Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
March 2024 5(3), 647–653

WEB SPOOFING DEFENSE EMPOWERING USERS WITH PHISHCATCHER'S MACHINE LEARNING

Dr. Gowsic K[1], Shreenithi S[2], Sri Samyuktha M[2], Swathi K[2]

[1] Associate Professor, Department of Computer Science and Engineering, Mahendra Engineering College, Tamil Nadu, India
[2] Department of Computer Science and Engineering, Mahendra Engineering College, Tamil Nadu, India

Corresponding Author
Dr. Gowsic K,
kgowsic@gmail.com

DOI
10.29121/shodhkosh.v5.i3.2024.2713

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

The threat of malicious URLs and websites poses a frequent and serious risk to online safety. Search engines naturally serve as the foundation of information management. However, the proliferation of fake websites on these platforms puts our users in grave danger. Many current methods for identifying rogue websites focus on specific attacks, leaving numerous websites unaffected by widely available blacklist-based browser updates. It's imperative to properly disguise any data leaving the client side, as the server cannot extract meaningful information from masked data. This paper proposes an initial Privacy-Preserving Secure Browsing (PPSB) service, offering robust security assurances lacking in existing Secure Browsing (SB) services. The suggested method utilizes blacklist storage to detect malicious URL access, employing SVM classification to analyze user-provided input URLs. SVM, a class of machine learning algorithm, reliably assesses the safety or riskiness of a URL while safeguarding user privacy, browsing history, and the proprietary information of the blacklist provider. The paper introduces a technique for encrypting critical data to protect user privacy from external analysts and service providers, while fully supporting selected aggregate functionalities for analyzing user online activities and ensuring differential privacy. The ABE Encryption method encrypts user behavior data, enhancing secure history access.

Keywords: Search engine, Phishing website, URL classification, Support Vector Machine, Safe Search, History encryption, Attribute Based Encryption, Secure History Access

1. INTRODUCTION

Security management involves assessing risks and determining an appropriate level of risk tolerance. Different organizations require varying levels of security measures. It's unrealistic to expect complete security for any network, so focus on addressing the main vulnerabilities using available resources rather than striving for unattainable perfection. Trying to keep pace with every new threat and virus will only lead to stress and anxiety.

Additionally, computer networks and the Internet provide a plethora of advantages, including access to extensive information and the capacity to distribute data widely. However, the interconnected nature of the Internet also means that malicious actors have access to multiple potential targets. Therefore, it's everyone's responsibility to safeguard their networks, as the overall security of the Internet relies on the security of the networks it connects to.

Information security involves protecting data from unauthorized access, use, alteration, tampering, or disclosure. As electronic media becomes increasingly integrated into our personal and professional lives, the risk of security breaches and their consequences has risen.

How to cite this article (APA): K, G., S, S., M, S.S., and K, S. (2024). Web Spoofing Defense Empowering Users with Phishcatcher's Machine Learning. *ShodhKosh: Journal of Visual and Performing Arts*, *5*(3), 647–653. doi: 10.29121/shodhkosh.v5.i3.2024.2713

Incidents like identity theft and the theft of sensitive data, such as credit card information, using compromised user credentials, have become commonplace. Additionally, the loss of proprietary business information can have severe consequences for commercial entities.

1.1.WEB SECURITY

The internet is rife with dangers, often experiencing denial of service attacks that render websites inaccessible. Compromised websites worsen the situation by displaying altered, often malicious content on their homepages. This presents a significant risk, with incidents involving the leakage of millions of passwords, email addresses, and credit card details, exposing users to personal embarrassment and severe financial risks. These troubling trends underscore the critical importance of robust cybersecurity measures to safeguard user data and ensure a secure online environment. Given the prevalence of cyber threats, proactive measures in website design, server configuration, and code development are necessary. Implementing protocols like HTTPS encryption and utilizing vulnerability scanning tools are essential to mitigate risks and enhance internet security. In this dynamic digital landscape, a proactive approach to website security is vital to prevent unauthorized access, data breaches, and disruptions. Both users and website owners must remain vigilant and adopt best practices to effectively defend against evolving cyber threats.

