USING CLASSICAL MACHINE LEARNING FOR PHISHING WEBSITES DETECTION FROM URLS

Iman Akour, University of Sharjah
Noha Alnazzawi, Yanbu University College
Ahmad Aburayya, Dubai Health Authority
Raghad Alfaisal, Universiti Pendidikan Sultan Idris
Said A. Salloum, University of Salford

ABSTRACT

Phishing is one of the various types of internet frauds that many people fall victim to. Scammers use phishing attacks to gain access to a user's sensitive information. This is done by creating fake websites that appear to be legitimate websites belonging to prestigious organizations. As a result, there is an urgent need to conduct research on the phenomenon of phishing attacks, which will prove extremely beneficial to individuals working in cyber security and phishing attack prevention firms. Blacklisting websites, retrieving characteristics from a website, raising awareness amongst people, and drawing parallels between phishing attacks and known patterns of prior phishing attacks are some of the ways currently used for detecting phishing attacks. To analyze and classify phishing websites, this paper employs classification models. The classification models are created by extracting phishing website features. The model was trained using machine learning algorithms such as "Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), and NAÏVE BAYES (NB)"classifiers on two different datasets containing 58,645 and 88,647 websites identified as phishing and legitimate URLs, respectively. SVM was discovered to be the best algorithm for the detection of phishing URLs, showing a degree of accuracy of 96.30 percent.

Keywords: Blacklisting Websites, Phishing Attacks, Machine Learning, Scammers

INTRODUCTION

As cyber security attacks continue grow in scale and sophistication, social engineering has become a frequently used method for counter attack and is regarded one of the most effective and simple strategies for obtaining inside and private data (Al Mansoori et al., 2021; Salloum et al., 2020; Salloum et al., 2021). Anti-Phishing Working Group (APWG) defines phishing as a criminal activity involving theft of financial records and identity through technical deception and social engineering (Yousuf, Al-Hamad, & Salloum, n.d.). Additionally, obfuscated messages and email addresses are used in a social engineering scheme in which innocent people who are unaware of the internet frauds are targeted (Alghizzawi et al., 2018; Alshurideh et al., 2018; Habes et al., 2019; Salloum, Khan & Shaalan, 2020). They are persuaded to believe that the emails sent and the senders as well are associated with a legitimate and trustworthy group or organization. The emails are created in such a way that they direct users to fraudulent websites, tricking them into disclosing their username and password, which is crucial financial information. Technical subterfuge, on the other hand, installs malicious software on a computer system with the goal of stealing data by tricking users into visiting fake websites or stealing their personal information (Anti-Phishing Working Group, 2020). Furthermore, software such as HT Track is freely accessible to customers, enabling them to duplicate any website and utilize it for any

Citation Information: Akour, I., Alnazzawi, N., Aburayya, A., Alfaisal, R., & Salloum, S.A. (2021). Using classical machine learning for phishing websites detection from URLs. *Journal of Management Information and Decision Sciences, 24*(S6), 1-9.

purpose, which is often to replicate and create bogus websites. Simultaneously, as far as preventative measures are concerned regarding the aforementioned assaults, it is imperative for organizations to educate users on how to spot phishing emails or fraudulent links. The lack of awareness regarding this subject is the reason why even educated individuals make the mistake of visiting a hostile website and assuming it for a real one, and ultimately end up revealing their personal and sensitive information.

Evidently, user education is just as important as employing computer-based models to detect and avoid phishing schemes when it comes to combating phishing attacks. Identifying malicious websites is one approach of detecting phishing attacks. The Uniform Resource Locators or URL of a website is its internet address and the most basic method of accessing it. URLs can be extremely handy in the phishing detection process. They help to distinguish between genuine and bogus websites, making them a useful tool for detecting fraudulent phishing websites. Therefore, setting up computer systems with programs that analyze URLs and identify malicious sites can be a viable phishing attack solution.

A program or method that can successfully and quickly classify novel URLs as linked with either phishing or legitimate websites appears to be the most suited solution to the problem (Salloum, Gaber & Vadera, 2021). While a similar strategy has been previously used, in which the anti-virus class established a blacklist of malicious URLs, the flaw of this approach was the frequent introduction of new malicious URLs that were not included in the blacklist and so were not identified. Training groups can utilize machine-learning technology to construct models that make detecting and classifying phishing websites easier.

