**TELKOMNIKA Telecommunication, Computing, Electronics and Control**
Vol. 19, No. 5, October 2021, pp. 1707~1714
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v19i5.15981

# A maximum entropy classification scheme for phishing detection using parsimonious features

**Emmanuel O. Asani[1], Adebayo Omotosho[2], Paul A. Danquah[3], Joyce A. Ayoola[4], Peace O. Ayegba[5], Olumide B. Longe[6]**
[1,4,5]Department of Computer Science, Landmark University, Omu-Aran, Nigeria
[1]SDG 11: Sustainable Cities and Communities Research group, Landmark University, Omu-Aran, Nigeria
[2]Internet Technologies and Internet Systems Research Group, Hasso Plattner Institute, Potsdam, Germany
[3]Council for Scientific and Industrial Research-Institute for Scientific and Technological Information, Accra, Ghana
[6]School of Computational Sciences and Informatics, Academic City University College, Accra, Ghana

**Article Info**

***Article history:***

Received Mar 03, 2020
Revised Jun 29, 2021
Accepted Jul 13, 2021

***Keywords:***

Classification
Parsimonous features
Phishing
Social engineering

## ABSTRACT

Over the years, electronic mail (e-mail) has been the target of several malicious attacks. Phishing is one of the most recognizable forms of manipulation aimed at e-mail users and usually, employs social engineering to trick innocent users into supplying sensitive information into an imposter website. Attacks from phishing emails can result in the exposure of confidential information, financial loss, data misuse, and others. This paper presents the implementation of a maximum entropy (ME) classification method for an efficient approach to the identification of phishing emails. Our result showed that maximum entropy with parsimonious feature space gives a better classification precision than both the Naïve Bayes and support vector machine (SVM).

*This is an open access article under the CC BY-SA license.*

***Corresponding Author:***

Emmanuel O. Asani
Department of Computer Science
Landmark University
Omu-Aran, Nigeria
Email: asani.emmanuel@lmu.edu.ng

## 1. INTRODUCTION

Owing perhaps, to its ubiquity and limitless potentials for communications and interconnectivity, the internet has become a home for divergent tendencies that engender both positive and negative practices; while internet technology is at the cutting edge of great innovations and revolutionary findings, criminals are equally able to deploy this technology for easy propagation and perpetration of their criminal agenda with universal reach [1]. The e-mail, being the predominant means of communication with over 3 billion active users, has become a veritable medium of choice for cybercriminals [2]. Thus, cybercrimes proliferate very rapidly and have the potentials to cause immense damage to both individuals and corporate organizations [3] [4]; phishing is perhaps the most popular of these crimes.

Phishing is an attack vector that deploys technical subterfuge and social engineering to surreptitiously obtain otherwise personal and sensitive information such as credit card pins, and user IDs [5]. Unsuspecting users are lured by criminal elements, masquerading as legitimate entities via electronic communication media to divulge vital, personal, often, financial information, which may, in turn, be used illegally by the criminals without the knowledge or consent of the real owners. Phishing is an instance of identity theft [6]. The phishing cycle often starts with an email that replicates the identity of a trusted associate or organization often with a bogus but juicy claim to a reward for the unsuspecting recipient, or in other instances, a dubious revalidation exercise by elements posing as financial institutions, demanding that users supply their authentication details. For instance, an attacker may describe a problem (usually generic), which in some cases may apply to the unsuspecting recipient. They then proceed to propose a solution, which usually include a link for filling out sensitive details, and a link to reject the offer, to give the email some modicum of authenticity. Having taken the bait, the user is made to fill out personal data such as bank account PIN, social security number, or some other useful authentication details, which may be used by the criminals to perpetrate illegal transactions later. Figure 1 shows the phishing e-mail lifecycle, while Figure 2 shows a oftypical information component in a phishing email.

