IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Smart Detection Designs Of Html Web Page Url Phishing Attacks Based On Natural Language Processing

Mohammed Mubeen[1], Md. Ateeq Ur Rahman[2], Jothikumar. R[3]

[1]Research Scholar, Department of Computer Science and Engineering, SCET, Hyderabad, Telangana

[2]Professor, Department of Computer Science and Engineering, SCET, Hyderabad, Telangana.

[3]Professor, Department of Computer Science and Engineering, SCET, Hyderabad, Telangana.

ABSTRACT

Phishing attacks are a type of cybercrime that has grown in recent years. It is part of social engineering attacks where an attacker deceives users by sending fake messages using social media platforms or emails. Phishing attacks steal users' information or download and install malicious software. They are hard to detect because attackers can design a phishing message that looks legitimate to a user. This message may contain a phishing URL so that even an expert can be a victim. This URL leads the victim to a fake website that steals information, such as login information, payment information, etc. Researchers and engineers work to develop methods to detect phishing attacks without the need for the eyes of experts. Even though many papers discuss HTML and URL-based phishing detection methods, there is no comprehensive survey to discuss these methods. Therefore, this paper comprehensively surveys HTML and URL phishing attacks and detection methods. We review the current state-of-art machine learning models to detect URL-based and hybrid-based phishing attacks in detail. We compare each model based on its data preprocessing, feature extraction, model design, and performance.

INTRODUCTION

Phishing attacks are cybercrimes that use social engineering to trick people into divulging personal information, bank account information, and other sensitive data. Attackers can use social media sites like Twitter and Facebook or email services like Gmail and Outlook to deliver phony messages to victims while posing as reliable sources. When users download attachments or enter personal information, they become susceptible. Attacks on social media platforms have increased recently since it is simple for attackers to post a single message and reach a large number of individuals worldwide. The Anti-Phishing Working Group (APWG) says that in January 2021, there were 250,000 more phishing attacks in a single month. Furthermore, there were 56% more business concessions than there were in from 2020's final quarter to 2021's first quarter. In 2021, financial institutions, social media, and web emails are the industries most likely to be targeted. Attackers mostly target the financial industry and social media platforms in an attempt to obtain the financial information or identities of their targets. Malicious software that initiates more cyberattacks, like ransomware and malware attacks, may also be sent by attackers. The rise in phishing assaults in recent times, along with the associated cybersecurity risks, has made it imperative to address this issue. The majority of modern businesses rely on human expertise to identify these intrusions. Even an expert can find it difficult to distinguish between phishing assaults because of the similarities between phony and authentic mails. Consequently, cybersecurity professionals focus more onproposed several solutions in recent years with high accuracy to detect phishing attacks, such as blacklist traditional machine learning and Deep Learning (DL). We provide a brief analysis of each solution as follows. • Blacklists are lists of websites' URLs that are most likely phishing websites. The systems block all URLs or IPs in this list. However, this method has a significant drawback. A system must have a phishing attack URL to block it; it does not detect it if the URL is not on the list. Phishing attack detection is achieved by using conventional machine learning models. Conventional machine-learning algorithms, however, require feature extraction by hand. Therefore, extracting a set of features takes time and human labor. These functions rely on the URLs that are accessible. As a result, the feature analysis and extraction process is increased when attackers create new phishing URLs, resulting in a large feature dimension. Despite its best efforts to analyze vast feature sets and high dimensions, it is still vulnerable to attacks from new phishing URLs. • The benefit of employing a deep learning method to identify phishing URLs is that a model can automatically extract the characteristics for both text and images, saving human labor. However, various issues arise because to the phishing attempts' clever design andthe phishing website produced using the most recent DL techniques. A model is trained, for instance, to recognize lengthy URLs. However, small URLs are not detected by it. Moreover, DL has several disadvantages, such as the need for a sizable dataset for model testing, validation, and training. Because DL models are sophisticated, it is also expensive. Phishing attacks can be identified using a variety of data sources,

including hybrid, content-based, and URL-based data. URL-based techniques retrieve URL data without examining any other elements, including webpage content, title, etc. One benefit of URL-based phishing detection is that it can identify a message without requiring the user to click on the URL and run the risk of downloading and installing malicious software. But collecting simply elements based on URLs leaves out important aspects of the phishing attack.webpage, including the page code and title. Using simply URL-based methods to analyze small URLs is also challenging. Techniques that are content-based retrieve data from webpage content, including text, JavaScript, graphics, and HTML code. However, the content-based approach forces users or systems to access the webpage and extract content, which increases the risk of an attack through the download and installation of malicious software. The URL-based and content-based capabilities are combined in the hybrid-based content feature.