In essence, URLs are made up of features, of which two categories exist: host-based and lexical features. While host-based features are concerned with website characteristics such as location, installation date, and website manager. Lexical features, on the other hand, refer to the URL's textual characteristics. When machine learning techniques are used to identify phishing URLs, the lexical or host-based features or a combination of them extracted from a URL are used. Owing to the fact that URLs are only text strings that can be broken down into subparts like hostname, protocol, and path, a site's validity can be determined by any combination of those elements.

Machine learning techniques have traditionally been used to detect malicious URLs, and numerous studies have also studied and analyzed them to further investigate the subject of phishing detection. The study by Ma, et al., (Ma, Saul & Voelker, 2011), for example, includes the development of a URL classification system that can process labeled URLs in real-time. According to Ma, et al., URL classification is a binary problem which is why their URL classification model not only collects URL features in real time from a major Internet mail provider, but also uses lexical and host-based features, as well as trains the online classifier using the Confidence Weighted (CW) algorithm. Sadeh, et al., (2007), on the other hand, employed PILFER to classify phishing using special features in order to detect fraudulent techniques used to defraud individuals. The data of 860 phishing and 6950 legitimate emails was classified using the Support Vector Machine (SVM) classifier. The classifier was assessed using ten-fold cross validation, yielding 92 percent accuracy. Finally, Parkait, et al., (Purkait, 2012) examined 358 papers published on the issue of phishing, including preventative strategies employed to counter it and their efficacy. They further compiled a comprehensive literature analysis that divided anti-phishing methods into eight categories. In their research, they also identified more complex anti-phishing techniques.

Machine-learning algorithms such as “Support Vector Machine (SVM), K-Nearest neighbor (KNN), Logistic Regression (LR), and NAÏVE BAYES (NB)” classifiers were used to identify URLs as either phishing or real, in keeping with the goal of this study, which was to deploy machine-learning algorithms on phishing datasets. Unfortunately, because training datasets containing phishing URLs can be hard to come by in the public domain, implementing the machine-learning approach whose performance is significantly dependent on the collected data sets becomes a challenge. This study makes use of the data collection by extracting components from the data URLs and making class labels available.

The following sections of this paper go over the study's various aspects: Section 2 of this paper deals with classifying phishing URLs, section 3 discusses the specific characteristics of the data set and the methodology used, section 4 delves into the findings of the research, and section 5 elaborates on the limitations that arose for the current study as well as the future prospects of this research.

URLS AND ATTACKERS’ TECHNIQUES

The elements that make up a URL might be observed in order to gain insight into the brains of phishing attackers. Figure 1 depicts the general structure of a URL. The URL of a webpage begins with a protocol name which is stated in its standard format. The identity of the organization, as well as the information about the server that hosts it, may be found in the subdomain and the Second Level Domain (SLD). Finally, the Top-Level Domain (TLD) name functions to identify the domains in the internet's DNS root zone. This section of the paper delves into the many different tactics used by phishers to remain undetected by the system admins and the security systems.

