites may be hosted in less reputable hosting centres, on machines that are not usual Web hosts, or through not so reputable registrars.

URL-Based Features

Lexical features are the textual properties of the URL itself, not the content of the page it points to. URLs are human-readable text strings that are parsed in a standard way by client programs. Through a multistep resolution process, browsers translate each URL into instructions that locate the server hosting the site and specify where the site or resource is placed on that host. To facilitate this machine translation process, URLs have the following standard syntax.

Page Based Features:

The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain. In our datasets and due to its short life span, we find that 98% of phishing dataset items have no links pointing to them. On the other hand, legitimate websites have at least 2 external links pointing to them Page Rank is a value ranging from "0" to "1". Page Rank aims to measure how important in our datasets, we find that about 95% of phishing web pages have no Page Rank. Moreover, we find that the remaining 5% of phishing web pages may reach a Page Rank value up to "0.2".

Content Based Features:

Mainly there is use of Natural Language Processing (NLP) and other machine learning techniques. Moreover much use more technical features and process them using machine learning algorithms has been imposed.

VI OUTPUT SCREENS

SCREEN-1

ISSN: 2366-1313

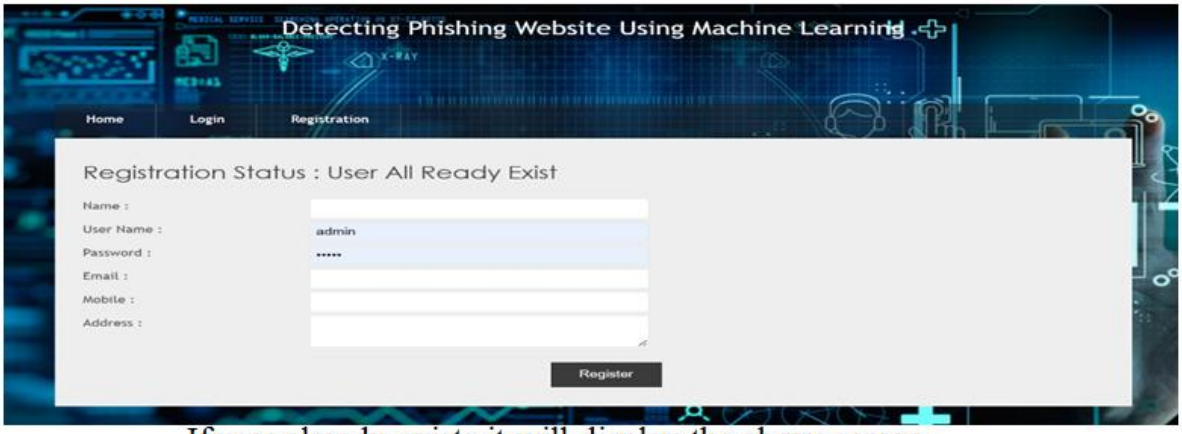
If user wants to search for any specific topic they can search here.

SCREEN-2

New user registration process.

SCREEN-3

ISSN: 2366-1313



If user already exists it will display the above screen

SCREN-4

If publisher posts any suspicious url the url will not be posted.

SCREEN-5

ISSN: 2366-1313

It displays that it is not safe to use.

VII CONCLUSION

We have study all 36 features in order to reduce time computation and providing high performance with the least combination of the powerful features. However, because of time shortage and hardware limitation, we chose random features to process its combination. We concluded after some observation that the combination of features computed takes the shape of normal distribution curve, it starts with least combination of features with low probability of combination and time consuming, then picks up accordingly, then goes down as it reach final number of 36 features.

VIII FUTURE ENHANCEMENT

In future ways can test this phishing websites through many ways according to our technology development. The future works can be to fix the antivirus also into the tool in which the user will be comfort access all pages and be secure.

REFERENCES

[1] AO Kaspersky lab. (2017). The Dangers of Phishing: Help employees avoid the lure of cybercrime. [Online] Available: https://go.kaspersky.com/Dangers-Phishing-Landing-Page- Soc.html [Oct 30, 2017].

[2]"Financial threats in 2016: Every Second Phishing Attack Aims to Steal Your Money" Internet: https://www.kaspersky.com/about/pressrel

eases/2017 financial-threats-in-2016. Feb 22, 2017 [Oct 30, 2017].

[3] Y. Zhang, J. I. Hong, and L. F. Cranor," Cantina: A Content-based Approach to Detecting Phishing Web Sites," New York, NY, USA, 2007, pp. 639-648.

[4] M. Blasi," Techniques for detecting zero day phishing websites." M.A. thesis, Iowa State University, USA, 2009.

[5] R. S. Rao and S. T. Ali,"Phish Shield: A Desktop Application to Detect Phishing Web pages through Heuristic Approach," Procedia Computer Science, vol. 54, no. Supplement C, pp. 147-156, and 2015.

[6] E. Jacobson, and E. Myers, Phishing and Counter-Measures: Understanding the Increasing Problem of Electronic Identity Theft. Wiley, 2006, pp.2–3.

[7] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen,"Detecting phishing web sites: A heuristic URL-based approach," in 2013 International Conference on Advanced Technologies for Communications (ATC 2013), 2013, pp. 597-602.

[8] Z. Zhang, Q. He, and B. Wang,"A Novel Multi-Layer Heuristic Model for Anti-Phishing," New York, NY, USA, 2017, p. 21:1-21:6.

[9] N. Sanglerdsinlapachai and A. Rungsawang,"Web Phishing Detection Using Classifier Ensemble," New York, NY, USA, 2010, pp. 210-215.

[10] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "CANTINA+: A Feature Rich Machine Learning Framework for Detecting Phishing Web Sites," ACM Trans. Inf. Syst. Secure., vol. 14, no. 2, pp. 21:1-21:28, Sep. 2011.

[11] Prasadu Peddi (2015) "A review of the academic achievement of students utilisinglarge-scale data analysis", ISSN: 2057-5688, Vol 7, Issue 1, pp: 28-35.

[12] Prasadu Peddi (2017) "Design of Simulators for Job Group Resource Allocation Scheduling In Grid and Cloud Computing Environments", ISSN: 2319-8753 volume 6 issue 8 pp: 17805-17811.