Detection and Analyzing Phishing Emails Using NLP Techniques

Rian Sh. Al-Yozbaky
Computer Sciences Department
College of Computer Sciences & Mathematics
Mosul University, Mosul-Iraq
rian.21csp85@student.uomosul.edu.iq

Mafaz Alanezi
Computer Sciences Department
College of Computer Sciences & Mathematics
Mosul University, Mosul-Iraq
mafazmhalanezi@uomosul.edu.iq

2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) | 979-8-3503-3752-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/HORA58378.2023.10156738

***Abstract*— The most common detrimental technique used by attackers to deceive victims into disclosing personal information is phishing, in which they pose as trustworthy individuals or organizations often via email. Although fake email attacks are a common tactic used by cybercriminals, their use has recently increased as attacker's profit from victims' anxiety. As a result, further study is required to determine how to recognize bogus emails. This paper proposed a new model to extract the Arabic email content and compare it using three determinants based on neural language programming (NLP) for the purpose of discovering whether it is a legitimate email or a phishing email. The first is a black list of Arabic common phishing words, the roots of a black list of Arabic common phishing words, and a list of Arabic common phishing sentences, the best two results for applying the above conditions were (99% Legal and 96% Phishing) when using the three conditions together and (99% Legal and 94% Phishing) when using a blacklist of common words of phishing, and then will present and discuss the results obtained.**

Keywords*— *Phishing email Dataset, Natural Language Processing, Phishing Email Detection, NLP Features, NLTK

I. INTRODUCTION

Phishing is a risk associated with social engineering that preys on the gullibility of uninformed internet users to fool them into disclosing personal information. Attackers and phishers adopt the guise of genuine internet users. Phishers attempt to access a victim's accounts without authorization in order to steal sensitive data, the victim's identity, and other personal information. Therefore, it is necessary to stop hacking and the related illegal activity. The Anti-scam Working Group (APWG) estimates that by the third quarter of 2020, there were 128,926 scam emails, up from 44,008 in the first quarter of 2020[1]. Phishing emails are one form of phishing where the phisher sends the recipient an email using a false email address to trick them into reading the email[2]–[5]. This enables the phisher to sway the user and profit from their personal data[6]. In order to combat the issue, a number of anti-phishing solutions have gained popularity, including phishing blacklists[7], utilizing NLP and ML for phishing email detection [8], [9]–[13], and more.

In [14]. suggested a method to spot phishing assaults that exposed a client's credentials to risky substances and resulted in a security breach. The advised course of action is to use an online search tool to compare the web address of the questioned website with the locations returned by our search query. Here, the catchphrases and domain names, which also include the title, body content, and meta description data, are relevant to their search [15].

At first, they observe how the typical TF-IDF ("Term Frequency-Inverse Document Frequency") determines whether a website is trustworthy or fake. Next, apply a weighted heuristic that was thoroughly suggested in their study by varying the weights of the label information. After that, alter the TF-IDF result similarly to enhance how their phishing indication is displayed [15].

The subject line of phishing emails, as well as the objects that the messages' content refers to, are crucial for recognizing them [16]. Utilizing the semantic web, messages' words are transformed into occasions. These things are known as certain occasions, and a pair of occasions is known as an occasion pair, which discusses the connection between these two events. Another calculation for phishing message discrimination that is dependent on occasion matching is suggested in their research. Building the knowledge base for the semantic web—which offers the connections between words and occasions—is the first stage in this computation. The process of organizing phishing communications after creating the classification information base follows. The last section explains how to recognize phishing by utilizing the semantic web knowledge base and class data set.

URI and CSS coordinating-based phishing detection tools (CUMP) were developed by Mishra A. and Gupta B. B. to identify zero-day phishing attacks [17]. The URI "uniform resource identifier" and CSS "cascading style sheet" concepts are the foundations of their structure. This theory is used in the hope that even seasoned users won't be able to recognize phishing sites by perception. The phisher attempts to imitate the visible plan and URI design repeatedly. Phishers often duplicate the aesthetic appearance using the same CSS style. It is challenging to implement an analogous concept without utilizing the same CSS. Their architecture utilized the fundamental components of any phishing attack for URI and CSS coordination in order to defend against phishing site risks, particularly "zero-day" attacks.