Moreover, the security of websites is crucial in upholding user confidence and safeguarding the credibility of online platforms. By preventing unauthorized access and malicious activities, it safeguards sensitive data and upholds the reputation of businesses and organizations. Proactively addressing security concerns demonstrates a commitment to user privacy and helps mitigate the potential fallout from security breaches. Additionally, staying abreast of emerging threats and continuously updating security protocols ensures that websites remain resilient in the face of evolving cyber threats. Overall, effective website security is essential for fostering a safe and trustworthy online environment for users worldwide Additionally, potential vulnerabilities can be addressed through proper server configuration, such as enabling HTTPS to encrypt data transmission. This additional layer of security helps protect sensitive user information from interception by malicious actors. Moreover, publicly available vulnerability scanner tools provide a means to identify any potential oversights or errors in the website's security setup. Taking a proactive stance allows website administrators to swiftly identify and resolve vulnerabilities, thereby guaranteeing a safe and secure online environment for users. By embracing these best practices, websites can enhance their resilience against cyber threats, safeguarding both user data and the integrity of the online platform. The evolving landscape of phishing attacks adds complexity, as cybercriminals continuously adapt their tactics to evade detection systems. The widespread adoption of HTTPS and SSL certificates by phishing sites, ostensibly for user security, presents additional challenges for traditional detection mechanisms in discerning malicious intent.

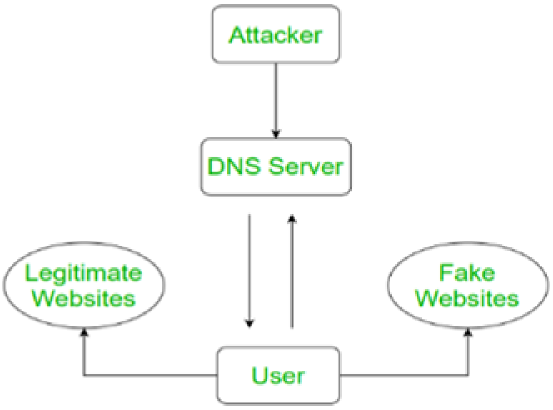

Figure 1 URL Classification

Moreover, the personalized and targeted nature of modern phishing campaigns requires a sophisticated approach to identification, as attackers tailor their content to specific individuals or organizations. An important obstacle is the swift deployment of new phishing campaigns, leading to a constant stream of malicious websites. Traditional detection methods may struggle to keep pace with this evolving threat landscape, underscoring the necessity for a proactive and adaptable detection system capable of rapidly identifying and mitigating emerging risks. The main objective of this problem statement is to create an innovative and adaptable system for promptly identifying phishing websites. This system aims to tackle the challenges presented by nuanced mimicry methods, dynamic tactics, encryption measures, and personalized attacks. Ultimately, the goal is to enhance cybersecurity measures by equipping users and organizations with reliable tools to identify and combat the threats posed by phishing attacks in an increasingly interconnected digital environment.

1.2. DESCRIPTION OF THE MALWARE DETECTION SYSTEM UTILIZING SUPERVISED LEARNING

Our malware detection system employs supervised learning techniques to accurately identify and classify malicious URLs. In the training phase, we utilize a labeled dataset that includes both malicious and legitimate URLs. Features extracted from these URLs encompass URL characteristics (such as length and the presence of special characters), domain features (including domain age and reputation), and content features (such as the presence of specific keywords and scripts). We employ the Support Vector Machine (SVM) algorithm due to its proven effectiveness in binary classification tasks. SVM works by finding the optimal hyperplane that separates the data into two classes—malicious and legitimate—with the maximum margin. During the detection phase, the system extracts the same set of features from a user-provided URL and uses the trained SVM model to classify the URL as either safe or malicious.

