Global Scientific JOURNALS

GSJ: Volume 11, Issue 10, October 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

A Study of Machine Learning Approach for Malicious Emails

Nisar Ali[1]

[1] Department of CS & IT, University of Engineering & Technology, Peshawar, 25000, Pakistan

Email: 17pwbcs0579@uetpeshawar.edu.pk

Abstract. In the current landscape, phishing attacks pose a significant threat to a wide range of internet users, including governmental and business entities. This paper tackles the issue by providing a comprehensive overview of Machine Learning and exposing the tactics used by phishers in various phishing techniques. Our survey highlights the alarming effectiveness of phishing emails in specific sectors, prompting a comparative analysis. Recognizing the growing threat, there's a pressing need for advanced phishing detection technology. The study explores the application of Machine Learning Models and technical solutions to mitigate the pervasive problem of phishing, along with essential awareness insights for users to successfully detect and thwart phishing scams. Through meticulous investigation, the research delved into the realm of data analysis by employing eight distinct datasets, subjecting them to rigorous testing phases labeled as TF-IDF, CV, and TF-IDF with EPOCH. The comprehensive analysis didn't stop there; the research evaluates TF-IDF, CV, and both TF-IDF plus EPOCH individually on each dataset before amalgamating them. This fusion allowed for a holistic evaluation, shedding light on the superior algorithm for each dataset and elucidating the reasons behind their effectiveness.

Keywords: Machine Learning (ML), Artificial Intelligence (AI)

1 Introduction

Email communication, serving as an essential aspect of contemporary living, enables seamless interactions for both personal and professional purposes. However, the extensive use of email has turned it into a primary target for cybercriminals aiming to exploit its vulnerabilities. The risks posed by malicious emails, encompassing phishing attacks, malware distribution, and social engineering scams, extend to individuals, businesses, and government organizations. These cybercriminals deploy email to send unsolicited or advertising messages to a group of recipients, where unsolicited emails indicate the recipient has not granted permission [1]. The malevolent intent behind these emails is to deceive recipients, leading to financial losses, data breaches, and compromised security.

In the context of enhancing malicious email detection, a novel deep-learning framework is proposed, leveraging entire email content. The comprehensive evaluation of this framework demonstrates superior results with an AUC of 0.993, surpassing state-of-the-art methods, including human expert feature-based machine learning models, by a TPR of 5% [2]. In another reference [3], the authors present a method employing Natural Language Processing to detect phishing emails by extracting keywords from the message body. However, both methods face the challenge of feature loss during feature extraction, hindering the accurate detection of phishing emails by machine learning algorithms.

Research Problem Traditional rule-based email filtering systems have long been the standard approach to combat malicious emails. Although effective against known threats defined by pre-established rules and signature databases, these systems often struggle to identify new and sophisticated attacks that continuously evolve to elude detection. As cybercriminal tactics adapt, there is an urgent need to explore advanced and adaptive solutions for the effective detection of malicious emails.

This document provides a thorough examination of various machine learning algorithms employed in the categorization of phishing websites. In contrast to prior studies, our review encompasses a detailed scrutiny and comparison of diverse methodologies for identifying phishing websites. Notably, this study introduces a novel approach by utilizing three distinct datasets to train, test, and validate multiple classification algorithms, including DT [4], SVM [5], RF [6], NB [7], KNN [8], and ANN [9], to distinguish phishing websites from legitimate ones. Additionally, we leverage the widely-accepted Principal Component Analysis (PCA) [10] for dimensional reduction, achieving classification performance that is either equivalent to or surpasses that of using the complete feature set within the dataset. Furthermore, we explore the significance of all attributes/features in the eight datasets through PCA-based component loading. The remaining sections of the manuscript are structured as follows: Section 2 delves into existing work in this domain, Section 3 outlines the proposed methodology, Section 4 discusses and compares the results obtained, and Section 5 concludes the study while suggesting avenues for future research.