In this paper proposed a new model to extract the Arabic email content. Furthermore, ccompare it uses three determinants based on NLP methods. Then text analysis to identify phishing emails-based content in three different scenarios (a blacklist of popular terms, the roots of a popular words blacklist, and a blacklist of popular sentences). When combined the three conditions, the results were reached to 99% for identifying legal emails as legal and 96% for identifying phishing emails as phishing.

II. PHISHING EMAIL DETECTION USING NLP

NLP is a branch of computer science that focuses on making it possible for computers to understand human discourse. The goal is to create a computer that is human-like in that it can understandably analyze data and commands given to it in natural language[18].

The use of NLP technology, rather than DL methods, ignores the differences between anti-phishing email and other goals and partly ignores environmental information, which is preventing the discovery of phishing emails from progressing. The current literature analysis demonstrates the significance of the problems in understanding the hacking detection study tendency using NLP and ML. To thoroughly evaluate these research works. To summarize the findings on identifying hacking using NLP and ML, primarily performed a poll. Diverse ML techniques have been studied in order to recognize fake emails. For the purpose of classifying emails as malevolent or secure emails, many characteristics have been created [19]–[24].

Spoofing emails are one form of phishing where the phisher sends the recipient an email using a false email address to trick them into reading the email [8], [3]–[5], [25]. This enables the phisher to sway the user and profit from their personal data [6]. In order to combat the issue, a number of anti-phishing technologies have acquired popularity, including phishing blacklists [8] and NLP and ML-based detection of phishing emails [3], [8], [11], [13].

III. PHISHING EMAILS DETECTING FEATURES

The purpose of phishing emails is to get the recipients' private information. Most users were phished because they were negligent in their internet browsing. Companies should train their employees on the tricks and tactics used by phishers. This part will cover both how to recognize fraud as well as how to protect against phishing assaults. Some spam blockers utilize hundreds of characteristics to sort out fake emails. These characteristics [8] for phishing email detection are categorized as follows:

- Characteristics based on the email body: These characteristics are derived from the email content. Binary elements like forms, HTML, and particular words and URLs can be found in the email text.
- Features that can be inferred from the topic of an email include references to earlier emails and the use of words similar "verify" or "debit."
- URL-based traits: These characteristics examine things like whether a domain name is used in place of an IP address. How many photos are linked both internally and externally, whether @ is present in links, and so forth, how many cycles there are in links, and other things.
- Script-based features: These functions examine emails for script-based components such as JavaScript, code for pop-up windows, on-click operations, and others.
- Author-based traits: These characteristics provide insight into the author., like the discrepancy between the sender's address and the recipient's response to the address[8].

Sheng et al. [7] conducted an investigation into the effectiveness of the hacking database. The approach is built on originator and link blacklists. The originating address and the link address from the email must be extracted and cross-referenced with the blacklists in order to detect whether the email comprises a phishing attempt. Sender blacklists and link blacklists are both types of blacklists. The continuous updating of the ban and user reports that identify a website as a scam website are the main drawbacks. Two of the most popular fraud websites are PhishTank [26] and OpenPhish [5] among the lists of scam websites. Blacklists play a major role in the effectiveness of blacklisting for scam email identification.

IV. PHISHING LIFE CYCLE

A false attack begins when an email is sent to an internet client as shown in Figure 1. The malicious link in this email directs users to a fake website that the sender has cloned from the legitimate website on which it is based. This persuades the uninformed email recipient of the email's and website's legitimacy.

Fig. 1. a new example of a phishing email [27]

Figure 1. shows the fundamental elements of a scam email. The University of Massachusetts Amherst provided this data [27] and is meant to show how to protect internet users from fraud.

