WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Phishing Website Detection Using Machine Learning

Sagar Khomane[a],Sandesh Bhosale[b],Sanket Karpe[c],Dnyneshwar Tanpure[d]

[abcd]4th year, BE Computer Engineering, Student SVPMs college of engineering Malegaon, India

A B S T R A C T

Phishing is a common way to trick unsuspecting people into revealing certain information. personal information such as your username, password and online financial transaction data is collected through phishing sites. Fishermen use their own websites with the same design and language quality on the official website. As technology advances to prevent phishing attacks, we must apply ant phishing measures to detect phishing attacks. This article is about identification machine learning detection methods and techniques. Often used by attackers phishing because it is easier to trick victims into clicking on malicious links that appear legitimate trying to avoid information security

Keywords:Phishing, Personal Information, Malicious Spam Links, Phishing Domain Characteristics, Cyber Security.

1. INTRODUCTION

Phishing is turning becoming a significant issue for security researchers due to how simple it is to create a phoney website that looks remarkably similar to an actual one. Even if professionals can recognise phoney websites, not all users can, which is why some individuals fall for them. phishing fraud. Getting access to bank account login information is the attacker's primary objective. Because people are oblivious to phishing assaults, they are becoming more and more successful. Because phishing assaults take use of user weaknesses, they are exceedingly challenging to stop, but developing phishing detection techniques is essential. Phishing is a popular kind of extortion when a malicious website poses as a reputable company in order to get personal information like passwords, login credentials, or MasterCard numbers. Despite the fact that there are several ways and tools that may spot potential phishing attempts in messages and traditional phishing content on websites, phishers still utilise inventive and hybrid tactics to bypass open frameworks and programming. Phishing is a type of fraud that uses social engineering techniques to disseminate private and sensitive information, such as passwords and public credit information, while pretending to be someone else. To trick people into visiting fake websites through links on phishing websites, fake communications that appeargenuine and claim to be from legitimate sources, such as financial institutions, online businesses, etc., are created.

2. STATE OF THE ART (LITERATURE SURVEY)

H. Huang et al., (2009) proposed frameworks that distinguish the phishing utilizing page section similitude that breaks down universal resource locator tokens to create forecast preciseness phishing pages usually keep their CSS vogue like their objective pages.

S.Marchal et al., (2017) proposed this technique to differentiate Phishing websites depending on examining authentic site server log knowledge. An application Off-the Hook application or identification of phishing websites. Free, displays a couple of outstanding properties together with high preciseness, whole autonomy, excellent language freedom, speed of selection, flexibility to dynamic phishing, and flexibility to advancement in phishing ways.

Mustafa Aydin et al. proposed a classification algorithm for phishing website detection by extracting websites' URL features and analyzing subset-based feature selection methods. It implements feature extraction and selection methods for the detection of phishing websites. The extracted features about the URL of the pages and composed feature matrix are categorized into five different analyses Alphanumeric Character Analysis, Keyword Analysis, Security Analysis, Domain Identity Analysis, and Rank Based Analysis. Most of these features are the textual properties of the URL itself and others are based on third parties services. Samuel Marshal et al. present Phish Storm, an automated phishing detection system that can analyze realtime any URL in order to identify potential phishing sites. Phish story is proposed as an automated real-time URL phishing ness rating system to protect users against phishing content. Phish Storm provides a phishing ness score for URLs and can act as a Website reputation rating system

.Fadi Thabtah et al. experimentally compared a large number of ML techniques on real phishing datasets and with respect to different metrics. The purpose of the comparison is to reveal the advantages and disadvantages of ML predictive models and to show their actual performance when it comes to phishing attacks. The experimental results show that Covering approach models are more appropriate as ant phishing solutions.
Muhemmet Baykara et al. proposed an application which is known as Anti Phishing Simulator, it gives information about the detection problem of phishing and how to detect phishing emails. Spam emails are added to the database by the Bayesian algorithm. Phishing attackers use JavaScript toplace a legitimate URL of URL onto the browser's address bar. The recommended approach in the study is to use the text of the e-mail as a keyword only to perform complex word processing.

