
ISSN: 2321-2152
IJMECE
International Journal of modern
electronics and communication engineering
E-Mail
editor.ijmece@gmail.com
editor@ijmece.com
www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

PHISHING WEBSITE DETECTION USING LIGHT GBM AND SVM

[1] M. KAVITHA, [2] G. SUCHARITHA, [3] S. SHIVA KUMAR, [4] K. MANOJKUMAR, [5] CH. MEGHANA
[2,3,4,5] U.G. Scholor, Department of DS, Sri Indu College Of Engineering & Technology, Ibrahimpatnam, Hyderabad.
[1]Assistant Professor, Department of DS, Sri Indu College Of Engineering & Technology, Ibrahimpatnam, Hyderabad.

ABSTRACT

Malicious URL (or) malicious website is a common and serious threat to cyber security. Naturally, search engine becomes the backbone of information management. Nevertheless, the flooding of large number of malicious websites on search engine has posed tremendous threat to our users. Most of exiting systems to detect malicious websites focus on specific attack. At the same time, available browser extensions based on blacklist are powerless to countless websites.Therefore, it is essential that any data leaving theclient side should be effectively masked such that the servercannot interpret any valuable information from the maskeddata. Here propose the first PPSB service. It provides strongsecurity guarantees that are missing in existing SB services.In particular, it inherits the capability of detecting unsafeURLs, while at the same time protects both the user's privacy(browsing history) and blacklist provider's proprietaryassets (the list of unsafe URLs).In this work, proposed a model which encrypts the users' sensitive data to prevent privacy from both outside analysts and service provider. Also, completely supports selective aggregate functions for online user behavior analysis and guaranteeing differential privacy. Homomorphic RSA algorithm is used for encrypting users' online behavior data.Implementation is done and its performances are evaluated based on a real time behavior set.

Keywords: Malicious URL Detection, Blacklist Creation, History encryption using AES, URL Recommendation, Key verification, History Access.

1. INTRODUCTION

The Internet is a risky location, with extremely good regularity, customers listen about websites becoming unavailable due to denial of provider attacks, or showing modified (and regularly unfavorable) information on their homepages. In different excessive-profile cases, hundreds of thousands of passwords, electronic mail addresses, and credit score card details were leaked into the public area, exposing website customers to both private embarrassment and financial danger.

The purpose of internet site safety is to prevent those (or any) kinds of attacks. The more formal definition of website safety is the act/practice of protective websites from unauthorized access, use, modification, destruction, or disruption.Effective internet site security requires layout attempt throughout the entire of the website: to your net

utility, the configuration of the web server, your rules for growing and renewing passwords, and the patron-facet code. While all that sounds very ominous, the coolest information is that in case you're the usage of a server-aspect net framework, it will almost actually enable "with the aid of default" robust and properly-thought-out protection mechanisms in opposition to some of the extra not unusual assaults. Other attacks can be mitigated thru your net server configuration, for instance with the aid of permitting HTTPS. Finally, there is publically available vulnerability scanners gear that let you find out in case you've made any obvious mistakes.

1.1 SAFE BROWSING

Malicious SB service issuer wants to recognize whether a person is journeying a specific web web page, e.g., some political information. One way to gain that is that the web browser sends all the visited URLs to a far off server, either inside the plaintext, hash cost or encrypted layout. However, this behavior may be detected by tracking and analyzing the browser, e.g., the usage of the taint evaluation technique. Specifically, as a way to track a particular URL the SB carrier issuer can insert the 32-bit hash prefixes of all its decompositions, e.g., c01e362f, after which push this newly up to date prefix filter out to the customers. Later, once a user visits the internet page (or comparable URLs that percentage a few decompositions), the matched hash prefixes might be sent to the far flung SB server. Based on the prior information of the prefix filter (i.e., the mappings among the hash prefixes and their corresponding URLs), the server can infer the URL (or area) navigated by means of the user. It gives strong protection ensures that are lacking in present SB services. In precise, it inherits the capability of detecting dangerous URLs, at the same time as on the identical time protects both the person's privacy (surfing records) and blacklist provider's proprietary belongings (the list of risky URLs). This approach has a few disadvantages along with; developing metadata of URLs fails while the server gets multiple prefixes for a URL and there may be a threat that other URLs may additionally have the equal hash prefixes this makes collision among URLs.