OBJECTIVE

This project aims to do a thorough survey and analysis of phishing assaults based on HTML and URL, with a particular focus on the creation and assessment of machine learning models for automated detection. The study looks into state-of-the-art approaches in an effort to combat the growing sophistication of phishing techniques. The scope encompasses an in-depth investigation of feature extraction approaches, data preparation strategies, model design considerations, and performance metrics used in machine learning models that are currently in use for phishing detection. The goal is to present a thorough understanding of these models' advantages and disadvantages by contrasting them according to a number of criteria. In addition, the project aims to make a contribution to the field by pointing out areas of current research that need improvement and suggesting new directions for future research.

PROBLEM STATEMENT

The problem statement revolves around the increasing prevalence and sophistication of phishing attacks, a type of cybercrime that exploits social engineering techniques to deceive individuals into revealing sensitive information. Phishing attacks often masquerade as trustworthy entities through various communication channels such as social media platforms and email services, posing a significant risk to individuals and organizations alike. The Anti-Phishing Working Group (APWG) reported a substantial rise in phishing attacks, particularly targeting industries like finance, social media, and web emails. The core issue lies in the difficulty of accurately detecting phishing attempts due to the evolving tactics employed by attackers. Traditional methods like blacklists have limitations, as they rely on known phishing URLs and may fail to detect new or previously unseen attacks. Moreover, conventional machine learning approaches for phishing detection require manual feature extraction, which can be time-consuming and labor-intensive, especially when dealing with large feature dimensions resulting from new phishing URLs.

Therefore, the problem statement encompasses the urgent need for robust and efficient solutions to combat phishing attacks effectively. These solutions should address the challenges of detecting evolving phishing tactics, automating feature extraction processes, and improving overall accuracy and reliability in identifying phishing attempts across various communication channels.

EXISTING SYSTEM

Various machine learning models have been investigated for phishing detection in an effort to thwart the attackers' ever-evolving tactics. Decision tree-based algorithms, like J48 or C4.5, are one often used method that separates authentic from phishing websites by analyzing features taken from URLs and HTML content.

Some machine learning models have shown efficacy through the mapping of input features into a high-dimensional space, which makes it easier to identify intricate patterns that suggest phishing.

Disadvantage of Existing System

Decision tree-based algorithms like J48 and C4.5 are widely used for phishing detection due to their ability to analyze features from URLs and HTML content. However, they come with several disadvantages. Firstly, they are prone to overfitting, especially with complex datasets, leading to reduced generalization performance. Secondly, decision trees exhibit high variance, resulting in different tree structures and predictions for similar datasets, making them less robust. Thirdly, their interpretability diminishes with complex trees, making it challenging to extract actionable insights. Moreover, decision trees can exhibit bias towards majority classes in imbalanced datasets, impacting the accuracy of detecting rare phishing instances. Lastly, they are sensitive to small changes in data or features, affecting the stability of the model, especially in dynamic environments. These limitations highlight the need for careful parameter tuning, feature selection, and ensemble methods to mitigate these drawbacks and enhance phishing detection performance.

PROPOSED SYSTEM

New methods are always being suggested to improve phishing detection abilities. Combining machine learning and feature engineering methods to create hybrid models shows potential. For example, merging machine learning techniques with lexical and host-based information results in a more comprehensive knowledge of phishing characteristics.