Figure 2. shows the characteristics of the phony website, which asks the email receiver for private information that the perpetrator then unlawfully acquires and uses [1], these characteristics are:

- Although the sender claims to be "UMass Amherst it@umass.edu>," UMass is not affiliated with the sender, and the email address used is not one of the institution's email addresses.
- Spelling and grammatical errors in phishing emails are obvious. For instance, the comma before the colon in this email is improper.
- The language used in a phishing email will also create a fictitious sense of urgency. This causes the recipient to respond without much thought.
- One illustration from this email is the danger of a "permanent" problem if the recipient waits too long to respond.
- The message URL has been designed to resemble a legitimate UMass Amherst page address. But when you mouse over it, you see that it takes you to a different page.

- When a link is hovered over, it reveals where it leads, and in this case, it is not to a reputable UMass website. Check the links twice before clicking them as a result!
- The letter's claim to be from both Microsoft Corporation and UMass Amherst is another problem. If the sender is dubious of their own identity or allegiance, they are a fraud.

The message's URL points to a phony SPIRE login page with the website address "tantechholdings.com."[28].

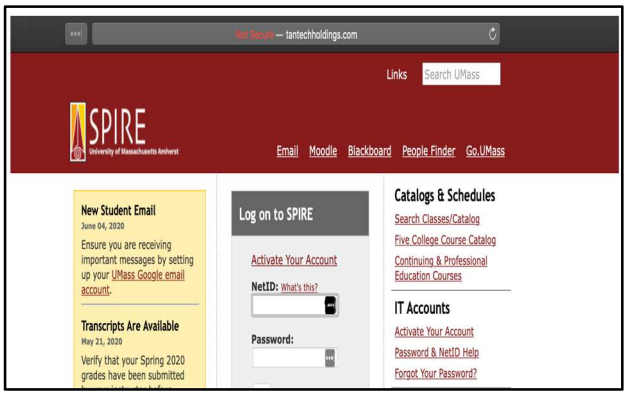

Fig. 2. Phishing website with annotations[27]

V. METHOD AND TOOLS

It demonstrates the proposed model's working mechanism in Figure 3. where the process begins by emailing a specific account with all of the electronic messages acquired in the planned database. After that, analysis checks to see if the messages are in Arabic or another language, and if they are, it examines the substance of the messages. After being read, messages are inspected and evaluated using three separate methods, with an emphasis on the message's content rather than its address or sender.

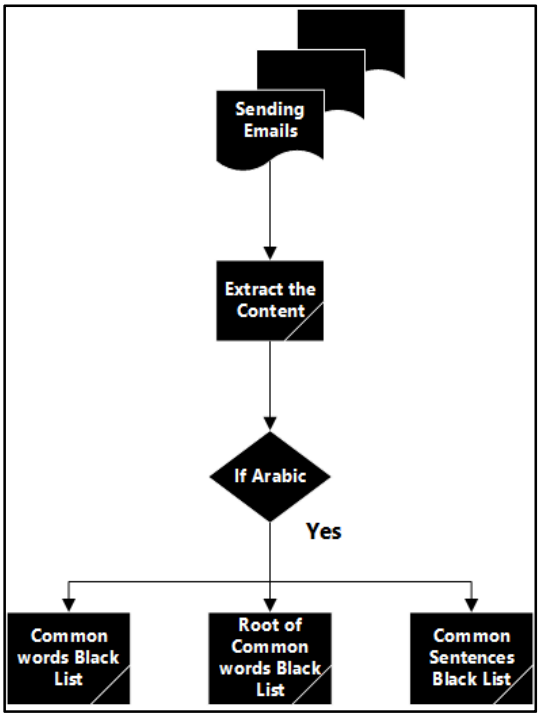

Fig. 3. Analyzing the email contents

Based on the matches for these phrases, calculate a value as follows:

- Examine each phrase's content separately and contrast it with a list of terms that are frequently used in phishing emails.
- Obtaining the root of each word in the email's content and comparing it to each word's root in the list of words from the first technique.
- Examine the phrases in the email text and compare them to a database of the terms that are most frequently used in phishing emails.

The results of the tests are thus compared to Figure 4. which displays the message's content and indicates whether it is a legitimate email or phishing email.