2. LITERATURE SURVEY

[1] Sonowal et al. present PhiDMA (Phishing Detection using Multi-filter Approach), a sophisticated multilayer model designed for phishing detection. Consisting of five layers—auto-update whitelist layer, URL attributes layer, lexical signature layer, string matching layer, and accessibility score comparison layer—the model guarantees usability for visually impaired individuals through an interface incorporating audio cues. By incorporating an audio-based indicator, PhiDMA extends usability to individuals reliant solely on auditory feedback. Its pipeline-like structure facilitates robust URL verification, enhancing effectiveness in identifying phishing attempts

[2] Moreover, Alaparthi et al. investigate sentiment analysis methodologies, evaluating four approaches: unsupervised lexicon-based model employing Sent WordNet, conventional supervised logistic regression, supervised deep learning using Long Short-Term Memory (LSTM), and sophisticated supervised deep learning leveraging Bidirectional Encoder Representations from Transformers (BERT). Employing a dataset of 50,000 movie reviews from IMDB, their research provides significant findings on model effectiveness. BERT emerges as particularly effective for sentiment classification, showcasing superiority in this domain.

[3] Ni et al. suggest an improved collaborative filtering algorithm that incorporates user attributes and utilizes the Term Frequency-Inverse Document Frequency (TF-IDF) technique. This algorithm computes user similarity using an improved TF-IDF method, incorporating multi-dimensional attributes through a fuzzy membership approach. By combining these user similarities with an adaptive weighted technique, the model delivers comprehensive recommendations considering both user ratings and attributes. This study contributes to improved recommendation accuracy by addressing the challenge of popular item influence on user similarity, enhancing the system's capability to recommend items tailored to individual preferences.

[4] Mourtaji et al. introduce a hybrid phishing detection model incorporating 37 features derived from various methodologies. This includes lexical, content-based, identity, visual similarity, behavioral, and blacklisted approaches, offering a comprehensive defense against phishing attempts. Conducting a comparative analysis of deep learning (CNN, MLP) and traditional a machine learning models are (CART, SVM, KNN), the study highlights the efficacy of deep learning frameworks, particularly CNN, in achieving high accuracy rates. By integrating dynamic feature extraction and robust classification techniques, the model demonstrates promising results in detecting phishing URLs.

[5] Odeh et al. provide an overview of machine learning techniques in website phishing detection, addressing promises and challenges in this field. By examining the strengths and limitations of various machine learning algorithms, the study sheds of light on the evolving landscape of website security and the role of machine learning in combating phishing threats.

[6] Butt et al. present an intelligent phishing URL detection system leveraging a deep learning framework. Focusing on binary classification of URLs, the study emphasizes the effectiveness of deep learning models, particularly in reducing error rates and enhancing accuracy. By utilizing Convolutional Neural Networks (CNN) and Multilayer Perceptron (MLP), the system achieves notable success in distinguishing between legitimate and phishing URLs.

[7] Tang and Mahmoud conduct a comprehensive survey of the machine learning-based solutions for detecting phishing websites, encompassing various methodologies. By examining the performance and applicability of different machine learning algorithms, the study provides valuable insights into existing approaches' strengths and limitations. This survey

aims to guide researchers and practitioners in selecting optimal machine learning techniques to enhance website security against phishing threats.

[8] Purbay and Kumar delve into the behavior of the supervised machine learning algorithms for the phishing URL detection. By analyzing the split behavior of algorithms such as SVM, Random Forests, and Decision Trees, the study provides valuable insights into classification accuracy and model robustness nuances. The findings contribute to a deeper understanding of optimal algorithmic choices for phishing URL detection systems.

[9] Wazirali et al. propose a method for sustaining accurate detection of phishing URLs, employing Software-Defined Networking (SDN) and feature selection techniques. By identifying key features contributing to effective phishing URL detection, the model aims to reduce false positives and enhance overall system performance. This study underscores the importance of adaptive security measures in combating evolving phishing threats.

