https://doi.org/10.17576/apjitm-2022-1102-04
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 11 No. 2 December 2022: 49 - 58
e-ISSN: 2289-2192

Received: 21 March 2022
Accepted: 28 June 2022
Published: 1 December 2022

THE EFFECTIVENESS OF URL FEATURES ON PHISHING EMAILS CLASSIFICATION USING MACHINE LEARNING APPROACH

AHMAD FADHIL NASWIR
LAILATUL QADRI ZAKARIA
SAIDAH SAAD

ABSTRACT

Phishing email classification requires features so that the performance obtained produces good accuracy. One of the reasons for the lack of development of models for detecting phishing emails is the complexity of the feature selection. Feature selection is one of the essential parts of getting a good classification result, commonly used features are header, body, and Uniform Resource Locator (URL). Besides the email body text content, the URL is one of the leading indicators that the phishing attack successfully happened. The URL is commonly located on the body of the phishing email to get the victim's attention. It will redirect the victim to a fake website to obtain personal information from the victim. There is a lack of information about how the URL features affect the phishing email classification results. Therefore, this work focuses on using URL features to determine whether an email is phishing or legitimate using machine learning approaches. Two public datasets used in this work are the Online Phishing Corpus and Enron Corpus. The URL features are extracted using the Beautiful Soup library. Two machine learning classifiers used in this work are Support Vector Machine (SVM) and Artificial Neural Network (ANN). The experiments were divided into two based on features used in the classifiers. The first experiment used raw email data with URL features, while the second only used raw email data. The first experiment shows higher accuracy in both classifiers, SVM and ANN. Hence, this research proves that the impact of selecting URL features will increase the performance of the classification.

Keywords: Phishing, Phishing Email Classification, Features Selection, URL Feature, Machine Learning

INTRODUCTION

Due to the complexity of current phishing attacks, the detection and classification of phishing attacks is a great challenge. Though many email filters have been developed for spam emails, few phishing email filters have been developed (Bagui & Nandi, 2019). Getting good-quality training data is one of the biggest problems in machine learning because data labeling can be a tedious and expensive task (Sumathi & Sujatha, 2019). From the dataset used by previous researchers, the process of evaluating the dataset is complex because of the limitations of the phishing email dataset and determining the dataset that is appropriate and the same as previous studies. Phishing email datasets used by previous researchers are Online Phishing Corpus, SpamAssassin, Enron, IWSPA, Phishtank, CSDMC 2010, and Custom (Ahmad et al., 2020).

Phishing emails, also known as cyber-attacks, cannot be separated from the existence of the sender; the attacker of deceptive phishing will create an email based on observations and different ways of writing. Their strategy is to create an urgent atmosphere that convinces the victim to react, for example, an account alert or promising reward (Kumar et al., 2020 & Abdullah & Mohd, 2019). The attacker's writing behavior aspects ensure the victims follow the flow of the attacker's email much as possible. Each part of the email has its criteria and characteristic, and each email will be unique because the sender will provide different

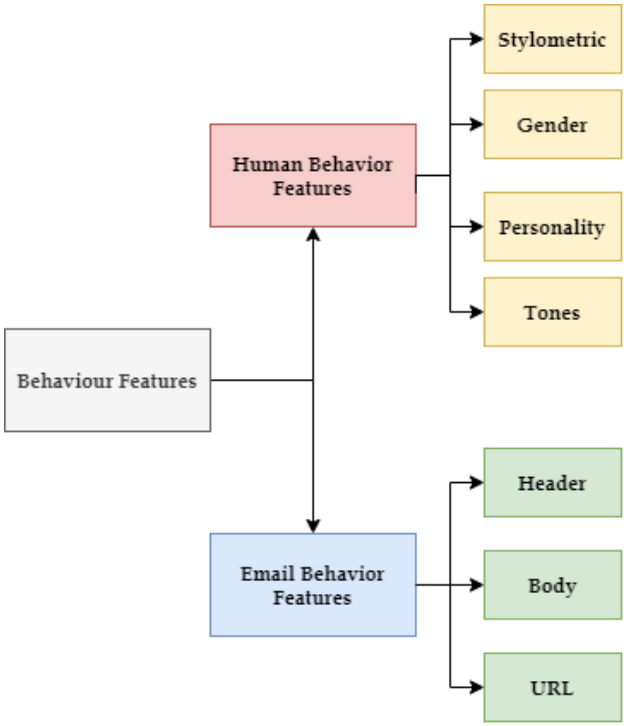
information and data. Using those parts as features in the classification process may have a greater chance of satisfying results. Commonly, phishing emails will include one or more URLs in the content to trick the victim into clicking on the URL and initiating the data phishing process (Shabudin, 2020). The common feature used in phishing email classification is the URL feature (Kumar et al., 2018). Although URL features can improve the classification performance, there is limited information about how significant are URL features in phishing email classification compared to other features. Previous researchers' two most used classifiers are Support Vector Machine and Naïve Bayes. Extraction and selection of features based on the email structures play an essential role in improving email classification performance results on specific content (Mujtaba et al., 2017).