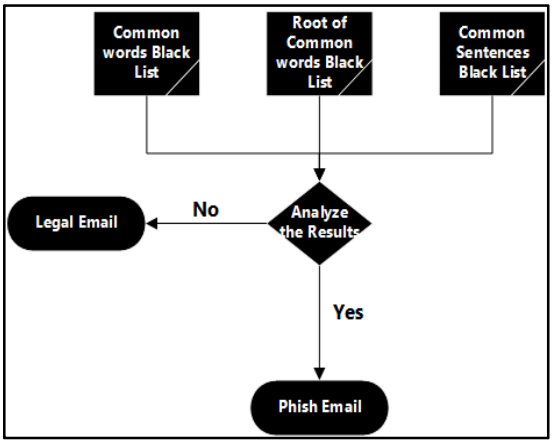

Fig. 4. Examining the email content

The tools and apps utilized were:

- To keep track of every phishing emails sent, a phishing system was developed using Kali Linux and the GoPhish tool.
- The Python programming language and libraries (NLTK, RE, Email, Pandas, and Imaplib), were used to compile and analyze the results and take the necessary action after determining that the messages were phishing emails.

VI. RESULT AND DISCUSSION:

In this section, the results of using the above three different ways will be presented with deep explained as well as classified into:

6.1 Using a list of common words for phishing emails

Figure 5. shows the results of comparing the list of common words for phishing e-mails with the content of the dataset emails.

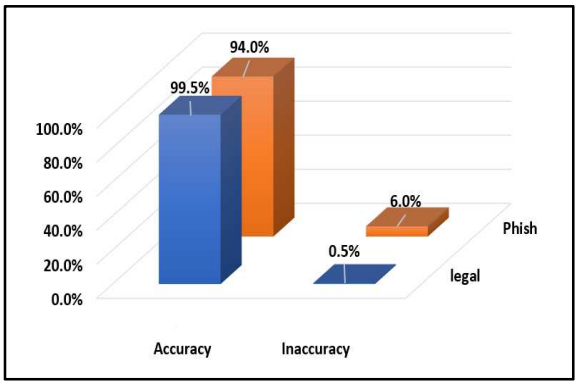

Fig. 5. A list of common words for phishing emails

In terms of detecting phishing emails, the results were accurate in detecting legitimate emails (99.5%) and phishing emails (94.0%).

6.2 Using Root of Words in the blacklist

The comparison of the prevalent terms used in phishing emails and the emails in the sample results in Figure 6.

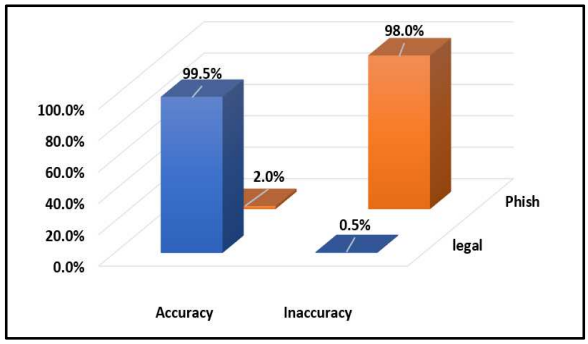

Fig. 6. Root of Words in the blacklist

99.5% of legitimate emails were correctly identified by the results, while just 2.0% of phishing emails were correctly identified.

6.3 Using a list of common sentences for phishing emails

The comparison between the common sentences for phishing emails and the emails in the dataset is shown in Figure 7.

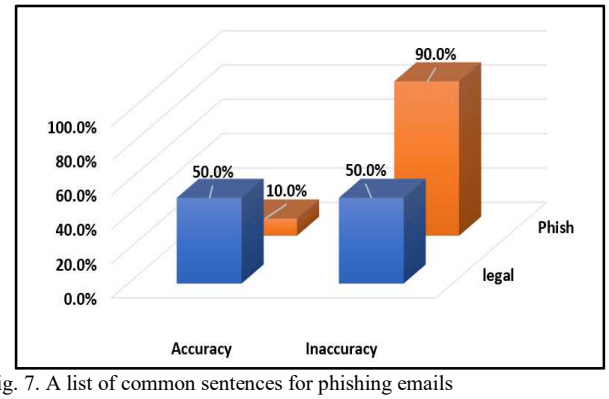

Fig. 7. A list of common sentences for phishing emails

The results' accuracy in identifying legitimate emails was 50.0%, whereas the results' accuracy in identifying phishing emails was 10.0%.