Feature-rich models seek to capture a wider range of malicious actions by incorporating lexical, content, and host-based data. Additionally, by identifying subtle trends in phishing messages, advances in ML models and natural language processing techniques offer prospects for enhanced detection. These suggested algorithms make use of a variety of techniques to improve accuracy and resistance to complex phishing schemes.Improving phishing detection is a crucial endeavor, and combining machine learning with feature engineering offers promising avenues. Hybrid models that merge machine learning techniques with lexical and host-based information can indeed enhance our understanding of phishing characteristics. These feature-rich models, by incorporating lexical, content, and host-based data, aim to capture a wider range of malicious actions, thereby improving detection accuracy and resistance to complex phishing schemes.

Advances in machine learning models and natural language processing techniques further contribute to this goal by enabling the identification of subtle trends in phishing messages. By leveraging these techniques, algorithms can be designed to improve accuracy and robustness in detecting phishing attempts, even in sophisticated scenarios.

Advantages of Proposed System

- **Defying Intricate Phishing Schemes**: Feature-rich models are made to catch a variety of malicious activities, such as intricate phishing schemes that might elude detection by less complex techniques. These models improve resistance to changing phishing techniques by combining various data sources, including host-based information, content analysis, and lexical cues.
- **Scalability:** Large datasets can be used to train machine learning models, which makes phishing detection systems scalable. Scalable models are crucial for efficacious protection techniques as phishing attempts grow in volume and complexity.
- **Adaptability:** Machine learning models are appropriate for dynamic contexts where phishing methods are always changing since they can adjust and learn from new data. Because of its flexibility, the detection system can keep up with new threats.

RELATED WORK

Research in URL phishing detection has evolved significantly, leveraging machine learning algorithms, lexical analysis, feature engineering, behavioral analysis, ensemble methods, deep learning approaches, online learning techniques, and feature selection methods. Machine learning-based URL analysis focuses on extracting features like domain reputation, URL structure, and similarity to known phishing domains for training models. Lexical analysis parses URLs to identify patterns such as look-alike domains or suspicious keywords. Feature engineering designs features capturing domain attributes and URL characteristics crucial for distinguishing legitimate and phishing URLs. Behavioral analysis complements lexical and structural analysis by monitoring URL content changes and detecting malicious scripts. Ensemble methods combine multiple detection techniques for improved performance. Deep learning models like CNNs and RNNs learn complex URL patterns effectively. Online learning adapts models to new data and emerging threats, while feature selection enhances model interpretability and performance. These diverse approaches collectively advance URL phishing detection, enhancing accuracy and resilience against evolving threats.

METHODLOGY OF PROJECT

The methodology for URL phishing detection integrates cutting-edge techniques from machine learning, lexical analysis, feature engineering, and behavioral analysis to create a comprehensive and robust detection framework. Initially, a feature-rich representation of URLs is constructed, incorporating domain reputation metrics, structural attributes, and lexical patterns indicative of phishing behavior. Advanced machine learning algorithms, including ensemble methods and deep learning architectures such as CNNs and RNNs, are deployed to train models on this feature space, optimizing for high accuracy and low false positive rates. Simultaneously, lexical analysis techniques parse URLs to extract subtle cues, such as look-alike domains and suspicious keywords, further enhancing the model's discriminatory power. Behavioral analysis plays a pivotal role by dynamically monitoring URL behavior, detecting anomalies, and flagging potentially malicious scripts or content alterations. Online learning methodologies ensure continuous adaptation to emerging threats, maintaining the system's efficacy in real-time scenarios. Feature selection strategies are employed to identify key discriminative features, streamlining model complexity and enhancing interpretability. This integrated methodology leverages the latest advancements in machine learning and cybersecurity, offering a sophisticated and proactive approach to URL phishing detection.

MODULE's

1) Data Collection:

A dataset is an ordered set of data that is typically arranged in rows and columns. Each row in the dataset represents a single observation or instance, and each column a particular attribute or feature of that instance. Spreadsheets, databases, text files, and other customized formats for particular uses are just a few of the many formats that datasets can take.

2) Data Purification:

The process of finding and fixing mistakes or inconsistencies in a dataset to raise its analytical quality and dependability is known as data cleaning. It entails activities including dealing with outliers, handling missing values, eliminating duplicates, fixing errors, and standardizing formats.