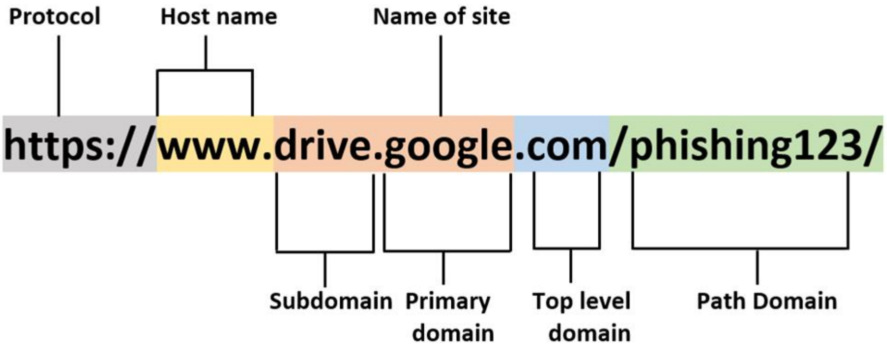

FIGURE 1
THE PARTS OF A URL

The internal address can be seen on the server page, whereas the domain name (host name) of the web page can be discovered in the previous sections, and the page name is visible in the HTML structure. When it comes to the names used for phishing, cyber security firms make a concerted effort to identify any illegitimate domains based on the name. While the structure of SLD and TLD creates a domain name, which is the most distinguishing and crucial portion of a URL, yet, blocking IP addresses proves to be a viable way to prevent access to web pages contained in the domain name if in any instance they are suspected of being phishing. Furthermore, in view of the fact that the SLD name can be produced only once, the attacker may easily be able to identify or acquire it for the purpose of phishing, despite the fact that the firm name and type of activity are normally stated on the SLD name. In addition, because the internal address structure is directly dependent on the attacker, attackers can build an unlimited number of URLs by expanding the SLD. This is achieved by means of the path and file names. Since certain methods can potentially contain key ways to weaken users, such as cybersquatting, typo squatting, changeable characters, and joint use of words, the detection system must be able to analyze and address the perpetrators' strategies. This helps refine their assaulting capabilities so as to acquire increasingly sensitive information.

METHODOLOGIES

This section delves into the discussion of the specifics of the experimental approach used to detect and identify phishing websites using URL features retrieved from websites. Such a method employs the following resources and techniques: feature selection, pre-selected data sources, performance evaluation metrics, and ML algorithms. This particular experimental approach begins with the selection of a phishing website from which the data is obtained, and the characteristics identified as the major ones are analyzed using feature selection algorithms. Once standardization is achieved for the features, the resulting features undergo training using machine learning models such as “SVM, KNN, LR, and NB” after which they are entered into the machine learning classifiers. The algorithm that demonstrates the highest efficacy (Hejazi, Khamees, Alshurideh & Salloum, 2021; Khamees, Hejazi, Alshurideh & Salloum, n.d.; Salloum, Wahdan, Salloum & Shaalan, 2021) is used to detect phishing websites using URLs (classifying the websites as either legitimate or phishing website). Figure 2 illustrates the proposed technique for this model.

