Efficient Email phishing detection using Machine learning

Mustafa Al_Fayoumi
Princess Sumaya University for Technology
Amman, Jordan
m.alfayoumi@psut.edu.jo

Ammar Odeh
Computer Science Department,
Princess Sumaya University for Technology, Amman, Jordan
a.odeh@psut.edu.jo

Ismail Keshta
Computer Science and Information Systems Department
College of Applied Sciences,
AlMaarefa University
Riyadh, Saudi Arabia
imohamed@mcst.edu.sa

Abobakr Aboshgifa
The Libyan Higher Technical Center for Training and Production
Tripoli, Libya
Abobakr.Aboshgifa@tpc.ly

Tareq AlHajahjeh
Greater Amman Municipality Amman, Jordan
tariq_hjahjh@yahoo.co.uk

Rana Abdulraheem
Princess Sumaya University for Technology
Amman, Jordan
ranafaek22@gmail.com

***Abstract*—Emails are frequently utilized as a way of personal and professional communication. Banking information, credit reports, login data, and other sensitive personal information are commonly transmitted over email. This makes them valuable to cybercriminals, who can exploit the knowledge for their gain. Phishing is a technique used by con artists to steal sensitive information from people by impersonating well-known sources. The sender of a phished email can persuade you to disclose personal information under pretenses. The detection of a phished email is treated as a classification problem in this research, and this paper shows how machine learning methods are used to categorize emails as phished or not. SVM classifier attains a maximum accuracy of 0.998 percent in email classification.**

Keywords—SVM, accuracy, naïve Bayes, Random Forests

I. INTRODUCTION

Phishing is the most common type of cybercrime that involves persuading victims to submit sensitive information such as account numbers, passwords, and bank account numbers. Cyber-attacks are commonly launched using email, instant messages, and phone calls[1, 2].

Despite continual updates to the procedures for preventing such cyber-attacks, the result is insufficient. On the other hand, phishing emails have expanded tremendously in recent years, indicating the need for more effective and modern measures to combat them.[3, 4]

Several approaches for filtering phishing emails have been developed. However, the problem still requires a comprehensive solution. This is the first poll we're aware of that focuses on applying Machine Learning (ML) approaches to detect phishing emails[4]. This research examines the various state-of-the-art machine learning (ML) algorithms currently used to detect phishing emails at different stages of the attack[5]. A comparative assessment and analysis of these methodologies are performed. This provides an overview of the topic, its immediate solution space, and potential future research possibilities[6-8].

The rapid advancement of internet technologies has changed the way people interact online while also posing new security risks. Newly growing global dangers attack the user's computer and have the potential to steal their identity and money[9].

Phishing is a term with thousands of references in scientific papers, a lot of press coverage, and scrutiny from banks and law enforcement agencies. However, this raises the question of what phishing is[10].

In some publications, the phenomenon of phishing is expressly described; in others, it is presented with an illustration, while others assume that the reader already understands what phishing is. Many academics have offered their definitions of phishing, resulting in a wide range of interpretations in the scholarly literature. Because the phishing issue is broad and covers a multitude of circumstances, the literature does not provide a detailed description of phishing attacks[11, 12].

The term phishing was coined in 1996 as a result of social engineering attacks by web scammers against America Online (AOL) accounts, according to the APWG. Detecting phished email in the proposed system can be regarded as a classification problem with two types, ham and phished. Machine learning is a branch of artificial intelligence. When a system is given the ability to learn, it is intelligent. Without explicitly programmed, supervised learning is a concept that we use in our model. For classification, machine learning techniques are utilized.[13, 14]

II. RELATED WORK

For identifying legal and fraudulent web pages, is-based phishing detection systems use two lists: whitelists and blacklists. Phishing detection systems that use whitelists create secure and genuine websites that deliver relevant information. Every website that isn't on the whitelist is regarded as potentially dangerous.[5] built a system that creates a whitelist by logging the IP address of each site that the user has visited with a Login user interface. When a user accesses a website, the system will alert them if its registered information is incompatible.

The authors of [15] classified phishing websites using URL parameters such as length, number of unique characters, directory, domain name, and file name to identify them. The system uses support Vector Machines to classify websites that

978-1-6654-8303-2/22/$31.00 ©2022 IEEE

are not online. Adaptive Regularization of Weights, Confidence Weighted, and Online Perceptron are utilized for online classification. According to the trials ' findings, using the Adaptive Regularization of Weights algorithm improves accuracy while reducing system resource requirements.