Phishing attacks pose serious risks to both individuals and corporate entities and have dire consequences on global security and the economy [7]. It is even more so dangerous, as it appears that phishers continue to perfect means to outmaneuver even the knowledgeable and security-conscious [8]; technology giants such as Google and Facebook have lost about $100 million to phishing emails from hackers who impersonated as hardware vendors in 2017. The economic effect of phishing attack is enormous; reports have shown that financial loss occasioned by phishing attacks exceeds $5 billion globally [9].

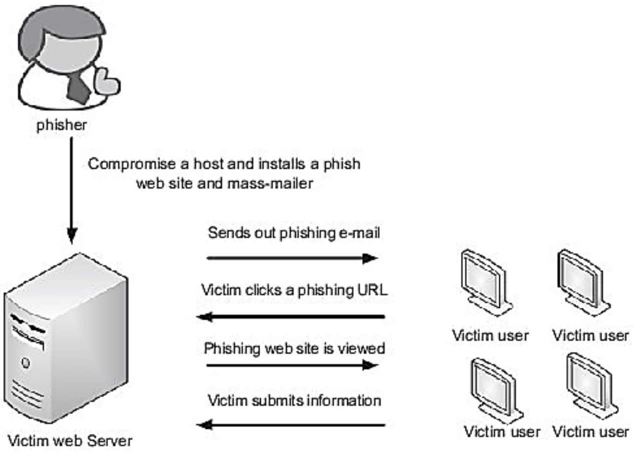
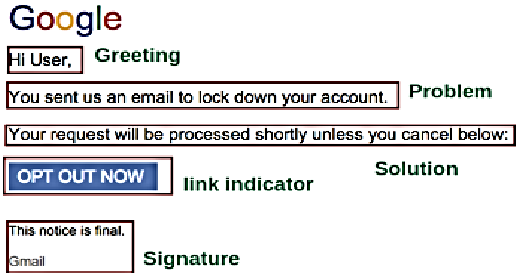

phisher
Compromise a host and installs a phish web site and mass-mailer
Sends out phishing e-mail
Victim clicks a phishing URL
Phishing web site is viewed
Victim submits information
Victim web Server
Victim user
Victim user
Victim user
Victim user

Figure 1. Lifecycle of a phishing email [10]



Hi User,

You sent us an email to lock down your account.

Your request will be processed shortly unless you cancel below:

This notice is final.
Gmail

Figure 2. Typical information component in a phishing email [11]

Phishing attackers are increasingly becoming more resilient over the years, due to the alarming increase in the volume of attack and the innovativeness with which the attacks are being implemented. Security specialists and phishers are in a vicious circle because apprehending phishers have become more and more complicated. Phishers are constantly changing their tactics to defeat anti-phishing techniques [12]. The aggregate number of distinctly recognized phishing attacks reached a peak of 263,538 attacks in the first quarter of 2018; an alarming upsurge from 180,577 reported in the last quarter of 2017 (APWG, 2018) [13]. The email has also been identified as the top phishing target; consequently, a phishing email attack aimed at individuals and corporate bodies is on the rise [14].

Several interventions have been made over the years to combat the phishing menace. Qabajeh *et al.* [15] identified some techniques both ‘traditional’ and ‘computerized’ in literature. Some traditional anti-phishing techniques like enforcing laws, equipping users with knowledge, and educating the public were mentioned. Computerized efforts include blacklists, filtering, associative classification, and rule induction as well as the use of machine learning approaches via different classification and model-based techniques. A variety of surveys and reviews of anti-phishing techniques have also been documented in the literature, to provide better understanding and enhance the development of better anti-phishing systems.