This study extracts email data and URL features to accurately identify and classify phishing emails. The rest of the paper is organized as follows: Section II discusses the related works. The framework for feature extraction of this experiment is in section III, and the feature selection part is in Section IV. Section V shows the experiment and the results and closes with a conclusion in Section VI.

RELATED WORK

Feature selection is a crucial task in conducting email classification research, as selecting the best and most relevant features in the experiment will give a better result. However, no such optimized features are equally applicable in all domains (Iqbal et al., 2010). In recent years, researchers have tried different feature selection, and extraction approaches. Popular features are used in phishing email classification, such as content-based, word embedding, word analysis, and header. By conducting several research reviews of research in phishing email classification in recent years, the most frequently used features are content-based and URL. The main reason is that the phishing email mainly contains links that will redirect the victim into another scheme (Kumar et al., 2018). In some previous research, the features of each category, both in terms of email and humans, play a vital role in email classification, whether ham, spam, or phishing. The following is a more detailed explanation of those features.

Based on the literature survey, there are two main features in the phishing email classification area: human behavior and email behavior. Figure 1 shows the behavior feature categories. Human behavior features are defined from the context of the phishing email, which is based on stylistic features (Tariq, 2018), as the email behavior is based on the structure of the phishing email. Human behavioral features are used mainly to determine authorship, gender, personality, and others related to interpreting a text or spoken language. Features included in the human behavioral features include Lexical, Structural, Syntactic, Content-specific, and Idiosyncratic. Those features are categorized as stylistic features that analyze the input data's content to specific steps and categories (Xiujuan et al., 2019).



Source: Ahmad et al. (2022)

FIGURE 1. Behaviour Features Classification

Each part of an email can be extracted as a feature to improve email classification. The structure and content of the overall email are considered the behavior of the email (Gangavarapu, 2020). The email behavior features can be divided into header, body, and URL. The explanation of each feature is as follows:

1. Header Features

Email header contains metadata information of the email such as email sender, subject, time, Number of CCs, whether it is a reply email, and date used as header features. Features extracted from the header can be numeric, Boolean, or categorical according to the type of features extracted.

2. Body Features

Email body features, commonly called content-based features, mainly focus on the context of email. Common content-based features extracted from the email body are total characters, word occurrences, presence of certain words, presence of Javascript, and body text format (HTML format or text format). Another content-based feature used in several kinds of research is the human stylistic category. It includes the writing style, grammar, word choices, tones, spelling errors, and vocabulary used.

3. URL Features

Phishing email attacks mostly contain URLs in the content. The URLs will trap the victim into a fake website, and phishing attacks will begin to occur. The URLs have several characteristics which differ from standard URLs in the general email, for example, the abnormal length of the URL, unknown domain, or content of the URL. The URL features include the structure of the URL, for example, URL length, URL has an IP address, double slash, Number of Dots, HTTP exists in URL, and "@" symbol exists in URL. URL is the most common feature category used

in phishing email classification. It includes all the suspicious URLs with specific criteria in the part of the email, especially in the body (Ahmad et al., 2020).