2 Literature Review

This literature review explores the substantial risk posed by cyber attackers employing email as a conduit to distribute harmful software to recipients' devices, causing disruptions for both individuals and organizations. The challenging task of detecting and categorizing such mali-

cious emails, encompassing spear phishing and zero-day attacks, is addressed through the introduction of a deep-learning solution. This solution utilizes data from email headers and bodies, incorporating dynamic analysis information as features. The system undergoes testing on four distinct language email datasets to emulate real-world scenarios, achieving satisfactory accuracy in the identification of both zero-day malicious emails and regular spam.

This chapter provides a comprehensive literature review centered on investigating the utilization of machine learning for the efficient detection of malicious emails. The objective of the review is to identify existing research, methodologies, and advancements in email security and machine learning techniques tailored specifically to combat malicious emails. Through an analysis of relevant literature, the chapter establishes a foundation for comprehending the current state-of-the-art while pinpointing gaps and opportunities for further research.

The identification of phishing websites holds a crucial role in the battle against online fraud. Recent strides in utilizing machine learning (ML) and data science methods across various domains, such as aerospace, border security, healthcare technologies, speech processing, object recognition, cybercrime detection, and smart cities, have been significant. However, the field of cybersecurity requires substantial improvement, particularly in light of the increasing instances of phishing attacks due to new techniques employed by malicious actors. To counter these attacks, several detection strategies have been developed.

In reference [11], the author presents a comprehensive analysis of various machine learning algorithms, evaluating their performance across multiple datasets. Statistical findings reveal the superiority of both artificial neural networks and random forest algorithms over other classification methods, achieving an impressive accuracy of over 97% using the identified features. Additionally, [12] proposes a hybrid approach that combines a genetic algorithm (GA) and SVM for detecting phishing emails on Android devices. This approach involves feature selection through the GA process and SVM for classification. The results of the study indicate that the proposed hybrid approach achieved a high accuracy rate, outperforming other state-of-the-art methods.

Extensive research has been conducted on spam detection, with early content-based spam identification relying on defined rules to identify spam messages. Subsequent research focused on utilizing traditional machine learning classifiers like Naïve Bayes, SVM, Random Forest, Decision trees, etc., which required handcrafted features extracted from training data. Feature selection techniques for classification tasks were also discussed by Agarwal [13]. In a review paper, Khorsi [14] summarized various statistics-based methodologies used to filter spam emails and found that no single technique was sufficient to combat spam due to inherent limitations. Additional significant work on traditional spam classifiers has been explored in [15].

In their work [16], Mohammad et al. employed a rule-based classification method to detect phishing websites, utilizing four distinct classification algorithms. The study emphasized the effectiveness of combining a feature reduction algorithm with a classification-based association for optimal performance. Noteworthy is the study's exclusive focus on high-frequency features, recognizing the potential for misleading conclusions, as higher frequency does not necessarily imply higher importance.

In this research endeavor, we have employed machine learning techniques to capture the inherent characteristics of email text and other features for their classification as either phishing or non-phishing based on carefully selected datasets.

3 Research Methodology

3.1 Datasets

In this section, we conduct a detailed examination of the dataset used in our study, which focuses on leveraging machine learning for effective malicious email detection. The chapter provides a thorough exploration of the data sources, shedding light on the dataset's size and the distribution of benign and malicious emails. To ensure the dataset's relevance and representativeness in emulating real-world email traffic, stringent criteria were applied for its selection. The datasets were sourced from diverse platforms like Kaggle, GitHub, and URLs, displaying variations in formats, including null values, duplicates, multiple columns, and unique labels. Employing a meticulous conversion process, all datasets underwent standardization into a unified format, thereby eliminating duplicates and null values. The final step involved merging the datasets and eliminating any duplicated values, resulting in a consolidated dataset with two columns labeled as 'text' and 'target.'