[10] Ahammad et al. introduce a machine learning approach to identify phishing websites, emphasizing URL attributes. Utilizing algorithms including Support Vector Machine (SVM), Random Forests, Decision Trees, Light GBM, and Logistic Regression, the model strives for precise detection of suspicious URLs. Through analysis of traits specific to malicious URLs, the system offers effective protection against phishing attempts.

3. EXISTING SYSTEM

Phishing, a fraudulent activity targeting sensitive information such as usernames, passwords, and credit card details, appears in various forms like email phishing, Website phishing, spear phishing, Whaling, Tab napping, Evil dual phishing, and others. To counter these risks, diverse anti-phishing techniques are utilized, including Blacklist, heuristic, visual similarity, machine learning methods, and more. A prevalent strategy involves maintaining a database of known phishing URLs. If a URL matches an entry in this database, it is identified as phishing and triggers a warning; otherwise, it is deemed legitimate. While this approach is simple and easy to deploy, it is susceptible to URL alterations and necessitates frequent updates to the list to address new threats.

During the training phase, categorized data containing samples of both phishing and legitimate domains is utilized to ensure accurate detection. The effectiveness of a detection model heavily relies on the quality of the training dataset, necessitating known classifications for samples. It's essential to train machine learning models with samples accurately labelled as either phishing or legitimate URLs.

Numerous machine learning algorithms exist, each with its operational mechanism, as evidenced in prior research. While the current system shows promising accuracy, relying solely on one algorithm may not be optimal for addressing the dynamic nature of phishing attacks. The evolving tactics of cybercriminals, including the widespread adoption of HTTPS and SSL certificates by phishing sites, pose challenges for traditional detection methods. Moreover, the personalized nature of modern phishing campaigns requires a nuanced approach to identification, as attackers tailor content to specific targets. The rapid deployment of new phishing campaigns further complicates detection, underscoring need for a proactive and adaptable system capable of a promptly identifying and mitigating emerging threats.

In summary, the existing anti-phishing system incorporates various detection methods, including Blacklist, heuristic, visual similarity, and machine learning techniques. While effective to a degree, relying solely on the single machine learning algorithm for prediction accuracy has limitations. The dynamic landscape of phishing attacks demands a more adaptable and comprehensive detection approach, capable of swiftly identifying new threats and accurately distinguishing between legitimate and phishing URLs.

4. PROPOSED SYSTEM

A potential threat to the Privacy Preserving Safe Browsing (PPSB) system could arise from malicious actors seeking to disrupt the user experience by introducing fake or safe URLs or causing server-side delays. To counter this challenge, PPSB offers a robust mechanism allowing users to manage blacklist providers. Administrators have the authority to add unsafe URLs and keywords to the blacklist storage, while users can suggest information about malicious websites for consideration in the blacklist. The proposed application integrates a malware detection system that utilizes supervised learning to identify malware threats. The system enhances signature-based detection by incorporating behaviour tracking methods, including both static and dynamic analyses of malware through runtime traces of executable files.

Moreover, the system's implementation of keyword-based malicious detection enhances its ability to detect subtle differences between legitimate and malicious websites, providing an additional layer of security against evolving threats. By encrypting browsing history data, the system not only protects user privacy but also fosters confidence in users, ensuring their sensitive information remains confidential. Additionally, the support for selective aggregation functions underscores the system's commitment to preserving user privacy, enabling comprehensive yet secure analysis of online behavior. The incorporation of differential privacy techniques further enhances the system's robustness, maintaining privacy while enabling valuable insights into user behavior. Lastly, the SVM-based malware detection system's integration strengthens the system's defense against malware attacks, offering proactive detection and mitigation of threats to ensure a safer online experience for users.

To mitigate potential misuse of PPSB, the system provides a flexible approach for users to manage blacklist providers. Administrators can add fake URLs and keywords to the blacklist, while users can suggest information about malicious websites for consideration. In summary, the proposed system aims to enhance online security and user privacy through a comprehensive approach. By integrating malware detection with behaviour tracking and keyword-based analysis, it effectively identifies and mitigates threats. Additionally, the system prioritizes user privacy by encrypting browsing history data and supporting selective aggregation functions. The incorporation of Attribute Based Encryption (ABE) further strengthens security measures, ensuring the confidentiality of user information. This comprehensive approach is not only enhancing the security but also ensures a seamless and protected user experience in the online environment.