A malicious user would possibly leverage PPSB to degrade the consumer-facet consumer experience, like putting a number of faux or secure URLs or increasing the server-aspect delay. To cope with this capability difficulty, PPSB presents a flexible mechanism for customers to add or eliminate blacklist providers. Admin ought to add the fake URL and keyword to this blacklist storage. User can also allowed suggesting the malicious internet site info concerning black list. In this gadget malware detection machine makes use of a supervised machine gaining knowledge of technique for discovering malwares. The SVM classification with malware detection system extends the idea of signature primarily based detection system with a aggregate of conduct tracking approach. It utilizes static and dynamic evaluation of malwares with the aid of taking the run time traces of the executable. Image based malicious detection also provide to compare the

image functions based totally on original internet site and malicious website. This version also affords seek records security which encrypts the users' sensitive statistics to save you privateness from both outside analysts and the aggregation provider. Also, completely helps selective combination functions for on-line consumer conduct analysis and ensuring differential privateness.

2. LITERATURE SURVEY

2.1 HYBRID RULE-BASED SOLUTION FOR PHISHING URL DETECTION USING CONVOLUTIONAL NEURAL NETWORK

Mourtaji, et.al,Implement a phishing detection mode that incorporates 37 features extracted from six different methods including the black listed method, lexical and host method, content method, identity method, identity similarity method, visual similarity method, and behavioral method. Furthermore, comparative analysis was undertaken between different machine learning and deep learning models which includes CART (decision trees), SVM (support vector machines), or KNN (K-nearest neighbors) and deep learning models such as MLP (multilayer perceptron) and CNN (convolutional neural networks).The novelty in our solution is that it is based on reasonable rules to improve the logic of treatments as in certain cases it is probable to detect a URL without executing the whole process. Deep learning networks have shown an enormous capability to resolve the problem of training from data especially with huge data. It is usually used for computer vision as image detection and classification problems. However, with regards to the machine learning field, researches are based on good concept algorithms. Still, trainings associated to deep learning networks need more resources due to the huge mathematical calculations that can be made, and in the case of binary classification for phishing URL detection, deep learning models especially CNN have not only demonstrated a very good performance and accuracy but also helped in the reduction of error rate.

2.2 PHISHING URL DETECTION USING MACHINE LEARNING METHODS

Ahammad, et.alPresent a solution for detecting such websites with the help of machine learning algorithms focused on the behaviors and qualities of the suggested URL. The web security community has created blacklisting services to identify malicious websites. Due to their recentness, lack of evaluation, or incorrect evaluation, many malicious websites inadvertently escape blacklisting. To create a machine learning model for detecting whether a URL is malicious or not, algorithms such as Random Forests, Decision Trees, Light GBM, Logistic Regression, and Support Vector Machine (SVM) are used. A cybercriminal will create a site that looks like the real one, and all of its information will be identical to that of the absolute URL. The URL will appear as an advertisement on other websites, and the fraud will happen when the user enters their credentials. And another way is by sending the malicious URL to the user through email, and when the user tries to open the URL some nasty

virus will be downloaded, this allows the cybercriminals to access the information to commit their crimes. Malicious and benign URLs look similar, so to distinguish them we need to extract some features from them. Detecting malicious URLs requires extracting some of their features from them, then comparing these features to determine whether the URL is malicious or benign.