Aleroud and Zhou [16] documented some anti-phishing techniques in emails, websites, mobile devices as well as social networking sites. They then proposed a new taxonomy of phishing attacks with an emphasis on the target environment, attacking techniques, communication media, and countermeasures. Their work offered a robust approach to identifying phishing attacks. Goel and Jain [17] provided a classification of mobile phishing attacks and suggested better methods to identify and ensure protection against these attacks. It was shown that individuals that use mobile devices were more likely to be exposed to phishing attacks than desktop users. Also, due to differences in functionality and layout of the devices, they proposed a devise-centric method that considers the device in use and was able to counter mobile phishing attacks. Sumanthi and Damodaram [18] surveyed seventeen phishing detection schemes with a performance evaluation based on several parameters which included accuracy, precision, recall, true negative estimate, true positive estimate, false-negative estimate, and false-positive estimate. Results indicated the target validation method had the highest accuracy of 99.54%, the phishing alarm had the highest precision of 100% and the smart website combined with the categorization model method for phishing detection produced a 98.72% recall value. Chiew *et al.* [19] presented a robust, systematic review of phishing attack and their associated vectors. The review showed that phishers make use of technical approaches such as cloud computing, clickjacking, and malvertising in their attacks and the development of intelligent systems will be a countermeasure in the discovery of phishing threats. Also, a review of anti-phishing methods in literature was suggested for the development of a more robust technique. Available approaches in literature have been noted to either compromise precision to improve response time or improve precision at the expense of response time [20]. Software phishing detection models generally include the black/whitelist, heuristics, machine learning, visual similarity, and hybrid approaches.

Sonowal and Kuppusamy [21] proposed a phishing detection model called PhiDma which used a multifilter approach. The proposed model employed the use of an audio-based indicator making it accessible by people who are visually impaired. Legitimate data was gotten from phishload and phishing data from phishtank for the implementation. The result from the experiment showed that the model was successful since a true positive rate of 90.54% and 94.18% true negative rate was recorded. Also, an accuracy of 92.72%, a false positive and false negative rate of 5.82% and 9.46% respectively were achieved by the model. Volkamer *et al.* [22] presented a walk-through analysis of reasons why people fall prey to phishing and suggested a concept to circumvent the process: The tooltip-powered phish email detection (Torpedo). Torpedo showed the original uniform resource locator (URL) of email addresses with the domain highlighted for easy recognition of phishing emails. The tool was evaluated in an email environment since this can be adapted to suit other messaging environments. The efficiency of Torpedo was tested against the status quo status bar in Thunderbird and results showed that Torpedo detected fraudulent emails 85.17% more times than the status bar URL which had 43.31%. Moghimi and Varjani [23] proposed two features sets to improve phish detection in internet banking. The support vector machine (SVM) was used to classify webpages with a feature vector consisting of 17 features: 9 relevant features and 8 suggested features. The results of their experiment indicated a true positive value of 99.14% and false negative, 0.86%.

Sahingoz *et al.* [24] designed an anti-phishing system based on machine learning that combined seven classification algorithms with natural language processing features. A dataset of 73,575 URLs, consisting of 36400 originally correct URLs and 37175 phishing URLs were constructed to evaluate the performance of the system. Results revealed that the performance of the proposed system was increased by 2.24% and 13.14% for natural language processing (NLP) based features and word vectors respectively. Also, the Random forest algorithm with NLP features produced the highest accuracy rate of 97.98% when compared to the seven other algorithms, Naive Bayes algorithm, k-nearest neighbor (n=3), Adaboost, sequential minimal optimization, K-star, and decision trees. The use of parallel processing and deep learning was suggested for future research. With the use of reinforcement learning, Smadi *et al.* [25] developed a novel approach in the detection of phishing attacks against online systems. Their proposed system called phishing email detection system which used a feature evaluation and reduction algorithm adjusted regularly to reflect changes in the environment that is, explored new behaviors in a new dataset. For classification, a neural network was combined with reinforcement learning in the designed system and used 50 features. A dataset containing 7315 phishing emails, 4951 ham emails for training and 26722 phishing URL's was used. 44% of the emails were used as the training dataset. Results of the performance evaluation over 50 independent runs showed an accuracy of 98.63% with a true positive and true negative rate of 99.07% and 1.81%.