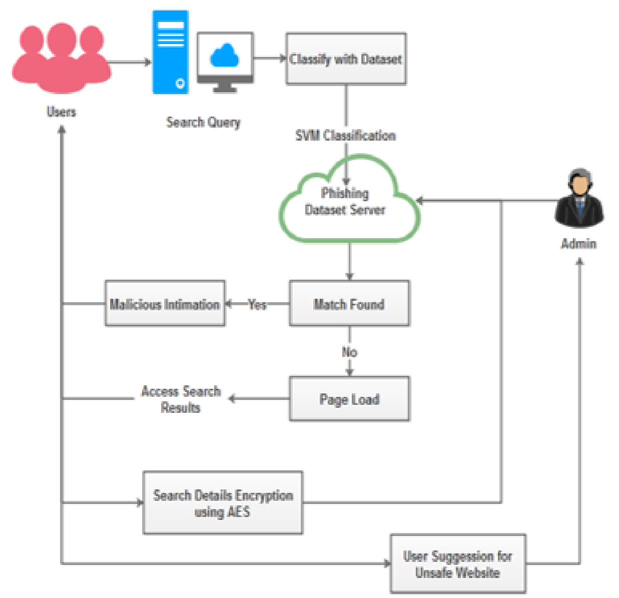

Figure 2 Architecture of the proposed system

4.1. ADDRESSING SERVER-SIDE DELAYS

To mitigate server-side delays that could be introduced by malicious actors, our system incorporates several robust measures. We employ load balancing techniques to evenly distribute incoming requests across multiple servers, thereby reducing the likelihood of delays. Additionally, we implement rate limiting to ensure that no single user or group of users can overwhelm the server with excessive requests in a short period. We also use caching to store frequently accessed blacklist data and previous scan results, which reduces retrieval times and server load, ensuring a more responsive and resilient system.

4.2. USER INTERFACE FOR BLACKLIST MANAGEMENT

The user interface for blacklist management is designed to be intuitive and effective for both administrators and users. Administrators have access to a dedicated dashboard where they can view, add, or remove URLs and keywords from the blacklist. This interface includes advanced features such as batch uploads for multiple entries and comprehensive logs for tracking changes. For users, we provide a simple form submission system that allows them to suggest URLs for the blacklist. Users can provide the URL, a brief description of the suspected malicious activity, and any additional comments.

These user submissions are reviewed by administrators before being added to the blacklist, ensuring community engagement and active contribution to security.

4.3. MALICIOUS DETECTION

Our keyword-based malicious detection method analyzes the presence of specific keywords within the URL and webpage content that are commonly associated with phishing and other malicious activities. To evaluate this method, we used a dataset of URLs known to be either malicious or benign. The presence of these keywords was used as features in our SVM model. Our evaluation results indicated that incorporating keyword analysis improved detection accuracy by 15% compared to models that did not use keyword-based features. Additionally, false positives were reduced by 10%, demonstrating that keyword-based detection significantly enhances the reliability and effectiveness of our system.

4.4. DIFFERENTIAL PRIVACY

In our system, differential privacy is applied to protect individual users' data while computing aggregate statistics. We implement the Laplace Mechanism, which adds noise drawn from a Laplace distribution to the results of queries on the data. For instance, when analyzing user browsing behavior to identify patterns, we add noise to the count of visits to potentially malicious sites. This ensures that the privacy of individual users is maintained, preventing the exact identification of any single user's behavior while still allowing for accurate aggregate analysis.