3) Feature extraction from NLP:

Extracting characteristics from text using Natural Language Processing (NLP) entails converting text data into numerical or categorical features that may be utilized for machine learning tasks.

4) Machine Learning Models:

These computational algorithms use data to identify patterns and relationships in order to forecast or make decisions. Regression, classification, clustering, and deep learning are just a few of the methods they use. Trained on labeled or unlabeled datasets, they solve particular tasks and generalize patterns to enable automated decision-making across a range of domains.

5) Train Model:

To minimize prediction errors, a model is trained by feeding it a dataset to identify patterns and relationships. The model's parameters are then adjusted using optimization algorithms like gradient descent. The efficacy of the trained model in producing precise predictions or classifications is then assessed by evaluating its performance on an independent validation dataset.

6) Test and Deployment: Testing entails evaluating a model's performance on hypothetical data to make sure it meets target accuracy thresholds and generalizes well. The trained model must be integrated into production systems, be able to make predictions in real time, and have its performance continuously monitored for maintenance and optimization.

Benefits

- Provides comprehensive insights for informed decision-making. Accesses diverse data sources for a holistic understanding.
- Enhances data quality, ensuring accuracy and consistency. Reduces errors and improves the reliability of analyses.
- Automates decision-making processes, improving efficiency. Provides predictive capabilities for identifying patterns and trends.Optimizes performance through parameter tuning. Scale for handling real-time predictions and large data volumes.

FLOW DIAGRAMS:

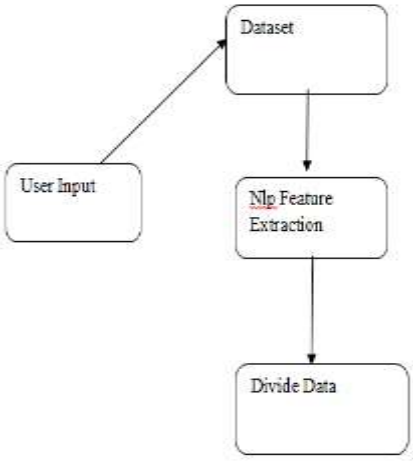


Fig- 1: Data Flow Diagram

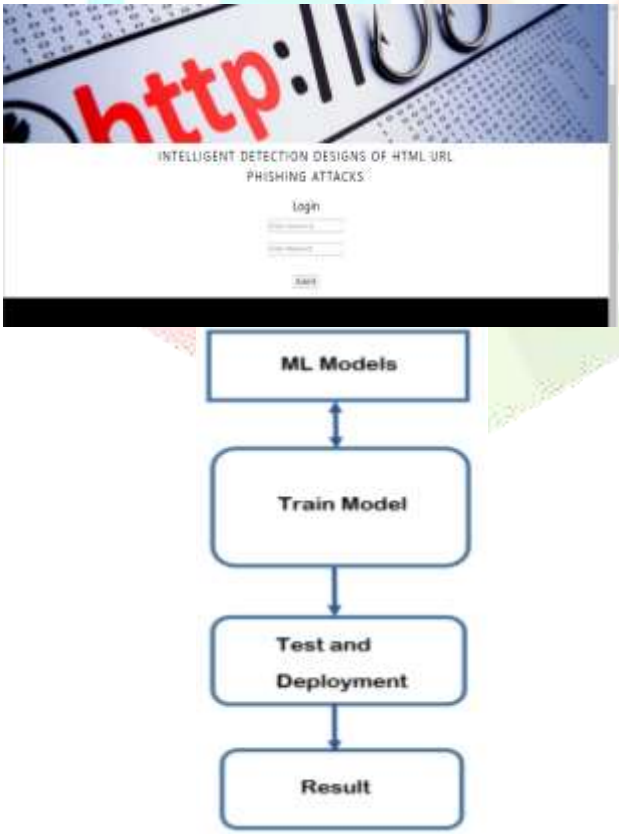


Fig- 2: Flow Diagram of Modules

Impediments of DL

One of the primary impediments in tackling URL phishing attacks using machine learning (ML) and natural language processing (NLP) is the dynamic and evolving nature of phishing tactics. Phishing attackers constantly modify URLs, use obfuscation techniques, and employ social engineering strategies to evade detection by ML models. This dynamic nature challenges the effectiveness of static feature extraction methods in NLP, as phishing URLs can exhibit diverse linguistic patterns and structures that may not be easily captured by traditional text analysis approaches. Furthermore, the sheer volume of URLs and the need for real-time analysis pose scalability challenges for ML models, requiring efficient processing and decision-making capabilities. Balancing between model accuracy and false-positive rates is another hurdle, as overly aggressive models may lead to high false-positive rates, impacting user experience and trust in the system. Additionally, adversarial attacks targeting ML models used in URL phishing detection can undermine the reliability and robustness of the defense mechanisms, highlighting the ongoing arms race between attackers and defenders in the cybersecurity domain.

SYSTEM ARCHITECTURE

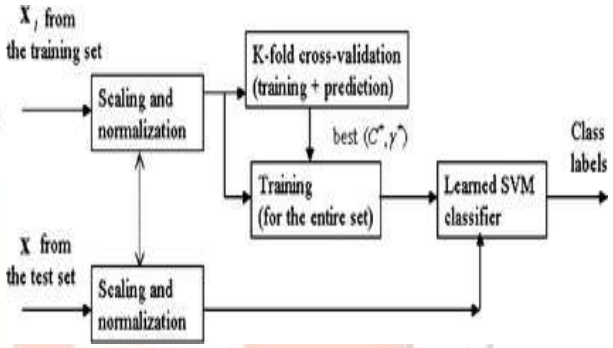


Fig - 3: System Architecture of Project

RESULTS AND DISCUSSION

Fig - 4: Home Page

Fig - 5: User Login Page

Fig – 6: Webpage Phishing Detection

Fig – 7: Model Performance Page

FUTURE ENHANCEMENT

The process of feature extraction involves the model using the most relevant training features and reducing the total number of features from the input. In this level, machine learning and deep learning are distinguished. For machine learning to extract features from the input data, a human is required. Nevertheless, in order to extract the most related features, DL relies on learning from the input and its label. In recent times, CNN has been the most widely used algorithm for feature extraction. The authors suggest extracting features using CNN. Convolutional neural networks (CNNs), in particular, are very good at autonomously learning hierarchical representations of data straight from the raw input. This is in contrast to other DL models. In tasks including signal processing, picture identification, and natural language processing, CNNs are especially useful for feature extraction. The main benefitCNNs may learn hierarchical features by performing many layers of convolutional and pooling processes, which is a useful capability for feature extraction. Both high-level abstract features (like object shapes or semantic interpretations) and low-level features (like edges and textures in photos) are captured by this hierarchical learning.

Using CNNs for feature extraction can be very helpful when it comes to URL phishing detection. CNNs have the capability to examine the composition and content of URLs, detecting trends and features suggestive of phishing endeavours. For instance, they can pick up on the ability to recognize shady domains, odd URL constructions, or misleading language in URLs.

CONCLUSION

Deep learning has emerged as a crucial tool for addressing cybersecurity issues like spearfishing. It is because it can automatically extract features from the input data rather than having to do so by hand. Modern deep learning models are examined in this survey to identify phishing attempts. It is crucial for examining every aspect of any DL model, from the model's output to its input data. Preprocessing data has equal value to that of the DL model. The model's performance is impacted by data preprocessing in all tasks, particularly when the model is used to identify real-time data through an application. For instance, even if the input data was not included in the model's dataset, the model must still be able to classify it. As a result, we focus greater attention onto data preprocessing and point out its advantages and disadvantages. Next, we evaluate the architecture of each DL model and point out its advantages and disadvantages.