A case-based reasoning phishing detection system developed by Abutair and Belghith [26] combined both online and offline detection of phishing attacks. The proposed system which used a relatively small dataset (572 cases) was very adaptive and able to predict a zero-hour phishing attack easily. The result showed the proposed system produced an accuracy of 95.62%. Hadi *et al.* [27] experimented on 11,055 phishing websites using a WEKA software environment. They proposed a fast-associative classification algorithm (FACA) for identifying phishing websites. The proposed algorithm outperformed other associative classification algorithms in classification accuracy and F1 evaluation measures. Zhang *et al.* [28] developed a modified deep neural network model in vaticinating phishing attacks. The hybrid deep neural network model combines an autoencoder with a convolutional neural network to be able to detect phish attacks in receivable time. The model was compared with the SVM, decision tree, and LinearSVC algorithm on a deep learning platform, Tensorflow. The results indicated that the developed architecture had an accuracy of 97.87%, precision of 98.69, and a recall of 97.20% which was the best of the four models considered. Similarly, [29] deployed the use of two classifiers, SVM and decision tree, to develop a learning-based aggregation analysis system in detecting phishing attacks on web pages. The results showed that this system enhanced the performance of existing anti-phishing methods. To safeguard sensitive information of users, it is crucial that an adequate means of identifying and apprehending phishing emails be developed. In this study, the maximum entropy (Max-Ent) classification method using parsimonious, but optimal features was implemented.

## 2. RESEARCH METHOD

In this section, the methodology used including data collection and the classification task is described. This includes the data collection and preparation process. The maximum entropy (ME) model was equally depicted using mathematical models.

### 2.1. Building corpora with parsimonious features

The dataset used for the study was from publicly available repositories by [30] (for phishing e-mail dataset) and [31] (for ham e-mail dataset). In total, we worked with 8266 e-mail corpora with 47 features which are commonly used in literature [32]-[34]. Of the 8266 corpora, we designated 6266 e-mails as our training data, leaving us with 2000 test data (1000 hams and 1000 phishes). Furthermore, we carried out a dimensionality reduction of the feature set from 47 to 27 using regression. This was based on the thinking that it is possible to get the parsimonious few ‘principal’ features, thus eliminating redundant features without much information loss.

### 2.2. Maximum entropy

The maximum entropy (ME) is a probabilistic model, based on the ‘principles of maximum entropy’. Maximum entropy has a well-established history in efficiently solving the text classifier problem. Additionally, maximum entropy is adaptable to a large feature set and its performance is not affected by the feature selection method [35]. Maximum entropy determines probabilities based on the principle of making minimal assumptions as follow:

Suppose that we have a set of features, a set of functions $f_1, \dots, f_m$ (by which we may determine the contribution of each feature to the model) and a set of conditions; we determine the probability distribution that satisfies the given conditions and minimizes the relative entropy (divergence of Kullback-Leibler) $D(p||p_0)$, with respect to the distribution $p_0$.

The conditional maximum entropy model is an exponential with the form:

$$p(x|y) = \frac{1}{Z(y)} \prod_{i=1}^{j} a_i^{f_i(y,x)}$$

where $p(x|y)$ denotes the probability of occurrence of outcome $x$, given context $y$ with constraint or feature functions $f_i(x|y)$.

ME model represents evidence with binary functions known as contextual predicates in the form:

$$f_{cp,y'}(x,y) = \begin{cases} 1 & if\ x = x' and\ cp(y) = true \\ 0 & otherwise \end{cases}$$

$cp$ is the contextual predicate which maps a pair of outcome o and context h to {true; false} [35].

## 3. RESULTS AND ANALYSIS

In this section, we report and evaluate the results of the maximum entropy classification techniques vis-à-vis the Naive Bayes (Baseline) and support vector machine (SVM). We describe some benchmark metrics used for the evaluation. We report the performances of the techniques using tables and charts.