(Shabudin, 2020) used a combination of several features, including URL features for Phishing Website classification. The URL features extracted in this experiment were URL length, URL using IP address, Presence "@" symbol, and double slash. These features were combined with several website/HTML-related features such as On-mouseover, RightClick, Favicon, and Page Rank. This experiment used several machine learning algorithms such as Random Forest, Multilayer Perceptron (MLP), and Naïve Bayer (NB). The Random Forest classifier produced the best result with 97% accuracy. The result concludes the significance of URL features in identifying Phishing Websites. The URL features have shown promising results and can be used in other phishing data sources such as Phishing Emails.

(Mahdieh, 2019) used several URL features to detect Phishing Websites, including URL Length, Number of Dash in URL, @ exists in URL, HTTP exists in URL, URL contains IP address and the number of all characters in the web-page URL. The research also used other features that have been extracted comes from UCI and Mendeley datasets, such as RightClickDisabled, PopUpWin, FakeLinkInStatusBar, and IframeOrFrame. Three classification algorithms used were Multilayer Perceptron (MLP), Random Forest, and Sequential Minimal Optimization. The best results obtained are the maximum F-measure value of 95% using the Random Forest Classification. In addition, there are three categories of feature sets extracted, including from UCI and Mendeley, that were selected in the feature selection task as universal features. All feature sets included the URL length feature, which shows the significant effects of URL length in phishing website classification.

(Niu, 2017) extracted 23 features from header, body, and URL features to detect phishing emails. The URL features extracted in this study include Presence of IP address, Number of URLs, Number of Dots greater than 3, Presence of "@" symbol, Length of URL, Presence of "-" character, and Number of "HTTP/HTTPS" in URLs more than 1. Niu used two datasets, the Online Phish Corpus (Nazario, 2006) and the Enron CALO project (Kaelbling, 2011). The algorithm used is SVM with the addition of Cuckoo Search (CS) for parameter optimization. The results obtained are very promising, with an accuracy of 91% using CS-SVM as the algorithm. This research also proves that the URL feature helps to classify phishing emails.

The related work has shown a promising result of URL features in identifying phishing websites and emails. Therefore, this paper analyzes the significance of URL features in improving phishing email classification using two datasets with SVM and ANN classifiers.

METHODOLOGY

The research methodology used is based on experimental research. The dataset was processed, and measurable and observable features were used. Various experiments were conducted to obtain satisfying results. Figure 2 shows the flow of the experiment. This research flow is divided into five main phases as follows:

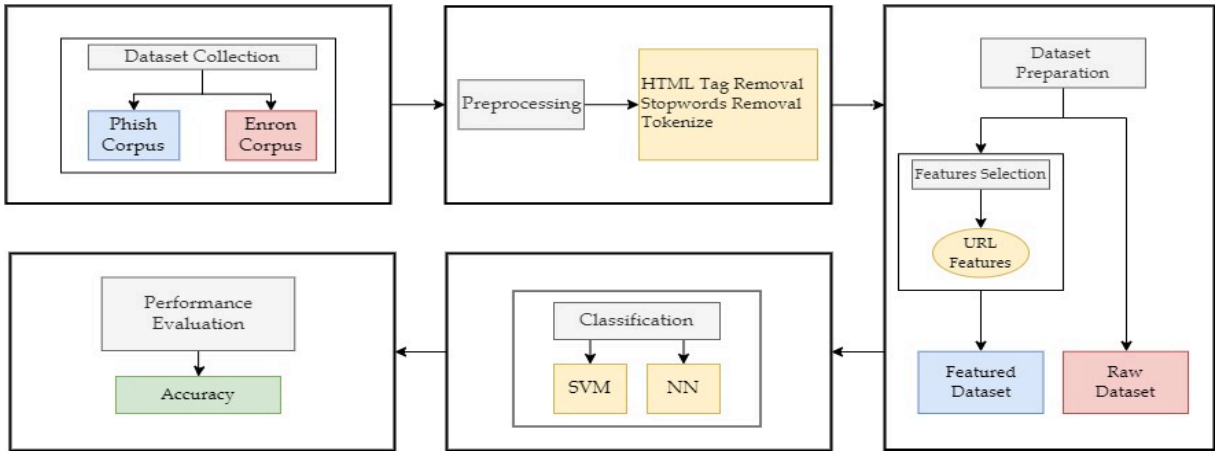


FIGURE 2. Experiment Flow

1. Dataset Collection

There are only a few publicly available phishing email datasets. The most widely used datasets in research related to phishing email classification are Online Phishing Corpus by Jose Nazario (Nazario, 2006) and the Enron Corpus CALO project (Kaelbling, 2011). These corpora have been used by Sumathi (2019), Xiujuan (2019), and Peng (2018). The dataset in this study is divided into two different types of labels/categories, namely phishing as the main label and Ham Email. A phishing email can be created by collecting other researchers' data or combining the email manually from public sources. There are approximately 2000 emails that can be processed/used from the Online Phishing Corpus. However, tedious data cleaning is required, such as deleting empty data rows. Initially, data from Online Phishing Corpus had a "mbox" format which had to be converted into Comma Separated Value (CSV). The data extracted and transformed into CSV files are "From", "To", "Date", "Subject", and "Body" consecutively. There are additional columns, namely Label and Label Number, to determine the type of email, which 1 is for phishing and 0 for ham or non-phishing email. This formatting is done with the help of Pandas Python Library.

The dataset is divided into three parts of 500, 1000, and 1500. The results of the divided data are compared to each other to comprehend the impact of data volume and URL features with carrying dataset volume on the phishing email classification. Moreover, to see how the effect of extracted URL features on datasets with varying amounts of data.

2. Preprocessing

Preprocessing is the process of cleaning data before the data can be used in further steps (Ali & Lailatul, 2016). The preprocessing step includes several processes, the first of which is HTML tag removal for some emails with HTML text format in the body using the BeautifulSoup library, except for the "<a>" tag. The tag will be used for extracting URL features. The following step is stopwords removal and data tokenization, which is done with the NLTK library.

3. Data Preparation and Features Selection

Features selection is an effective task to improve the classification performance or reduce the processing time (Suhaidi et al., (2021), Adel et al., (2019)). The features extracted in this experiment are URL features. The URL features are from "<a>" tags and count the number of links inside the email body using the Beautiful Soup library. There are two conditions considered for extracting the features. The first condition is the detection of the "<a>" tag, and the second is to count the total number of "<a>" tags. The output for each email will be counted

and assigned to new feature variables, namely URL Flag and URL Count. Table 1 shows a detailed explanation of the features extracted.

TABLE 1. URL Features Extracted

Feature	Observed Field	Value	Description
URL	BODY TEXT	URL Flag	Boolean value that represents the presence of URL in an email by detecting <a> tag
		URL Count	A total number of URLs found in the body text

The URL Flag feature is used for initial checking of the body content of the email. If "<a>" and "<a href>" tags, HTTP, and WWW are detected in the content, it will continue with the calculation of the number of URLs that are found, namely the URL Count in the email body. If the criteria are not found, it will be continued to the next email with the same process.

4. Algorithms Implementation

This research used Support Vector Machine (SVM) and Neural Network/Artificial Neural Network algorithms (ANN). Keras library is used to develop and implement ANN. The ANN structure consists of one input layer, one hidden layer, and one output layer. The classification is executed twice based on the features used: the raw dataset without features and with the URL features. Each classification execution will use three different email volumes: 500, 1000, and 1500. The division of the data volumes in each execution is done manually. Each experiment used a data split process for training and testing, with 80% for training data and 20% for testing data. This process is done using "train_test_split" provided by the Sklearn library.

5. Performance Evaluation

The performance is evaluated with several experiments to analyze the best accuracy of the various algorithms used. Figure 2 shows the flow of the experiment.