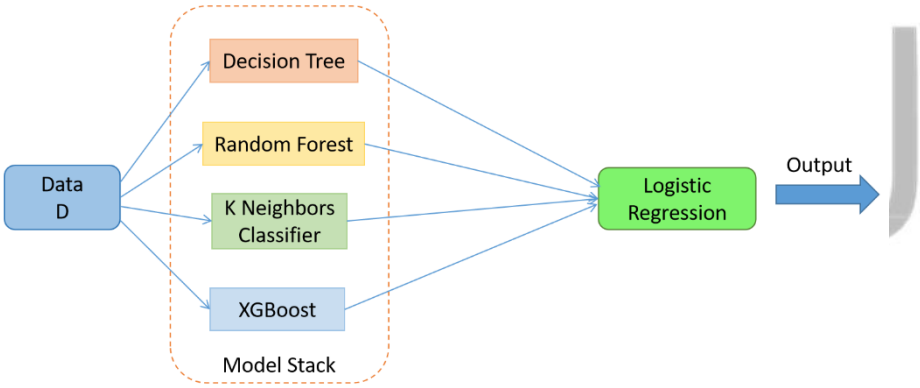

Figure 01: Stacking methodology

3.2 Model Selection

Initiating with the criteria for model selection, this section delineates the strategies employed to confront the task of proficiently identifying malicious emails. A codebase and dataset (5572, 5) were acquired from GitHub, setting the stage for a systematic process encompassing Data Cleaning, Exploratory Data Analysis (EDA), Text Pre-processing, Model Building, Evaluation, Improvement, Website Integration, and Deployment. The intricacies of the system's architecture and components are vividly depicted in Figure 6. This chapter serves as a prelude to an in-depth exploration of email classification and the efficient detection of malicious emails.

4 Conclusion and Future Work

Consistently, the Support Vector Classifier (SVC) emerged as the top-performing model, exhibiting superiority across diverse evaluation metrics and datasets. Notably, Logistic Regression (LR) demonstrated exceptional accuracy consistently, showcasing its versatility and robustness by employing an ensemble approach that leverages multiple decision trees effectively to handle diverse data patterns. The Random Forest (RF) model stood out for its remarkable ability to capture intricate relationships within the data. Examining precision, the Support Vector Classifier (SVC) proved to be the optimal model. SVC consistently demonstrated high precision scores across datasets, emphasizing its proficiency in making accurate positive predictions while minimizing false positives. The exceptional precision performance of SVC, especially in tasks where precision is critical (e.g., medical diagnoses or fraud detection), establishes it as a top choice.

Future work in the integration of Machine Learning and Neural Networks for cybersecurity against phishing attacks includes developing hybrid models, incorporating advanced natural language processing, and exploring real-time dynamic learning mechanisms. Evaluating the scalability and efficiency in large-scale environments, ensuring the interpretability of models, and extending research to multi-modal approaches are essential. Collaboration between academia, industry, and regulatory bodies is crucial for addressing ethical considerations and establishing standardized benchmarks.

In conclusion, the selection of the most suitable model necessitates a meticulous consideration of dataset characteristics, task-specific priorities, and constraints. While the Support Vector Machine (SVM) consistently demonstrates robust performance, factors such as model interpretability, computational efficiency, and real-world applicability should also be weighed in the decision-making process.

5 References

1. Kumar, N., & Sonowal, S. (2020, July). Email spam detection using machine learning algorithms. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 108-113). IEEE.
2. Qabajeh, I., Thabtah, F., & Chiclana, F. (2018). A recent review of conventional vs. automated cybersecurity anti-phishing techniques. Computer Science Review, 29, 44-55