### 3.1. Performance metric

The performance metrics used to evaluate our work were accuracy, precision, recall, and error rate. This was calculated based on the correctness or otherwise of the classified test data depicted by true positive, true negative, false positive, and false negative. True positive is the correctly classified phish, true negative is the correctly classified ham, false-positive depicts phishes wrongly classified as ham while false negative depicts hams wrongly classified as phishes. In the context of our study, we define them as follows:

$tp = total\ number\ of\ true\ positives$
$tn = total\ number\ of\ true\ negatives$
$fp = total\ number\ of\ false$
$fn = total\ number\ of\ false\ negatives$

We depict the confusion matrix and presentation of results in tabular form as in Table 1.

Table 1. Confusion matrix of the classification ($m = MaxEnt, s = SVM, n = Naive\ Bayes$)

| | | Phishes | Ham |
|---|---|---|---|
| Phishes | $m$ | 996 | 4 |
| | $s$ | 968 | 31 |
| | $n$ | 972 | 27 |
| | | 165 | 835 |
| | | 3 | 998 |
| | | 5 | 996 |

Accuracy: This is the percentage of correctly classified mails (hams as well as phishes). This is given as:

$$Accuracy = \frac{t_p + t_n}{total\ email\ test\ data}$$

Precision: This is the total number of true positives divided by the total number of emails identified as hams. This is given as;

$$precision = \frac{t_p}{t_p + f_n}$$

Recall: This is the percentage of correctly classified phishes. This is given as;

$$recall = \frac{t_p}{t_p + f_p}$$

Error rate, this is given as;

$$Error\ rate = 1 - accuracy$$

### 3.2. Performance measure and discussion

We measured the performance of our work with reduced feature space of 27, relative to Naïve Bayes (baseline) and SVM which has 47 features. We present the confusion matrix in Table 1 and the performance measure of the 3-classification scheme in Table 2. We present a plot of true value against the values predicted by our classification scheme which is depicted in the confusion matrix in Figure 3. The performance metrics, that is the number of a true positive, true negative, false positive and false negative is presented graphically in Figure 4. This was used to evaluate the performance of the classification models which is presented graphically in Figure 5.

Table 2. Performance measure of 3 classification schemes

| Parameter | Maximum Entropy | SVM | Naïve Bayes |
|---|---|---|---|
| True positive ($\boldsymbol{t_p}$) | 996 | 968 | 972 |
| True negative ($\boldsymbol{t_n}$) | 835 | 998 | 996 |
| False positive ($\boldsymbol{f_p}$) | 165 | 3 | 5 |
| False negative ($\boldsymbol{f_n}$) | 4 | 31 | 27 |
| Accuracy | 0.9155 | 0.983 | 0.984 |
| Precision | 0.99523242 | 0.969873664 | 0.973607038 |
| Recall | 0.835 | 0.997002997 | 0.995004995 |
| Error rate | 0.0845 | 0.017 | 0.016 |
| No of Features | 27 (dimensionality reduction from 47 to 27 through regression) | 47 | 47 |