6.4 Using two conditions (list of words and the root)

The comparison between the content of the dataset emails and the list of phrases that are frequently used in phishing emails and their roots is shown in Figure 8.

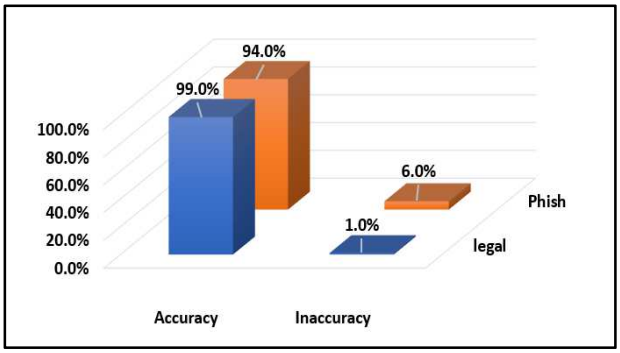

Fig. 8. Two conditions (list of words and the root)

The results were accurate in detecting phishing emails with a score of 94.0% and legitimate emails with a score of 99.0%.

6.5 Using two conditions (list of words and the sentences)

Figure 9. shows the results of comparing the list of common words for phishing e-mails and the common sentences for phishing e-mails with the content of the dataset emails.

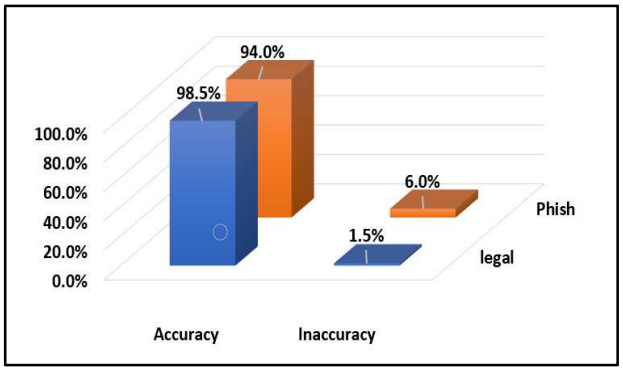

Fig. 9. Two conditions (list of words and the sentences)

The accuracy of the results in identifying legitimate emails was 98.5 percent, and the accuracy in identifying phishing emails was 94.0 percent.

6.6 Using two conditions (list of roots and the sentences)

The comparison of the root of the list of common terms and sentences for phishing emails with the content of the dataset emails yielded the findings shown in Figure 10.

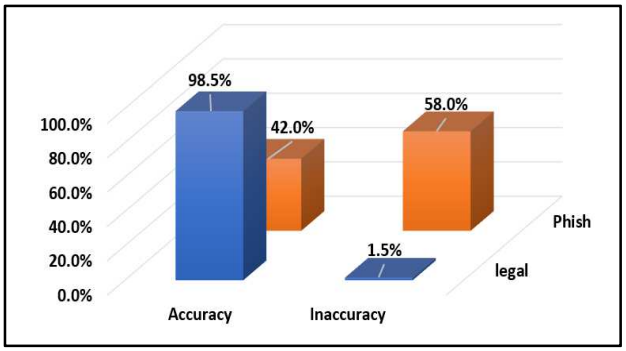

Fig. 10. Two conditions (list of roots and the sentences)

The accuracy of the results for recognizing legal emails was 98.5 percent, whereas the accuracy for phishing emails was 42.0 percent.

6.7 Using Three conditions (list of words, the roots, and sentences)

The comparison of the dataset emails' content with the list of common words, their roots, and common sentences for phishing emails yielded the results shown in Figure 11.