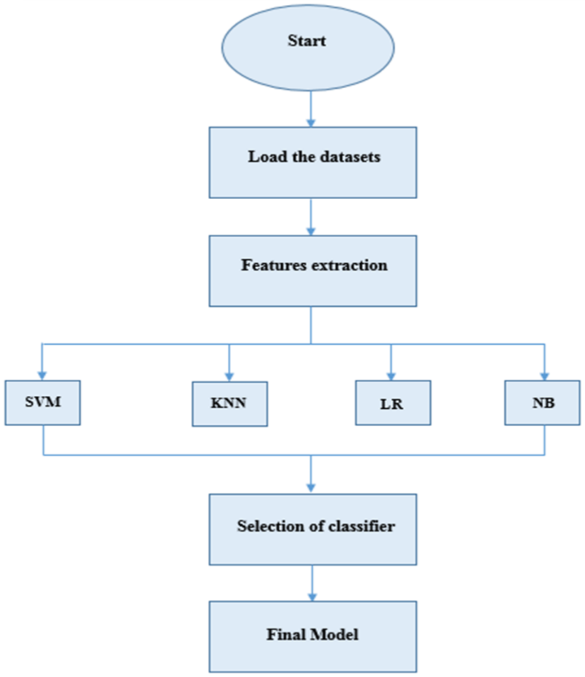

FIGURE 2
THE PROPOSED APPROACH

Dataset

The dataset can also be used to acquire more examples and warnings about phishing and legal proceedings which make gathering data not only an important but also the initial step in the process. The dataset stage is critical for maintaining the conclusion validity. The outcome acquired from the analyzed information is used to carry out further examination as well as the prediction and anticipation of any developments in phishing. (Vrbančič, Fister Jr, & Podgorelec, 2020) was used to compile all of the components. Figure 3 further shows that the value of the 4 categories is dependent on distinct sub-strings. The first group, however, is dependent on the values of the features across the whole URL string (Vrbančič et al., 2020). Google search index and URL resolve metrics are the two categories that comprise the final group of features. After removing the target phishing property, there remain in the dataset 111 features which function to indicate whether the particular instance is phishing (value 1) or legitimate (value 0). The dataset used is divided into two different versions, totaling 58,645 occurrences, with a nearly balanced equation across the target groups. There are 27,998 instances tagged as legitimate and 30,647 instances that are labeled as phishing websites (Vrbančič et al., 2020). The purpose is to mimic the reality of the original conditions while also allowing the development of other legal websites. The second form of the dataset has 88,647 cases, with 58,000 instances classed as legal and 30,647 instances tagged as phishing. For the sake of this method, we initially obtained a list of 30,647 labeled phishing URLs using the Phishtank website. 58,000 legitimate website URLs were also taken from the Alexa ranking website from their list of legitimate URLs. In addition to that, a collection of 27,998 community-labeled and arranged URLs was obtained (Lab, 2014), and for the reason that these URLs represent impartially reported news, they are considered legitimate. Furthermore, several variations of the datasets were created as indicated above using the URL lists of phishing and authentic websites. The larger and unpredictable dataset contains all of the examples from dataset small, as well as increased examples of feature selection from Alexa's top sites list of URLs. The smaller dataset, on the other hand, contains examples of features derived from Phishtank URLs as well as the cases of extracted features using community labeled and categorized URLs that are made to appear legitimate (Vrbančič et al., 2020).

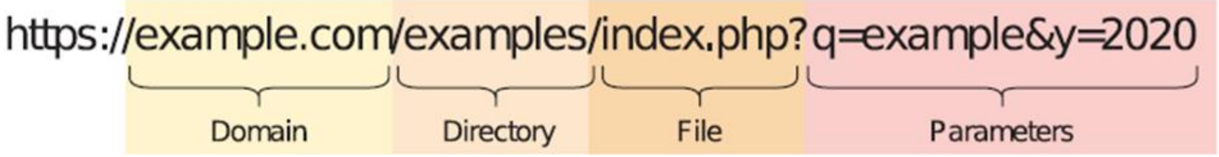

FIGURE 3
URL COMPONENTS (VRBANČIČ ET AL., 2020)

Feature Extraction

Because a direct association was identified between both machine-learning algorithms and features and the performance of the trained system, this research conducted a thorough literature analysis in order to find essential features. Furthermore, studies analyzing features in other groups, such as email content analysis and website analysis, were combined with those studies that were examining the URL, while the URL's attributes were also analyzed independently via hostname, path sections, and domain. Scripts written in the Python programming language were used to achieve the 111 different features observed previously in our investigation (Vrbančič et al., 2020).

System Implementation and Performance Evaluation

Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Naïve Bayes (NB)” were used in the machine learning-based system experiment. In the Python programming language, the Scikit library trained the models built by the aforementioned algorithms.

In the case of a dependent variable having been classified into two distinct classes, the algorithm that is capable of generating predictions most efficiently is Logistic Regression. However, there are limitations in this algorithm in terms of feature compatibility, outlier values, and feature recurrence which significantly weakens its prediction capability. The speed and efficacy of the KNN algorithm are the defining characteristics of this algorithm. It provides estimations depending on the distance between the k neighbors, which requires a great deal of memory space to calculate. It is also critical to have an appropriate k value in order to obtain reliable and valid findings. Although the SVM algorithm is not a difficult or complex system to operate and also has the ability to process a high number of independent variables to provide plausible solutions by using nonlinear problems, it is still not considered the best model when it comes to working with huge datasets. Furthermore, too much noise has a negative impact on its performance. Naive Bayes is a conditional probability-based classification algorithm and its operation is governed by Bayes theorem. The reason that Naive Bayes is favored as an algorithm is because it is simple and easy to implement and requires minimal training to operate. This algorithm, however, is not desirable for a variety of reasons, for example, poorer estimation when working with fewer data and the assumption that the traits are unrelated to one another.

EXPERIMENTAL AND EVALUATION

The experiment was done entirely on a Lenovo (HP Pavilion 15 Gaming, i5-10300H 2.5 GHz, 8 GB, 256 GB SSD, NVIDIA GTX 1050/3 GB), and the methodologies utilized in the experiment were subjected to a 80 percent split and ten-fold cross validation. Once the entire dataset was securely put onto Jupyter Notebook in an Anaconda setting, the four algorithms exhibiting the highest accuracy were used to categorize the URL features. These algorithms were tested on the Hudders field phishing datasets, and their efficacy was analyzed using the following criteria: precision, F-measure, accuracy, and recall. The reason for using these variables was that they are the most accurate indicators of good quality.