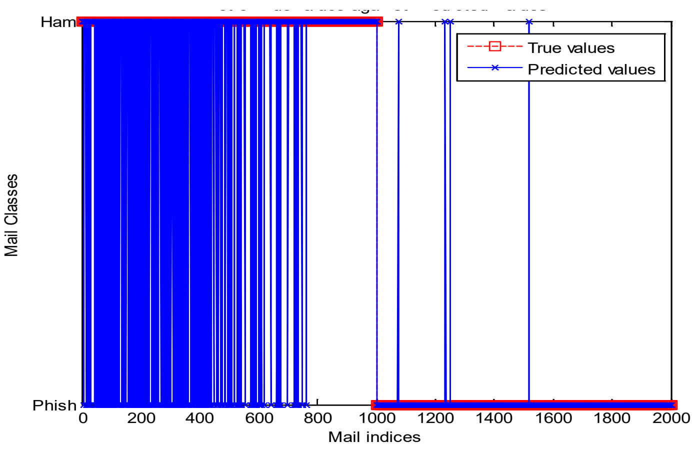


Figure 3. Plot of true value against predicted value

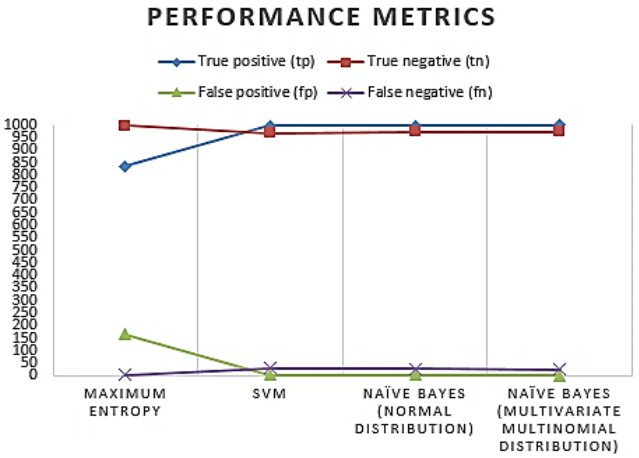


Figure 4. Performance metric

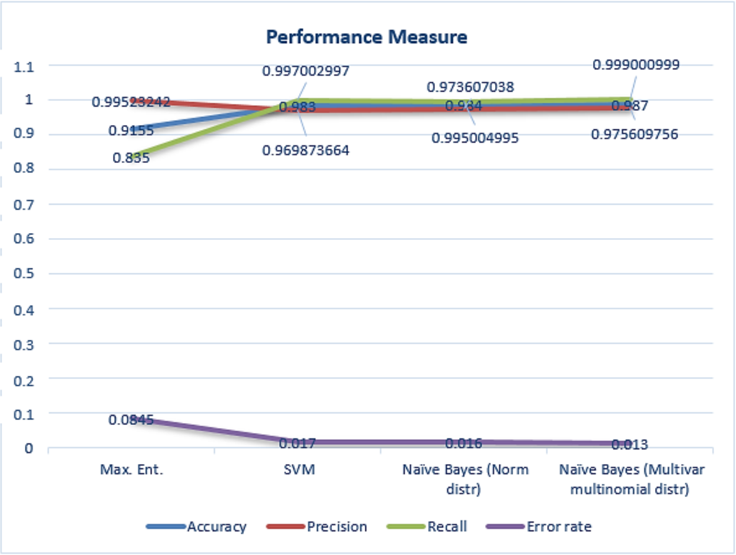


Figure 5. Performance evaluation of maximum entropy, SVM and Naïve Bayes

## 4. CONCLUSION

The maximum entropy with parsimonious feature space clearly outperformed both the Naïve Bayes and SVM in terms of precision. Naïve Bayes, however, had the highest accuracy rate in our experiment. Future work may include hybridizing maximum entropy which has the highest precision and Nave Bayes which has the highest accuracy. We note that maximum entropy is generative while Naïve Bayes is discriminative; it will be interesting to see the result of the hybrid of both.

## ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the Information Systems Modelling, Simulation & Security Research Group (iSyMS)-www.isyms.tech and the Society for Multidisciplinary & Advanced Research Techniques (SMART) for access to research repositories, literary materials and useful insights from partners. We equally appreciate the funding provided by the Landmark University Center for Research Innovation and Development (LUCRID).

## REFERENCES

[1] E. O. Asani, A. Omotosho, O. B. Longe, J. O. Omonigho and B. Gbadmosi, “A Real-time Gesture Engineered CAPTCHA,” *International Journal of Mechanical Engineering and Technology* vol 9, no 12, pp. 618-629, 2018.

[2] S. S. Nair, M. M. Sherin and T. Santha, "Deduplication Enabled Secure E-mail Server on Cloud Environment using Virtual Data Optimizer," *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 270-275, doi: 10.1109/ICACCS48705.2020.9074463.

[3] J. Hong, “The state of phishing attacks,” *Communications of the ACM,* vol 55, no 1, pp. 74-81, 2012, doi: 10.1145/2063176.2063197.