3. PROJECT DESCRIPTION

To create our design, we incorporated a website that offers a platform to all drug users. This adaptable and interactive website will be used to evaluate the reliability of websites. This website was created using, among other web design languages, HTML, CSS, JavaScript, and Django. HTML was used to create the website's first framework. CSS is used to add the website's products, which ups its charm and stoner appeal. Because it is intended for all types of drug users, the website must be simple to use and barrier-free for stoners. Even the most ignorant individual must be able to access this website so they may use it and benefit from it. The website contains details about our services. Information on the unethical behaviour in the This adaptable and interactive website will be used to evaluate the reliability of websites. This website was created using, among other web design languages, HTML, CSS, JavaScript, and Django. HTML was used to create the website's first framework. CSS is used to add the website's products, which ups its charm and stoner appeal. Because it is intended for all types of drug users, the website must be simple to use and barrier-free for stoners. Even the most ignorant individual must be able to access this website so they may use it and benefit from it. The website contains details about our services. There is also information on unethical behaviour in the modern technology world. The website's creators anticipated that users would the ability to identify between reliable and false websites, as well as a greater understanding of the societal inequities that prevail in today's society. They have the right to protect themselves from anybody attempting to profit from their personal information, such delivery addresses, words, credit card data, CVV, bank account information, and so on. The dataset comprises a number of characteristics that need to be taken into consideration while figuring out if a website address is legitimate or malicious.

4. ALGORITHMS USED

Support Vector Machine:
Classification and regression problems are resolved using Support Vector Machine, or SVM, one of the most used supervised learning techniques. It is mostly used, nevertheless, in Machine Learning Classification problems. In order to swiftly categorise new data points in the future, the SVM algorithm aims to define the best line or decision boundary that can split n-dimensional space into classes. The name of this best choice boundary is a hyperplane.
The extreme vectors and points that help create the hyperplane are chosen via SVM. The SVM approach is based on support vectors, which are utilised to represent these extreme situations. Look at the image below to see how two separate groups are shown by

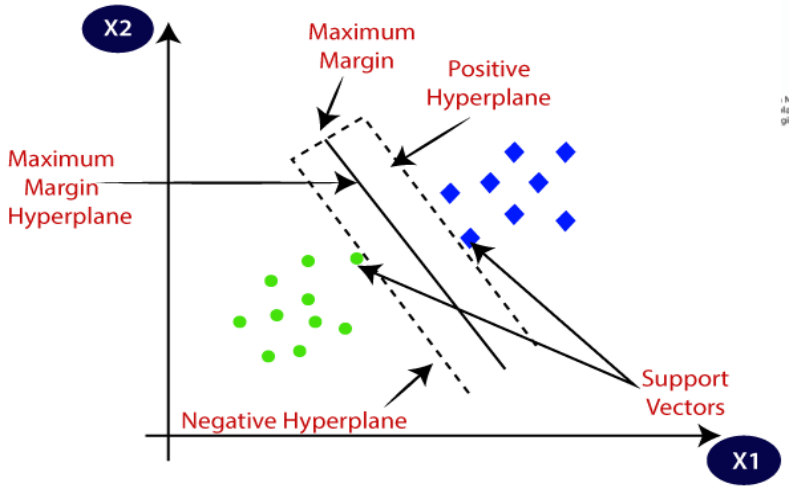

Fig1 SVM Working

5.PROJECT REQUIREMENT

1.Hardware Requirements:-
- 2GB RAM (minimum)
- 100GB HDD (minimum)
- Intel 1.66 GHz Processor Pentium 4 (minimum)
- Internet Connectivity

2. Software Requirements:
- WINDOWS 7 or higher
- Python 3.6.0 or higher
- Visual Studio Code
- Tkinter
- Dataset of Phishing Website

6. WORKING

- Taking an unstructured dataset from Kaggle, GitHub, etc. is our initial step. •
- Because we are using unstructured data, data pretreatment is a crucial stage in the modeling process. Hence, we take certain information from the URL, such as the domain name, URL length, and age of the URL.
- As machine learning only comprehends numerical data, we assign some value to the characteristics after extracting them from the URL, such as 0 and 1.
- In the following phase, we train the model using the SVM methods, and we compare how well they performed.
- We can observe that random forests provide improved accuracy after utilizing two methods.

7.RESULTS

after using SVM algorithms it gives 96% accuracy and it predicts whether the URL is phishing or not.