2.3 MALICIOUS UURL DETECTION BASED ON A PARALLEL NEURAL JOINT MODEL.

Yuan, et.al Propose a parallel neural joint model algorithm for analysis and detection of malicious Uniform Resource Locator (URL). The URL-based method is significantly faster than other methods because it does not require page parsing. Most companies use Blacklist to identify malicious URLs. By detecting and analyzing malicious URL's characteristics, the semantic and visual information will be extracted. First, a visualization algorithm is used to realize the visualization of the URL mapping to a gray image with texture characteristics. Second, the lexical feature and character feature of URL are extracted and further processed through word vector technology. These extracted features are transformed into lexical embedding vectors and character embedding vectors.The order of characters in the URL is informative. Normally, numerous long URLs of the same class have the same or similar character sequence. A new technique of malware visualization based on image processing technology was proposed indicating that the image texture features are available.Through observation, it can be found that the URLs used by the same organization or generated by the same phishing attack tool have a similar structure. Thus, a parallel joint neural network model is proposed able to capture URLs' visual and semantic information. To encode image texture features and indRNN for encoding the URL text features. Then, we merge the two extracted features and use the attention mechanism to further filter them while focusing on the effective features enhancing the classification accuracy.

2.4 A TRANSFORMER-BASED MODEL TO DETECT PHISHING URLS.

Xu, Pingfan, et.alIntroduce a transformer-based malicious URL detection model, which has significant accuracy and outperforms current detection methods. Proposed approach to the transformer model design is not identical to the standard structure. The model design of Rudd and Abdallah from FireEye Inc. inspires the transformer model design of proposed approach. In proposed solution, the transformer model is very similar to the design of OpenAI's GPT model, one of the famous variants of the Transformer model. Proposed Transformer model applied left to right (L-R) modeling and only contained the encoder part from the standard transformer model.Specifically for the input text pre-processing part, we tokenize the input URL first by splitting it into single-character tokens. The most frequently appeared single-character tokens in the URLs of the training dataset are used to form the token repository, which is also known as the vocabulary of the tokenizer.This token and position embedding layer consists of two sub-

layers which are both embedding layers. One of these two separate embedding layers is for tokens, while the other is for token index (positions). The layer next to the token and position embedding layer is a transformer block. The transformer block is the same as the encoder layer of the standard transformer model, which consists of sublayers: multi-head attention and point-wise feed-forward networks. The transformer block has the same embedding dimension as the token and position embedding layer.

2.5 TOWARDS LIGHTWEIGHT URL-BASED PHISHING DETECTION

Butnaru, et.alProposed work uses supervised machine learning to block phishing attacks, based on a novel combination of features that are extracted solely from the URL. We evaluate our performance over time with a dataset which consists of active phishing attacks and compare it with Google Safe Browsing (GSB), i.e., the default security control in most popular web browsers.Here propose and evaluate a phishing detection engine, which uses supervised machine learning in order to detect phishing attacks based on a novel combination features that are extracted from the URL. This allows us to avoid any delays which stem from the computation of features that need access to third-party resources, such as access to WHOIS records. In summary, our work makes the following contributions: Train, optimise and evaluate a phishing detection engine which relies on supervised machine learning, based on features that stem from the URL. Our feature selection process includes features that have been proven suitable by the literature, coupled with new ones that we propose and evaluate. To the best of our knowledge, we are the first to use the Levenshtein distance as a similarity index feature for training a range of machine learning algorithms in this domain. And also, revisit the use of suggestive vocabulary, as one of our features. Evaluate the performance of our phishing detection engine over time by classifying active phishing attacks that were reported on PhishTank, without model retraining. Find that the performance of the classification is not affected by time, as well as it significantly outperforms the protection that is offered by GSB.

3. EXISTING SYSTEM

Phishing is the fraudulent activity to get sensitive records inclusive of usernames, passwords and credit score card info, frequently for malicious motives, with the aid of disguising as a sincere entity in an digital conversation. Phishing attack can be carried out in various form like Email phishing, Website phishing, spear phishing, Whaling, Tab napping, Evil dual phishing and many others. To avoid this phishing attack various anti-phishing applications have to be use. There are diverse anti phishing solutions inclusive of Blacklist, heuristic, visible similarity, machine learning techniques and many others.