3. Muralidharan, T., & Nissim, N. (2023). Improving malicious email detection through novel designated deep-learning architectures utilizing entire email. Neural Networks, 157, 257-279
4. Decision Trees — scikit-learn 0.22.1 documentation, Scikit-learn.org. [Online]. Available: https://scikit- learn.org/stable/modules/tree.html. [Accessed: 10- Jun- 2019]
5. Support Vector Machines — scikit-learn 0.22.1 documentation, Scikit-learn.org. [Online]. Available: https://scikit- learn.org/stable/modules/svm.html. [Accessed: 10- Jun- 2019]
6. Ensemble methods — scikit-learn 0.22.1 documentation, Scikit-learn.org. [Online]. Available: https://scikit-learn.org/stable/modules/ensemble.html
7. Naive Bayes — scikit-learn 0.22.1 documentation, Scikit-learn.org. [Online]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html.
8. Nearest Neighbors — scikit-learn 0.22.1 documentation, Scikit-learn.org. [Online]. Available: https://scikit- learn.org/stable/modules/neighbors.html
9. Nicholson, C.: 2019. "A Beginner's Guide to Neural Networks and Deep Learning", Pathmind. [Online]. Available: https://pathmind.com/wiki/neural-network. [Accessed: 14- Jun- 2019]
10. Jolliffe, I., Cadima, J.: "Principal component analysis: a review and recent developments", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, 2016. DOI: 10.1098/rsta.2015.0202
11. N. Ahmed, R. Amin, H. Aldabbas, D. Koundal, B. Alouffi, and T. Shah, "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges," Secur. Commun. Networks, vol. 2022, p. 1862888, 2022, doi: 10.1155/2022/1862888.
12. Mohammad, R., McCluskey, L., Thabtah, F.: "Intelligent rule-based phishing websites classification", IET Information Security, vol. 8, no. 3, pp. 153-160, 2014. DOI: 10.1049/ietifs.2013.0202
13. Karnik, R., Bhandari, D. G. M.: 2016. "Support vector machine-based malware and phishing website detection". Available: https://pdfs.semanticscholar.org/ffea/603ec9f33931c9de630ba1a6ac71924f1539.pdf?_ga=2.226066713.262761491.1579621617110277422 6.1578838444
14. Babagoli, M., Aghababa, M. P., Solouk, V.: 2018. "Heuristic nonlinear regression strategy for detecting phishing websites". DOI: https://doi.org/10.1007/s00500-018-3084-2
15. Sahingoz, O. K., Buber, E., Demir, O., Diri, B.: 2019. "Machine learning-based phishing detection from URLs". DOI: https://doi.org/10.1016/j.eswa.2018.09.029
16. Tahir, M.A.U.H., Asghar, S., Zafar, A., Gillani, S.: 2016. "A hybrid model to detect phishing sites using supervised learning algorithms". DOI: 10.1109/CSCI.2016.0214

17. Nicholson, C.: 2019. "A Beginner's Guide to Neural Networks and Deep Learning", Pathmind. [Online]. Available: https://pathmind.com/wiki/neural-network. [Accessed: 14- Jun- 2019]
18. A. Mughaid, S. AlZu'bi, A. Hnaif, S. Taamneh, A. Alnajjar, and E. A. Elsoud, "An intelligent cyber security phishing detection system using deep learning techniques," Cluster Comput., vol. 25, no. 6, pp. 3819–3828, 2022, doi: 10.1007/s10586-022-03604-4.

19. [6] H. F. Atlam and O. Oluwatimilehin, "Business Email Compromise Phishing Detection Based on Machine Learning: A Systematic Literature Review," Electronics, vol. 12, no. 1, p. 42, 2022.
20. T. Muralidharan and N. Nissim, "Improving malicious email detection through novel designated deep-learning architectures utilizing entire email," Neural Networks, vol. 157, pp. 257–279, 2023.
21. S. AlZu'bi, S. Al-Qatawneh, M. Alsmirat: Accelerating Statistical Segmentation Time with Transferable HMM Trained Matrices. In: Proceedings of the 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 172–176. IEEE (2018)
22. S. Al-Zubi, B. Hawashin, A. Mughaid, T. Baker: Efficient 3D Medical Image Segmentation Algorithm over a Secured Multimedia Network. Multimedia Tools and Applications, 80(11), 16887–16905 (2021)
23. S. AlZu'bi, Y. Jararweh: Tracing the Evolution of Data Fusion in Autonomous Vehicles Research: From Imaginary Idea to Smart Surrounding Community. In: Proceedings of the 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC), pp. 306–311. IEEE (2020)
24. A.A. AlKhatib, T. Sawalha, S. AlZu'bi: Overview of Load Balancing Techniques in Software-Defined Cloud Computing. In: Proceedings of the 2020 Seventh International Conference on Software Defined Systems (SDS), pp. 240–244. IEEE (2020)
25. I. Fette, N. Sadeh, A. Tomasic: Learning to Detect Phishing Emails. In: Proceedings of the 16th International Conference on World Wide Web, pp. 649–656 (2007)