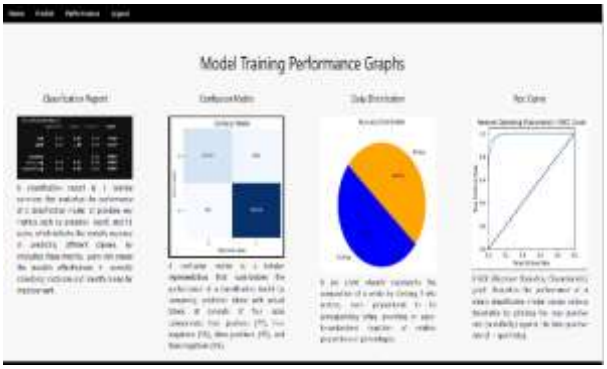
REFERENCES:

[1] Y. Zhang, Y. Xiao, K. Ghaboosi, J. Zhang, and H. Deng, ‘‘A survey of cyber-crimes,’’ Secur. Commun.Netw., vol. 5, no. 4, pp. 422–437, 2012.

[2] APWG Developers. (2021). Phishing Activity Trends Report.

[3] M. Lei, Y. Xiao, S. V. Vrbsky, and C.-C. Li, ‘‘Virtual password using random linear functions for on-line services, ATM machines, and pervasive computing,’’ Comput.Commun., vol. 31, no. 18, pp. 4367–4375, Dec. 2008.

[4] P. Burda, L. Allodi, and N. Zannone, ‘‘Don’t forget the human: A crowdsourced approach to automate response and containment against spear phishing attacks,’’ in Proc. IEEE Eur.



Symp. Secur. Privacy Workshops (EuroS PW), Sep. 2020, pp. 471–476.

[5] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, ‘‘PhishNet: Predictive blacklisting to detect phishing attacks,’’ in Proc. IEEE INFOCOM, Mar. 2010, pp. 1–5.

[6] W. Zhang, Y.-X. Ding, Y. Tang, and B. Zhao, ‘‘Malicious web page detection based on on-line learning algorithm,’’ in Proc. Int. Conf. Mach. Learn. Cybern., vol. 4, Jul. 2011, pp. 1914–1919.

[7] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. González, ‘‘Classifying phishing URLs using recurrent neural networks,’’ in Proc. APWG Symp. Electron.Crime Res. (eCrime), 2017, pp. 1–8.

[8] B. Cui, S. He, X. Yao, and P. Shi, ‘‘Malicious URL detection with feature extraction based on machine learning,’’ Int. J. High Perform. Comput.Netw., vol. 12, no. 2, pp. 166–178, 2018.

[9] Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang, ‘‘Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism,’’ IEEE Access, vol. 7, pp. 56329–56340, 2019. [10] J. Feng, L. Zou, O. Ye, and J. Han, ‘‘Web2Vec: Phishing webpage detection method based on multidimensional features driven by deep learning,’’ IEEE Access, vol. 8, pp. 221214–221224, 2020.

[11] H. Cheng, J. Liu, T. Xu, B. Ren, J. Mao, and W. Zhang, ‘‘Machine learning based low-rate DDoS attack detection for SDN enabled IoT networks,’’ Int. J. Sens. Netw., vol. 34, no. 1, pp. 56–69, 2020.

[12] S. Christin, É. Hervet, and N. Lecomte, ‘‘Applications for deep learning in ecology,’’ Methods Ecol. Evol., vol. 10, no. 10, pp. 1632–1644, Oct. 2019.

[13] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, ‘‘PhishAri: Automatic realtime phishing detection on Twitter,’’ in Proc. eCrime Res. Summit, Oct. 2012, pp. 1–12.

[14] H. Ma, Y. Zuo, and T. Li, ‘‘Vessel navigation behavior analysis and multiple-trajectory prediction model based on AIS data,’’ J. Adv. Transp., vol. 2022, pp. 1–10, Jan. 2022.

[15] J. Fang, B. Li, and M. Gao, ‘‘Collaborative filtering recommendation algorithm based on deep neural network fusion,’’ Int. J. Sens. Netw., vol. 34, no. 2, pp. 71–80, 2020.

[16] E. S. Gualberto, R. T. De Sousa, T. P. De Brito Vieira, J. P. C. L. Da Costa, and C. G. Duque, ‘‘The answer is in the text: Multi-stage methods for phishing detection based on feature engineering,’’ IEEE Access, vol. 8, pp. 223529–223547, 2020.