Malicious SB service provider wants to know whether a user is visiting a particular web page, e.g., some political news. One way to achieve this is that the web browser sends all the visited URLs to a remote server, either in the

plaintext, hash value or encrypted format. However, this behavior can be detected by monitoring and analyzing the browser, e.g., using the taint analysis technique. Specifically, in order to track a particular URL the SB service provider can insert the 32-bit hash prefixes of all its decompositions, e.g., c01e362f, and then push this newly updated prefix filter to the clients. Later, once a user visits the web page (or similar URLs that share some decompositions), the matched hash prefixes would be sent to the remote SB server. Based on the prior knowledge of the prefix filter (i.e., the mappings between the hash prefixes and their corresponding URLs), the server can infer the URL (or domain) navigated by the user. It provides strong security guarantees that are missing in existing SB services. In particular, it inherits the capability of detecting unsafe URLs, while at the same time protects both the user's privacy (browsing history) and blacklist provider's proprietary assets (the list of unsafe URLs).

4. PROPOSED SYSTEM

A malicious party might leverage PPSB to degradethe client-side user experience, like inserting a numberof fake or safe URLs or increasing the server-side delay.To address this potential issue, PPSB provides a flexiblemechanism for users to add or remove blacklist providers. Admin could add the fake URL and keyword to this blacklist storage. User can also allowed suggesting the malicious website details regarding black list. In this system malware detection system uses a supervised machine learning approach for discovering malwares. The SVM based malware detection system extends the idea of signature based detection system with a combination of behavior monitoring approach. It utilizes static and dynamic analysis of malwares by taking the run time traces of the executable.This model also provides search data security which encrypts the users' sensitive data to prevent privacy from both outside analysts and the aggregation service provider. Also, completely supports selective aggregate functions for online user behavior analysis and guaranteeing differential privacy. Proposed system also provides the privacy of user's search history information. To provide security with the help of RSA encryption approach.

The blacklist providershave the incentive to collect and publish unsafe URLs and keywords forhelping users to avoid websites that contain malware orphishing and deceptive content, e.g., for the better marketing purpose. Assume that these blacklist providers and thecorresponding PPSB servers are semi-trusted. They faithfullyperform the designed procedures, i.e., the database preparation/update. Butthey should not be aware of the queried URLs from users.In proposed work a service providerthat owns a high-quality blacklist, which may be morefrequently updated or simply contains more items. User also allowed to directly sharingblacklists with servers in an uncontrollable way couldmake these dataset be obtained by every user, including thecompetitors. The client needs to search into the list ofunsafe URLs and keywords. The searched

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

URL could be matched with blacklist providers. Once match could be found, that will show the malicious alert to the user. Otherwise the page will be loaded to show the details to searched user. In further the searched URLs or Keywords are stored in encrypted format, which will not reveal the users sensitive information to the server or unknown person.