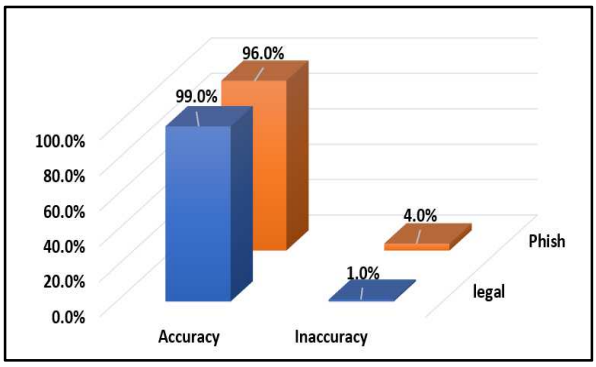

Fig. 11. Three conditions (list of words, the roots and sentences)

The findings for detecting legitimate emails were accurate (99.0%), and the results for detecting phishing emails were accurate (96.0%).

VII. CONCLUSION:

Due to the great development in relying on electronic economic transactions in various fields of life in the world, especially in the Arab countries during the past three years, the electronic crimes associated with electronic phishing have grown frighteningly, due to the scarcity of research presented in the Arabic language. find there is a large gap in finding solutions for treatment and elimination. On phishing emails in the Arabic language, especially by methods that rely on (NLP) in analyzing the content of messages, and there is also a great scarcity in the availability of a database of phishing emails in the Arabic language for the purpose of conducting research and studies. Researchers may build on and explore additional research and studies utilizing the findings of this study to locate content-based phishing emails for messengers in Arabic and other languages using the same test and conclusion approach. The best first results for utilizing the suggested techniques to recognize phishing emails when the three approaches were combined were (96%) Compared to (94%) when utilizing one method, illustrating the effectiveness of combining more than one method to create successful results.

ACKNOWLEDGMENT

The authors would like to thank the College of Computer Science and Mathematics, University of Mosul for supporting this work.

REFERENCES

[1] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques," IEEE Access, vol. 10, pp. 65703–65727, 2022, doi: 10.1109/access.2022.3183083.

[2] S. Gupta, A. Singhal, and A. Kapoor, "A literature survey on social engineering attacks: Phishing attack," Proceeding - IEEE Int. Conf. Comput. Commun. Autom. ICCCA 2016, pp. 537–540, 2017, doi: 10.1109/CCAA.2016.7813778.

[3] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues," IEEE Access, vol. 5, pp. 9044–9064, 2017, doi: 10.1109/ACCESS.2017.2702187.

[4] E. S. Gualberto, R. T. De Sousa, T. P. B. De Vieira, J. P. C. L. Da Costa, and C. G. Duque, "From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection," IEEE Access, vol. 8, pp. 76368–76385, 2020, doi: 10.1109/ACCESS.2020.2989126.

[5] G. Sonowal and K. S. Kuppusamy, "PhiDMA – A phishing detection model with multi-filter approach," J. King Saud Univ. - Comput. Inf. Sci., vol. 32, no. 1, pp. 99–112, 2020, doi: 10.1016/j.jksuci.2017.07.005.

[6] A. Zamir et al., "Phishing web site detection using diverse machine learning algorithms," Electron. Libr., vol. 38, no. 1, pp. 65–80, 2020, doi: 10.1108/EL-05-2019-0118.

[7] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," 6th Conf. Email Anti-Spam, CEAS 2009, 2009.

[8] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey," Procedia CIRP, vol. 189, no. 2019, pp. 19–28, 2021, doi: 10.1016/j.procs.2021.05.077.

[9] R. Verma and N. Hossain, "Semantic feature selection for text with application to phishing email detection," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8565, pp. 455–468, 2014, doi: 10.1007/978-3-319-12160-4_27.

[10] G. Park and J. M. Taylor, "Using Syntactic Features for Phishing Detection," 2015, [Online]. Available: http://arxiv.org/abs/1506.00037

[11] A. Vazhayil, N. B. Harikrishnan, R. Vinayakumar, and K. P. Soman, "PED-ML: Phishing email detection using classical machine learning techniques CENSec@Amrita," CEUR Workshop Proc., vol. 2124, pp. 69–76, 2018.