[4] S. A. Robila and J. W. Ragucci, "Don’t be a Phish: Steps in User Education,” *ACM SIGCSE Bulletin,* vol. 38, no. 3, pp 237-241, September 2006, doi: 10.1145/1140123.1140187.

[5] E. O. Asani and A. A. Adegun “Maximum Phish Bait: Towards Feature Based Detection of Phishing using Maximum Entropy Classification Technique,” *In proceedings of International Conference on Science, Technology, Education, Arts, Management and Social Sciences iSTEAMS*, 2014.

[6] G. Xiang, “Toward a Phish Free World: A Feature-type-aware Cascaded Learning Framework for Phish Detection,” Doctoral Dissertation, Language Technologies Institute School of Computer Science Carnegie Mellon University, Pittsburgh, 2013.

[7] R. Islam and J. Abawajy, “A multi-tier phishing detection and filtering approach,” *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 324-335, 2013, doi: 10.1016/j.jnca.2012.05.009.

[8] R. W. Lucky, “Clickphobia [Reflections],” *IEEE Spectrum*, vol 48 no, p. 25, 2011, doi: 10.1109/MSPEC.2011.5676377.

[9] A. K. Jain and B. B. Gupta “Phishing Detection: Analysis of Visual Similarity Based Approaches” *Security and Communication Networks,* vol. 2017, pp. 1-20, 2017, doi: 10.1155/2017/5421046.

[10] M. Alauthman, A. Almomani, M. Alweshahm, W. Alomoush and K. Alieyan, “Machine Learning for Phishing Detection and Mitigation,” B. B. Gupta, & Q. Z. Sheng, (Eds.). *Machine Learning for Computer and Cyber Security: Principle, Algorithms, and Practices* (1st ed.), CRC Press, pp 26-47, 2019, doi: 10.1201/9780429504044.

[11] N. Maleki and A. A. Ghorbani, “Generating Phishing Emails Using Graph Database,” In: Heng S. H., Lopez J. (eds) *Information Security Practice and Experience. ISPEC 2019. Lecture Notes in Computer Science,* vol. 11879, 2019. Springer, Cham. doi: 10.1007/978-3-030-34339-2_25.

[12] A. Hamid, I. Rahmi and A. Jemal, “Profiling Phishing Email Based on Clustering Approach,” *In proceedings of 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications,”* 2013, pp. 628-635, doi: 10.1109/TrustCom.2013.76.

[13] AWPG, “Phishing Activity Trends Report,” 2nd Quarter 2018.

[14] Phishlab, “2018 Phishing Trends and Intelligence Report: Hacking the Human,” 2018. [Online]. Available at: https://info.phishlabs.com/hubfs/2018%20PTI%20Report/PhishLabs%20Trend%20Report_2018-digital.pdf. Accessed 22/02/2019.

[15] I. Qabajeh, F. Thabtaha and F. Chiclana, “A recent review of conventional vs. automated cybersecurity anti-phishing techniques,” *Computer Science Review*, vol. 29, pp. 44-55, 2018, doi: 10.1016/j.cosrev.2018.05.003.

[16] A. Aleroud and L. Zhou, “Phishing environments, techniques, and countermeasures: A survey,” *Computers and Security*, vol. 68, pp. 160-196, 2017, doi: 10.1016/j.cose.2017.04.006.

[17] D. Goel and A. K. Jain, “Mobile phishing attacks and defence mechanisms: State of art and open research challenges,” *Computers and Security*, vol. 73, pp. 519-544, 2018, doi: 10.1016/j.cose.2017.12.006

[18] K. Sumanthi and R. Damodaram, “Survey and Analysis on Phishing Detection Techniques,” *International Journal of Advanced Research in Computer Science*, vol. 9, no. 1, 2018.

[19] K. L. Chiew, K. S. C. Yong and C. L. Tan, “A survey of phishing attacks: their types, vectors and technical approaches,” *Expert Systems with Applications*, vol. 106, pp. 1-20, 2018, doi: 10.1016/j.eswa.2018.03.050.