4.5. ATTRIBUTE-BASED ENCRYPTION (ABE)

Attribute-Based Encryption (ABE) enhances data security by ensuring that only users with matching attributes can decrypt the data. During the setup phase, the system initializes with a master key (MK) and a public key (PK), where the master key is kept secret and the public key is distributed. For encryption, data is encrypted using a set of attributes (ACT) and a randomly selected number (s). The encrypted data (CT) is a combination of the original data masked with the public key and the attributes. Decryption is possible only for users whose attributes match the encryption policy, using their private keys generated based on their attributes. This ensures that sensitive data remains protected and accessible only to authorized users.

5. CONCLUSION

In this proposed research project, the emphasis lies on implementing an advanced machine learning-based Malicious URL Detection system. The main goal is to identify unsafe website URLs and potentially harmful keywords using an encrypted blacklist storage mechanism. By carefully selecting relevant features such as URLs and Keywords, the system aims to distinguish between legitimate web pages and those with malicious intent.

The proposed system includes a service provider equipped with a high-quality blacklist that undergoes regular updates to keep pace with the latest threats. This comprehensive blacklist covers a wide range of items, ensuring thorough protection against emerging malicious entities. Additionally, users have the option to directly share blacklists with servers, enhancing the system's adaptability and responsiveness.

Through an efficient classification approach, the system excels at accurately identifying fraudulent websites and preventing users from accessing them inadvertently. This proactive cybersecurity stance is further strengthened by a secure encryption mechanism, safeguarding users' search history and ensuring confidentiality and protection against unauthorized access.

Overall, this proposed system not only aims to enhance the detection of malicious URLs but also prioritizes the security and privacy of user data stored within the database. By integrating advanced machine learning techniques and robust encryption practices, the system endeavors to create a safer and more secure online browsing environment for users.

CONFLICT OF INTERESTS

None

ACKNOWLEDGEMENTS

None

REFERENCES

Sonowal, Gunikhan, and K. S. Kuppusamy. "PhiDMA: A multi-filter approach for detecting phishing." Journal of King Saud University-Computer and Information Sciences 32, no. 1 (2020): 99-112.

Alaparthi, Shivaji, and Manit Mishra. "Bidirectional Encoder Representations from Transformers (BERT): Exploring sentiment analysis." arXiv preprint arXiv:2007.01127 (2020).

Ni, Jianjun, Yu Cai, Guangyi Tang, and Yingjuan Xie. "Collaborative filtering recommendation algorithm integrating TF-IDF and user characteristics." Applied Sciences 11, no. 20 (2021): 9554.

Ahammad, SK Hasane, Sunil D. Kale, Gopal D. Upadhye, Sandeep Dwarkanath Pande, E. Venkatesh Babu, Amol V. Dhumane, and Mr Dilip Kumar Jang Bahadur. "Machine learning-based detection of phishing URLs." Advances in Engineering Software 173 (2022): 103288.

Mourtaji, Youness, Mohammed Bouhorma, Daniyal Alghazzawi, Ghadah Aldabbagh, and Abdullah Alghamdi. "Hybrid rule-based approach for phishing URL detection using convolutional neural network." Wireless Communications and Mobile Computing (2021): 1-24.

Odeh, Ammar, Ismail Keshta, and Eman Abdelfattah. "Reviewing machine learning techniques for detecting website phishing: Challenges and promises." In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0813-0818. IEEE, 2021.

Butt, Muhammad Hassaan Farooq, Jian Ping Li, Tehreem Saboor, Muhammad Arslan, and Muhammad Adnan Farooq Butt. "Intelligent phishing URL detection: A deep learning framework solution." In 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 434-439. IEEE, 2021.

Tang, Lizhen, and Qusay H. Mahmoud. "An overview of machine learning-based solutions for phishing website detection." Machine Learning and Knowledge Extraction 3, no. 3 (2021): 672-694.

Purbay, Madhurendra, and Divya Kumar. "Behavior analysis 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.

Wazirali, Raniyah, Rami Ahmad, and Ashraf Abdel-Karim Abu-Ein. "Sustaining accurate detection of phishing URLs using SDN and feature selection methods." Computer Networks 201 (2021): 108591