Authors in [16]used a nonlinear regression technique to detect whether a website is phishing or not in a recent study. They train the system using harmony search and support vector machine meta-heuristic techniques. Harmony search, they claim, has a higher accuracy rate of 94.13 percent and 92.80 percent for train and test procedures, respectively, thanks to the use of around 11,000 web pages.

In [17] created a phishing detection system that uses adaptive self-structuring neural networks to classify the data. It has 17 features, some of which are reliant on third-party services. As a result, real-time execution takes substantially longer; yet, it can achieve higher accuracy rates. It only has 1400 items in its dataset, yet it has a reasonable acceptance rate for noisy data.

Yank in [18] provides an anti-phishing strategy that employs machine learning to identify phishing websites from legal ones by extracting 19 features from the client side. They used PhishTank (2018) and Openfish (2018) phishing pages and 1918 authentic web pages from Alexa popular websites, online payment gateways, and prominent banking websites. Their proposed approach achieved a 99.39 percent true positive rate using machine learning[4].

III. Proposed Work

The attackers add subdomains to the links to make them appear authentic. The number of dots in the link rose as subdomains were added. As suggested by In a valid email, the number dots should not be used. More than three [three]. This is a binary feature, meaning it determines whether or not a link exists. It would be in the mail if the number of dots was more prominent than three. This is a phished email.

The total number of links is: In general, phishing emails provide more information. In comparison to ham, the transmitter attempts to send many links. By tricking the user, you might direct him to an illicit website. This is a recurring feature.

The presence of JavaScript in an email indicates that the sender is either trying to conceal information or activate specific browser changes [18]. This is a one-of-a-kind feature. The presence of the script> tag in an email indicates that it has been phished.

Form tag: Phishing emails feature forms integrated into them to acquire information from users. This is a binary characteristic, meaning that the presence of a form tag indicates that the email is phished.

HTML emails allow the sender to include embedded graphics and URLs, which are not possible with plain text emails. If the email has an HTML tag, it is considered phishing. This is a one-of-a-kind feature.

The use of action words in emails shows if the sender expects the recipient to do a specific action, such as clicking on a link, filling out a form, or submitting detailed information. This is a recurring feature.

The word PayPal indicates that the sender is posing as a member of a recognized organization. The word "PayPal" appears in the mail's links or the "from" section, implying that the sender is affiliated with PayPal. This is a one-of-a-kind feature.

The presence of the term bank is a binary indicator indicating the message is about banking. Either the sender is posing as a member of the financial organization, or the reader is looking at the reader's credentials.

The word account appears in the email, indicating that it seeks emails tied to an account. It could be a social media account, a bank account, or something else entirely. It's a one-of-a-kind feature.

A. Support Vector Machine

SVM is a supervised technique often used for text categorization because of its speed and accuracy. It generates a hyper_plane, a two-dimensional line that best separates the categories, based on the training data. The decision boundary is the name given to this hyper_plane[19].

In phishing detection, the input is represented by a collection of features, such as the presence or absence of a specific term, and the output, which is 1 or -1, shows whether or not the email is phished.

B. The naive Bayes

The naive Bayes classifier[20] is a probabilistic technique that uses the Bayes theorem to classify sample data. The Bayes theorem asserts that, given a hypothesis H and evidence E, the relationship between the probability of the hypothesis P(H) before acquiring the evidence and the likelihood of the hypothesis P(H|E) after having the evidence is:

$$P(H|E) = \frac{P(E|H)}{P(E)} P(H)$$

Each category's probability is calculated, and the one with the highest possibility is chosen.

C. Random forests

Random forests, also known as random decision forests, are an ensemble learning method for classification, regression, and other tasks that works by training a large number of decision trees and then outputting the class that is the mode of the categories (classification) or the mean prediction (regression) of the individual trees. Random decision forests address the problem of decision trees over-fitting to their training set.[21]

IV. Result and Analysis

The confusion matrix is used to assess the performance of the three intelligent classification algorithms. The classifier's performance on the input dataset is represented in this matrix. This matrix can derive various performance metrics, such as accuracy and F-measure