[12] I. R. A. Hamid and J. Abawajy, "Hybrid feature selection for phishing email detection," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7017 LNCS, no. PART 2, pp. 266–275, 2011, doi: 10.1007/978-3-642-24669-2_26.

[13] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J. P. Niyigena, "An effective phishing detection model based on character level convolutional neural network from URL," Electron., vol. 9, no. 9, pp. 1–24, 2020, doi: 10.3390/electronics9091514.

[14] A. K. Jain, S. Parashar, P. Katare, and I. Sharma, "PhishSKaPe: A Content based Approach to Escape Phishing Attacks," Procedia Comput. Sci., vol. 171, no. 2019, pp. 1102–1109, 2020, doi: 10.1016/j.procs.2020.04.118.

[15] M. Alanezi, "Phishing Detection Methods: A Review," Tech. Rom. J. Appl. Sci. Technol., vol. 3, no. 9, pp. 19–35, 2021, doi: 10.47577/technium.v3i9.4973.

[16] H. Che, Q. Liu, L. Zou, H. Yang, D. Zhou, and F. Yu, "A Content-Based Phishing Email Detection Method," Proc. - 2017 IEEE Int. Conf. Softw. Qual. Reliab. Secur. Companion, QRS-C 2017, pp. 415–422, 2017, doi: 10.1109/QRS-C.2017.75.

[17] A. Mishra and B. B. Gupta, "Intelligent phishing detection system using similarity matching algorithms," Int. J. Inf. Commun. Technol.,

vol. 12, no. 1–2, pp. 51–73, 2018, doi: 10.1504/IJICT.2018.089022.

[18] S. Aggarwal, V. Kumar, and S. D. Sudarsan, "Identification and detection of phishing emails using natural language processing techniques," ACM Int. Conf. Proceeding Ser., vol. 2014-Septe, no. January, pp. 217–222, 2014, doi: 10.1145/2659651.2659691.

[19] G. Sonowal, "Phishing Email Detection Based on Binary Search Feature Selection," SN Comput. Sci., vol. 1, no. 4, pp. 1–14, 2020, doi: 10.1007/s42979-020-00194-z.

[20] L. F. Gutiérrez, F. Abri, M. Armstrong, A. S. Namin, and K. S. Jones, "Phishing Detection through Email Embeddings," pp. 1–9, 2020, [Online]. Available: http://arxiv.org/abs/2012.14488

[21] A. Kumar, J. M. Chatterjee, and V. G. Díaz, "A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing," Int. J. Electr. Comput. Eng., vol. 10, no. 1, pp. 486–493, 2020, doi: 10.11591/ijece.v10i1.pp486-493.

[22] 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, doi: 10.1109/ACCESS.2020.3043396.

[23] A. Ora, "Spam Detection in Short Message Service Using Natural Language Processing and Machine Learning Techniques," Natl. Coll. Ireland, Pp.1-27., 2020.

[24] S. R. Mirhoseini, F. Vahedi, and J. A. Nasiri, "E-Mail phishing detection using natural language processing and machine learning techniques," Academia, vol. 23, no. 7, pp. 1–9, 2020, [Online]. Available: https://scholar.smu.edu/datasciencereview/vol6/iss2/14/

[25] P. N. Astya, Galgotias University. School of Computing Science and Engineering, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section. SP/C Joint Chapter, and Institute of Electrical and Electronics Engineers, "Proceeding, International Conference on Computing, Communication and Automation (ICCCA 2016) : 29-30 April, 2016," pp. 1–4, 2016.

[26] 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, doi: 10.1109/ACCESS.2019.2913705.

[27] "University of Massachusetts Amherst," p. https://www.umass.edu/it/freshphish, 2020.

[28] N. A. Unnithan, N. B. Harikrishnan, R. Vinayakumar, K. P. Soman, and S. Sundarakrishna, "Detecting phishing E-mail using machine learning techniques CEN-SecureNLP," CEUR Workshop Proc., vol. 2124, pp. 50–56, 2018.