[20] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg and E. Almomani, "A survey of phishing email filtering techniques," *IEEE communications surveys & tutorials*, vol. 15, no. 4, pp. 2070-2090, 2013, doi: 10.1109/SURV.2013.030713.00020.

[21] G. Sonowal and K. S. Kuppusamy, "PhiDMA-A phishing detection model with multi-filter approach," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 1, pp. 99-112, 2017, doi: 10.1016/j.jksuci.2017.07.005.

[22] M. Volkamer, K. Renaud, B. Reinheimer and A. Kunz, "User experiences of TORPEDO: TOoltip-poweRed Phishing Email Detecti On," *Computers & Security*, vol. 71, pp. 100-113, 2017, doi: 10.1016/j.cose.2017.02.004.

[23] M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," *Expert Systems with Applications*, vol. 53, pp. 231-242, 2016, doi: 10.1016/j.eswa.2016.01.028.

[24] O. K. Sahingoz, E. Buber, O. Demir and B. Diri, "Machine learning based phishing detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345-357, 2019, doi: 10.1016/j.eswa.2018.09.029.

[25] S. Smadi, N. Aslam and L. Zhang, "Detection of online phishing email using dynamic evolving neural network based on reinforcement learning," *Decision Support Systems*, vol. 107, pp. 88-102, 2018, doi: 10.1016/j.dss.2018.01.001.

[26] H. Y. A. Abutair and A. Belghith, "Using Case-Based Reasoning for Phishing Detection," *Procedia Computer Science*, vol. 109, pp. 281-288, 2017, doi: 10.1016/j.procs.2017.05.352.

[27] W. Hadi, F. Aburub and S. Alhawari, "A new fast associative classification algorithm for detecting phishing websites," *Applied Soft Computing Journal*, vol. 48, pp. 729-734, 2016, doi: 10.1016/j.asoc.2016.08.005.

[28] X. Zhang, D. Shi, H. Zhang, W. Liu and R. Li, "Efficient Detection of Phishing Attacks with Hybrid Neural Networks," *2018 IEEE 18th International Conference on Communication Technology (ICCT),* 2018, pp. 844-848, doi: 10.1109/ICCT.2018.8600018.

[29] J. Mao, *et al.*, "Detecting Phishing Websites via Aggregation Analysis of Page Layouts," *Procedia Computer Science*, vol. 129, pp. 224-230, 2018, doi: 10.1016/j.procs.2018.03.053.

[30] J. Nazarios, "Phishing Corpus," 2018 url: https://monkey.org/~jose/phishing/ Accessed 26/01/2019.

[31] SpamAssasin, "Public Corpus," 2018. url: https://spamassassin.apache.org/old/publiccorpus/. Accessed 26/01/2019 url: https://docs.apwg.org/reports/apwg_trends_report_q2_2018.pdf. Accessed 22/02/2019.

[32] M. Khonji, A. Jones and Y. Iraqi, "A Study of Feature Subset Evaluators and Feature Subset Searching Methods for Phishing Classification," *In proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS '11)*, ACM New York, USA, 2011, pp. 135-144, doi: 10.1145/2030376.2030392.

[33] I. R. A. Hamid, J. Abawajy and T. H. Kim, "Using Feature Selection and Classification Scheme for Automating Phishing Email Detection," *Studies in Informatics and Control*, vol. 22, no. 1, 61-70, 2013, doi: 10.24846/v22i1y201307.

[34] N. Vaishnaw and S. R. Tandan, "Development of Anti-Phishing Model for Classification of Phishing E-mail," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 4, no. 6, pp. 39-45, 2015, doi: 10.17148/IJARCCE.2015.4610 39.

[35] L. Zhang and Y. Tian-shun, "Filtering junk mail with a maximum entropy model," *In Proceeding of 20th International Conference on Computer Processing of Oriental Languages (ICCPOL, 2003)*, 2003, pp. 446-453.