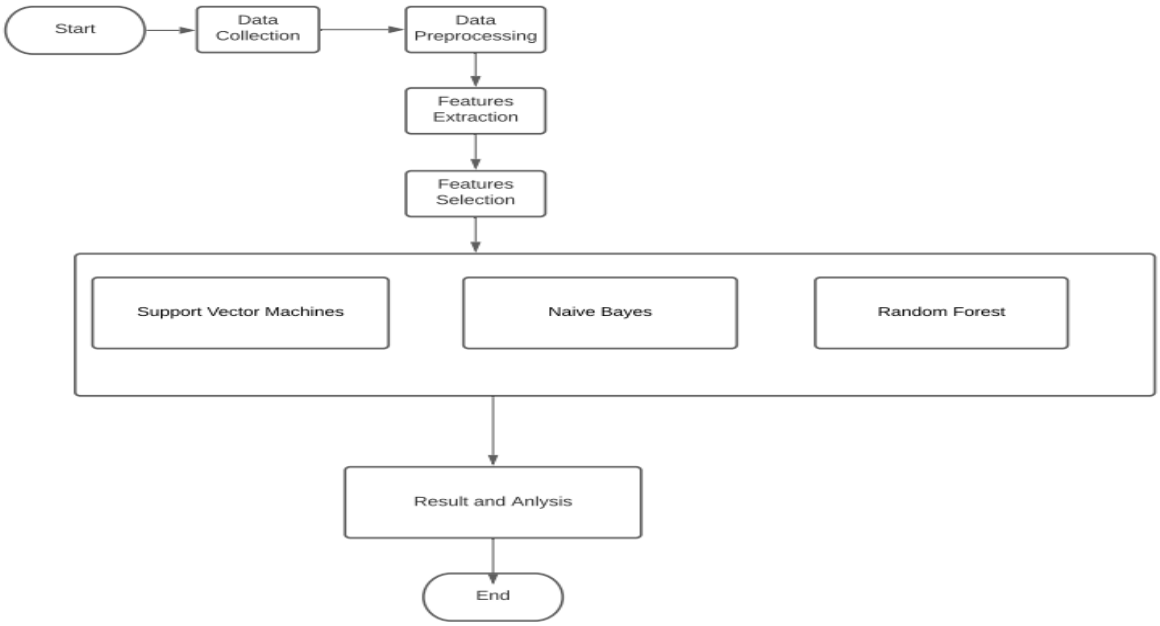

Figure 1. Block diagram of the proposed system.

.

Table 1. The accuracy and F measure for three classifiers in different testing ratio

Classifiers	Testing Ratio	Accuracy	F-measure
SVM	50:50	0.874453	0.8744535
	60:40	0.980363	0.9803729
	70:30	0.998002	0.998002
Naive Bayes	50:50	0.809524	0.8095238
	60:40	0.75	0.75
	70:30	0.797052	0.755814
Random Forests	50:50	0.90341	0.8975013
	60:40	0.855178	0.847619
	70:30	0.824666	0.8170676

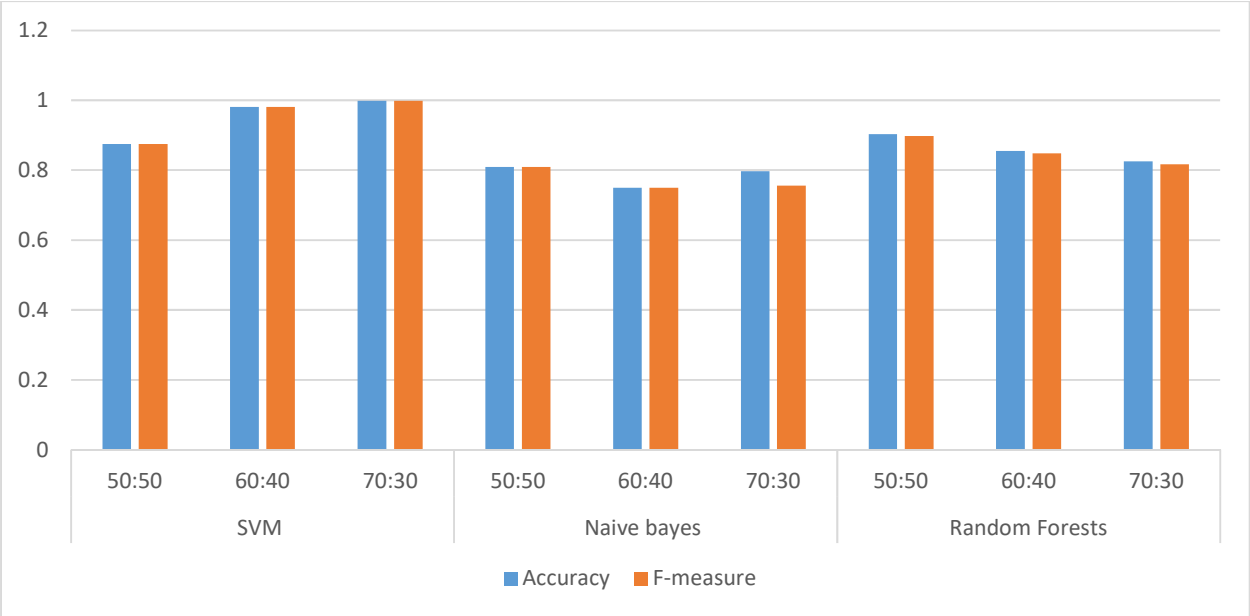

Figure 2. The histogram comparison between three classifier

It's vital to keep in mind that as the percentage of training data grows, so does the computational complexity and training time. As a result, 30% of the dataset is used for training to balance the percentage of correctly categorized data and the percentage of incorrectly classified data.