Fig 2. Result of url.

here we pasted the www.google.com URL and we can see our model predicted the Normal website. so www.google.com is not a phishing website.

8.CONCLUSION

Even while it has been demonstrated that relying just on URL lexical cues provides a high degree of accuracy (97%), phishers have realised that by carefully modifying URLs to evade detection, they may increase the accuracy of their URL targets. I discovered a way to make predictions more difficult. These traits were consequently mixed with others. A. Tips for Successful Hosting. I want to develop a scalable web application that combines phishing detection with online learning. This allows us to quickly learn new phishing attack patterns and extract additional attributes to improve model accuracy.

FUTURE SCOPE

Although it has been demonstrated that using only URL lexical cues yields good accuracy (97%), phishers have figured out how to make it challenging to anticipate a URL destination by meticulously altering the URL to avoid detection. Therefore, the best strategy is to combine these traits with others, including host. In order to learn new phishing assault patterns quickly and increase the accuracy of our models with greater feature extraction, we want to construct thephishing detection system as a scalable web service that incorporates online learning.

REFERENCES

1. HTTPS://WWW.IJERT.ORG/RESEARCH/DETECTION-OFPHISHING-WEBSITES-USING-MACHINELEARNINGIJERTV10IS050235.PDF

2. Y. SÖNMEZ, T. TUNCER, H. GÖKAL, AND E. AVCI, "PHISHING WEB SITES FEATURES CLASSIFICATION BASED ON EXTREME LEARNING MACHINE," 6TH INT. SYMP. DIGIT. FORENSIC SECUR. ISDFS 2018 - PROCEEDING, VOL. 2018– JANUA, PP. 1–5, 2018.

3. T. PENG, I. HARRIS, AND Y. SAWA, "DETECTING PHISHING ATTACKS USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING," PROC. - 12TH IEEE INT. CONF. SEMANT. COMPUT. ICSC 2018, VOL. 2018–JANUA, PP. 300– 301, 2018.

4. M. KARABATAK AND T. MUSTAFA, "PERFORMANCE COMPARISON OF CLASSIFIERS ON REDUCED PHISHING WEBSITE DATASET," 6TH INT. SYMP. DIGIT. FORENSIC SECUR. ISDFS 2018 - PROCEEDING, VOL. 2018–JANUA, PP. 1–5, 2018.

5. S. PAREKH, D. PARIKH, S. KOTAK, AND P. S. SANKHE, "A NEW METHOD FOR DETECTION OF PHISHING WEBSITES: URL DETECTION," IN 2018 SECOND INTERNATIONAL CONFERENCE ON INVENTIVE COMMUNICATION AND COMPUTATIONAL TECHNOLOGIES (ICICCT), 2018, VOL. 0, NO. ICICCT, PP. 949–952.

6. K. SHIMA ET AL., "CLASSIFICATION OF URL BITSTREAMS USING BAG OF BYTES," IN 2018 21ST CONFERENCE ON INNOVATION IN CLOUDS, INTERNET AND NETWORKS AND WORKSHOPS (ICIN), 2018, VOL. 91, PP. 1–5.

7. W. FADHEEL, M. ABUSHARKH, AND I. ABDELQADER, "ON FEATURE SELECTION FOR THE PREDICTION OF PHISHING WEBSITES," 2017 IEEE 15TH INTL CONF DEPENDABLE, AUTON. SECUR. COMPUT. 15TH INTL CONF PERVASIVE INTELL. COMPUT. 3RD INTL CONF BIG DATA INTELL. COMPUT. CYBER SCI. TECHNOL. CONGR., PP. 871– 876, 2017.

8. X. ZHANG, Y. ZENG, X. JIN, Z. YAN, AND G. GENG, "BOOSTING THE PHISHING DETECTION PERFORMANCE BY SEMANTIC ANALYSIS," 2017.

9. L. MACHADO AND J. GADGE, "PHISHING SITES DETECTION BASED ON C4.5 DECISION TREE ALGORITHM," IN 2017 INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION, CONTROL AND AUTOMATION, ICCUBEA 2017, 2018, PP. 1–5.

10. A. DESAI, J. JATAKIA, R. NAIK, AND N. RAUL, "MALICIOUS WEB CONTENT DETECTION USING MACHINE LEANING," RTEICT 2017 - 2ND IEEE INT.