To offer an equal simultaneous comparison between models, “SVM, KNN, LR, and NB” a series of experiments were conducted utilizing the typical machine learning approaches. The models were determined based on the extent of their output’s comparability and competitiveness. The findings that were concluded were accurately documented and it was made sure that no bias occurred in model selection. The findings for the complete dataset, obtained using eight different algorithms, show that the SVM classifier, which demonstrates a 96.30 percent accuracy rate, achieves a high-test classification result. Table 1 shows the training time, precision, accuracy, F-measure, and recall rate.

Table 1
TEST RESULTS OF CLASSIFIERS ON DATASET

Algorithm	Accuracy	Precision	Recall	F-Measure
SVM	0.963	0.962	0.963	0.963
KNN	0.94	0.94	0.94	0.94
LR	0.935	0.932	0.935	0.935
NB	0.897	0.895	0.896	0.897

CONCLUSIONS AND FUTURE WORKS

Much has been done in order to reduce the number of phishing attacks that occur, such as exposing phishers, instructing a novice, and improving the visualization and toolbars that are incorporated in the web browser (Abutair & Belghith, 2017; Chen, Guan & Su, 2014; Dogukan, Abdulwakil, & Aydin, 2017; Gupta et al., 2021; Jain & Gupta, 2018; Sahingoz, Buber, Demir & Diri, 2019; Saxe & Berlin, 2017). Phishing detection rates appear to be low while training users regarding phishing is an expensive process. Furthermore, it has recently been shown that a phishing-defeating procedure based on machine-learning algorithms is highly effective (Akinyelu & Adewumi, 2014; Al-Janabi, Quincey & Andras, 2017; Sanglerdsinlapachai & Rungsawang, 2010). This strategy, which is based on some features, use models derived from machine learning techniques in order to identify websites as phishy or authentic. As a result, we focus on implementing a phishing detection system in this study, using machine learning methods. Furthermore, the existing datasets in the literature were used to assess the proposed system, and the results were compared to the most recent study in the literature. The goal of this project is to identify the best machine learning algorithm for detecting phishing URLs by comparing each algorithm's false negative, wrong positive, and accuracy rates. This research investigated the effectiveness of machine learning in terms of phishing detection despite the use of unfavorable ML learning methodologies. This work manages the extraction and examination of numerous aspects of genuine and phishing URLs to detect phishing URLs using machine learning. Machine learning is a promising tool for distinguishing between legal and fraudulent websites. As previously stated, the goal of phishing websites is to steal one's sensitive information, such as credit card numbers, user names and passwords, and other private information. This is accomplished by deceiving and convincing them of the connection to legitimate websites. However, because this approach is susceptible to an adverse machine learning technique, machine learning can be used to reduce the accuracy of a trained classifier model. The following algorithms, for example, “SVM, KNN, LR, and NB” algorithms are used to detect phishing websites. When our model was evaluated using four different machine learning algorithms, SVM produced the best results, with a 96.30 percent accuracy rate. The recommended technique was shown to have high accuracy rates and boosted phishing detection efficacy based on the contrasting results. As a result, for future projects, the focus must be on developing a big and current dataset for URLs that rely on a Phishing Detection System. It is also important to use dataset and strive to enhance our system by applying specific hybrid algorithms as well as NLP-based features models as indicated in (Sahingoz et al., 2019). Lastly, we intend to use an SVM in the internet browser and a large number of beginners in a pilot study to perform active learning. Finally, the several attribute extraction models are combined with the proposed approach and carried out to examine its use in a realistic plan.

Citation Information: Akour, I., Alnazzawi, N., Aburayya, A., Alfaisal, R., & Salloum, S.A. (2021). Using classical machine learning for phishing websites detection from URLs. *Journal of Management Information and Decision Sciences, 24*(S6), 1-9.

REFERENCES

Abutair, H.Y.A., & Belghith, A. (2017). Using case-based reasoning for phishing detection. *Procedia Computer Science, 109*, 281–288.

Akinyelu, A.A., & Adewumi, A.O. (2014). Classification of phishing email using random forest machine learning technique. *Journal of Applied Mathematics*.

Al-Janabi, M., Quincey, E., & Andras, P. (2017). Using supervised machine learning algorithms to detect suspicious URLs in online social networks. *In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.*

Al Mansoori, S., Almansoori, A., Alshamsi, M., Salloum, S.A., & Shaalan, K. (n.d.). Suspicious activity detection of Twitter and Facebook using sentimental analysis.