V. CONCLUSION

This research proposes an intelligent approach for detecting phishing emails effectively. It examines the differences between Naive Bayes, Random Forests, and SVM. The goal is to find the most effective intelligent classification model for detecting email phishing. Different experiments were conducted on three benchmarking testing levels to evaluate the performance of the three classifiers.

We plan to test SVM's performance on different benchmarking datasets in the future. Performance comparison of SVM with various kernels, such as Gaussian or sigmoid kernels, will also be carried out.

VI. ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by Princess Sumaya University for Technology (PSUT) and AlMaarefa University while conducting this research work

VII. REFERENCES

[1] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," Computers & Security, vol. 68, pp. 160-196, 2017.
[2] I. Vayansky and S. Kumar, "Phishing–challenges and solutions," Computer Fraud & Security, vol. 2018, pp. 15-20, 2018.
[3] E. J. Williams, et al., "Exploring susceptibility to phishing in the workplace," International Journal of Human-Computer Studies, vol. 120, pp. 1-13, 2018.
[4] A. Odeh, et al., "Machine LearningTechniquesfor Detection of Website Phishing: A Review for Promises and Challenges," in 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 0813-0818.
[5] A. Odeh, et al., "Efficient Detection of Phishing Websites Using Multilayer Perceptron," 2020.
[6] A. Odeh, et al., "PHIBOOST-a novel phishing detection model using Adaptive boosting approach," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 7, 2021.
[7] K. L. Chiew, et al., "A survey of phishing attacks: Their types, vectors and technical approaches," Expert Systems with Applications, vol. 106, pp. 1-20, 2018.
[8] M. Al-Fayoumi, et al., "Intelligent association classification technique for phishing website detection," International Arab Journal of Information Technology, vol. 17, pp. 488-496, 2020.
[9] Y. Kwak, et al., "Why do users not report spear phishing emails?," Telematics and Informatics, vol. 48, p. 101343, 2020.
[10] A. Odeh, et al., "PHISHING WEBSITE DETECTION USING MULTILAYER PERCEPTRON."
[11] G. Sonowal and K. Kuppusamy, "PhiDMA–A phishing detection model with multi-filter approach," Journal of King Saud University-Computer and Information Sciences, vol. 32, pp. 99-112, 2020.
[12] A. ODEH, et al., "Efficient Prediction Of Phishing Websites Using Multilayer Perceptron (Mlp)," Journal of Theoretical and Applied Information Technology, vol. 98, 2020.

[13] I. Keshta and A. Odeh, "Security and privacy of electronic health records: Concerns and challenges," Egyptian Informatics Journal, vol. 22, pp. 177-183, 2021.
[14] R. Faek, et al., "Exposing Bot Attacks Using Machine Learning and Flow Level Analysis," in International Conference on Data Science, E-learning and Information Systems 2021, 2021, pp. 99-106.
[15] O. K. Sahingoz, et al., "Machine learning based phishing detection from URLs," Expert Systems with Applications, vol. 117, pp. 345-357, 2019.
[16] A. Oest, et al., "Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis," in 2018 APWG Symposium on Electronic Crime Research (eCrime), 2018, pp. 1-12.
[17] A. Abbasi, et al., "The phishing funnel model: A design artifact to predict user susceptibility to phishing websites," Information Systems Research, vol. 32, pp. 410-436, 2021.
[18] P. Yang, et al., "Phishing website detection based on multidimensional features driven by deep learning," IEEE Access, vol. 7, pp. 15196-15209, 2019.
[19] D. A. Pisner and D. M. Schnyer, "Support vector machine," in Machine Learning, ed: Elsevier, 2020, pp. 101-121.
[20] D. Berrar, "Bayes' theorem and naive Bayes classifier," Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands, pp. 403-412, 2018.
[21] S. Athey, et al., "Generalized random forests," The Annals of Statistics, vol. 47, pp. 1148-1178, 2019.