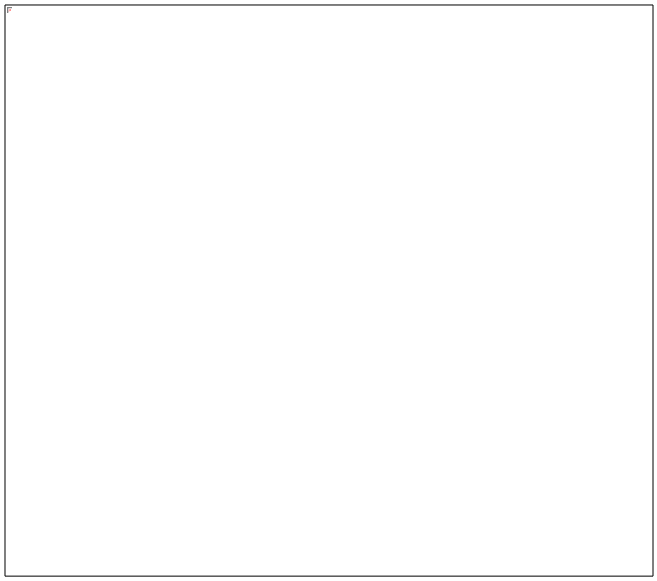

Fig 4.1: Architecture for Proposed Work

5. CONCLUSION

In this proposed work, implement a Malicious URL Detection process using machine learning techniques. This focuses on detecting unsafe website URLs and keywords with the help of encrypted blacklist storage. According to few selected features can be used to differentiate between legitimate and malicious web pages. These selected features are many such as URLs and Keywords. In proposed work a service provider that owns a high-quality blacklist, which may be more frequently updated or simply contains more items. User also allowed to directly sharing blacklists with servers in an uncontrollable way could make these dataset be obtained by every user. With the help of efficient classification approach will detect the fake websites accurately and prevent the users from accessing that websites. This also provides the secure encryption approach avoid the unknown access of search history. The security is provided to the search data which has been stored in the database.

6. REFERENCES

[1] Mourtaji, Youness, Mohammed Bouhorma, DaniyalAlghazzawi, GhadahAldabbagh, And Abdullah Alghamdi. "Hybrid Rule-Based Solution for Phishing Url Detection Using Convolutional Neural Network." Wireless Communications And Mobile Computing 2021 (2021): 1-24.

[2] Ahammad, SkHasane, Sunil D. Kale, Gopal D. Upadhye, SandeepDwarkanathPande, E. VenkateshBabu, Amol V. Dhumane, And MrDilip Kumar Jang Bahadur. "Phishing Url Detection Using Machine Learning Methods." Advances In Engineering Software 173 (2022): 103288.

[3] Yuan, Jianting, Guanxin Chen, ShengweiTian, AndXinjun Pei. "Malicious UURL Detection based on a Parallel Neural Joint Model." Ieee Access 9 (2021): 9464-9472.

[4] Xu, Pingfan. "A Transformer-Based Model to Detect Phishing Urls." Arxiv Preprint Arxiv:2109.02138 (2021).

[5] Butnaru, Andrei, AlexiosMylonas, AndNikolaosPitropakis. "Towards Lightweight Url-Based Phishing Detection." Future Internet 13, No. 6 (2021): 154.

[6] Odeh, Ammar, Ismail Keshta, AndEmanAbdelfattah. "Machine Learning

techniques for Detection Of Website Phishing: A Review For Promises And Challenges." In 2021 Ieee 11th Annual Computing And Communication Workshop And Conference (Ccwc), Pp. 0813-0818. Ieee, 2021.

[7] Butt, Muhammad HassaanFarooq, Jian Ping Li, TehreemSaboor, Muhammad Arslan, And Muhammad Adnan Farooq Butt. "Intelligent Phishing Url Detection: A Solution Based On Deep Learning Framework."In 2021 18th International Computer Conference On Wavelet Active Media Technology And Information Processing (Iccwamtip), Pp. 434-439.Ieee, 2021.

[8] Tang, Lizhen, AndQusay H. Mahmoud. "A Survey of Machine Learning-Based Solutions for Phishing Website Detection." Machine Learning And Knowledge Extraction 3, No. 3 (2021): 672-694.

[9] Purbay, Madhurendra, AndDivya Kumar. "Split Behavior of Supervised Machine Learning Algorithms for Phishing URL Detection." In Advances In Vlsi, Communication, And Signal Processing: Select Proceedings Of Vcas 2019, Pp. 497-505. Springer Singapore, 2021.

[10] Wazirali, Raniyah, Rami Ahmad, And Ashraf Abdel-Karim Abu-Ein. "Sustaining Accurate Detection Of Phishing Urls Using Sdn And Feature Selection Approaches." Computer Networks 201 (2021): 108591