Alghizzawi, M., Ghani, M.A., Som, A.P.M., Ahmad, M.F., Amin, A., Bakar, N.A., … Habes, M. (2018). The impact of smartphone adoption on marketing therapeutic tourist sites in Jordan. *International Journal of Engineering & Technology, 7*(34), 91–96.

Almansoori, A., Alshamsi, M., Abdallah, S., & Salloum, S.A. (2021). Analysis of cybercrime on social media platforms and its challenges. *In The International Conference on Artificial Intelligence and Computer Vision (pp. 615–625). Springer.*

Alshurideh, M., Al Kurdi, B., Abumari, A., & Salloum, S. (2018). Pharmaceutical promotion tools effect on physician's adoption of medicine prescribing: Evidence from Jordan. *Modern Applied Science, 12*(11), 210–222.

Lab, O. (2014). *URL testing lists intended for discovering website.* Censorship.

Chen, C.M., Guan, D.J., & Su, Q.K. (2014). Feature set identification for detecting suspicious URLs using Bayesian classification in social networks. *Information Sciences, 289,* 133–147.

Dogukan, A., Abdulwakil, A., & AYDİN, M.A. (2017). Detecting phishing websites using support vector machine algorithm. *PressAcademia Procedia, 5*(1), 139–142.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. *In Proceedings of the 16th international conference on World Wide Web.*

Gupta, B.B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. *Computer Communications, 175,* 47–57.

Habes, M., Salloum, S.A., Alghizzawi, M., & Mhamdi, C. (2019). The relation between social media and students' academic performance in Jordan: YouTube perspective. *In International Conference on Advanced Intelligent Systems and Informatics.*

Hejazi, H.D., Khamees, A.A., Alshurideh, M., & Salloum, S.A. (2021). Arabic text generation: Deep learning for poetry synthesis. *In Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021 (pp. 104–116).*

Jain, A.K., & Gupta, B.B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. *In Cyber Security (pp. 467–474). Springer*.

Khamees, A.A., Hejazi, H.D., Alshurideh, M., & Salloum, S.A. (n.d.). Classifying audio music genres using CNN and RNN. *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021*, 315.

Ma, J., Saul, L.K., Savage, S., & Voelker, G.M. (2011). Learning to detect malicious urls. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(3), 1–24.

Purkait, S. (2012). Phishing counter measures and their effectiveness–literature review. *Information Management & Computer Security.*

Sahingoz, O.K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications, 117,* 345–357.

Salloum S., Gaber T., & Vadera S. (2021). Phishing website detection from URLs Using classical machine learning ANN model. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 399.

Salloum, S.A. (n.d.). Classifying audio music genres using a multilayer sequential model. *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021,* 301.

Salloum, S.A., Alshurideh, M., Elnagar, A., & Shaalan, K. (2020). Machine learning and deep learning techniques for Cybersecurity: A Review. *In Joint European-US Workshop on Applications of Invariance in Computer Vision (pp. 50–57). Springer.*

Salloum, S.A., Khan, R., & Shaalan, K. (2020). A survey of semantic analysis approaches. *In Joint European-US Workshop on Applications of Invariance in Computer Vision (pp. 61–70). Springer.*

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2021). Phishing email detection using natural language processing techniques: A literature survey. *Procedia Computer Science, 189,* 19–28.

Sanglerdsinlapachai, N., & Rungsawang, A. (2010). Web phishing detection using classifier ensemble. *In Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services* (pp. 210–215).

Saxe, J., & Berlin, K. (2017). eXpose: *A character-level convolutional neural network with embeddings for detecting malicious URLs, file paths and registry keys.* ArXiv Preprint ArXiv:1702.08568.

Vrbančič, G., Fister Jr, I., & Podgorelec, V. (2020). Datasets for phishing websites detection. Data in Brief, 33, 106438.

Wahdan, A., Salloum, S.A., & Shaalan, K. (2021). Text classification of Arabic text: Deep learning in ANLP. *In Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021 (pp. 95–103).*

Yousuf, H., Al-Hamad, A.Q., & Salloum, S. (n.d.). An overview on CryptDb and Word2vec approaches.