EXPERIMENT

The classification is executed twice based on the features used: the raw dataset without features and with the URL features. Each classification execution will use three different email volumes: 500, 1000, and 1500. The classifiers in this experiment are SVM and ANN. The process carried out on the two algorithms is as follows; For the SVM algorithm, the data column containing texts is transformed into a vector using CountVectorizer. The data and label columns are assigned to variables: x for the data and y for the label. Next is the data train and test process with an 80:20 ratio. The kernel used for this experiment is Linear since the data used is linearly separable, and it is also one of the most common kernels to be used. The last step is to fit the model with the prepared data. The process in the ANN algorithm is slightly different due to the structure of the ANN, which has several layers. A neural network model is needed to input and train the data. A sequential model containing one dense layer as the input layer, one dense layer as the hidden layer, and one dense layer as the output layer was created for the classification model. Data that already contains features that have been extracted and done with test and train split will be entered into the input layer, which is connected to the next layer. The training process will start at this point, and the output layer will provide one output value, either phishing or non-phishing.

The two classifiers are evaluated by comparing their accuracy. Accuracy is the primary metric for comparing models. It also describes how the model performs across all classes, which is useful when all classes are of equal importance, in this case, Phishing and Non-

Phishing. The entire process for this experiment uses the Python language with several additional libraries, namely Pandas for the data frame, NLTK for the Tokenize process, Beautifulsoup4 for the HTML section, Sklearn for the data preparations, and Keras library for the Neural Network model. Table 2 shows the results of comparing experiments with and without the URL features for phishing email classification.

TABLE 2. Experiment Results

Technique	Accuracy					
	Dataset without URL Features			Dataset with URL Features		
	500	1000	1500	500	1000	1500
SVM	54%	55%	56%	95%	96%	97%
ANN	60%	64%	56%	92%	94%	90%

As shown in Table 1 above, the classifier's accuracy improves significantly by extracting when the URL features are used. It can be seen in the results table that the accuracy of using SVM is better than the standard neural network because the number of datasets used in this experiment is relatively small. The neural network structure, which is the basis of deep learning, requires a large number of datasets so that the learning process can be effective and optimal. This analysis also proves that selecting and using URL features will increase the performance of the classification of phishing emails. The selected feature comes from the URL feature category, namely the Number of URLs used in previous research. From these results, it can also be said that the selection and use of features are essential in helping to get good accuracy results, especially in phishing email classification. The comparison results in the form of a graph can be seen in Figure 3, and the comparison with previous research can be seen in Figure 4:

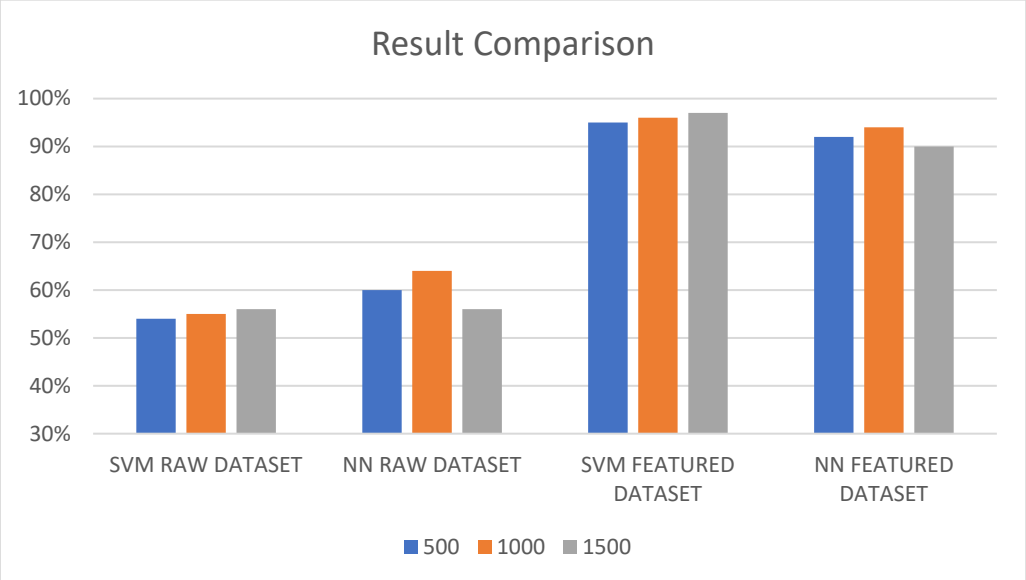


FIGURE 3. Result Comparison on Initial Analysis

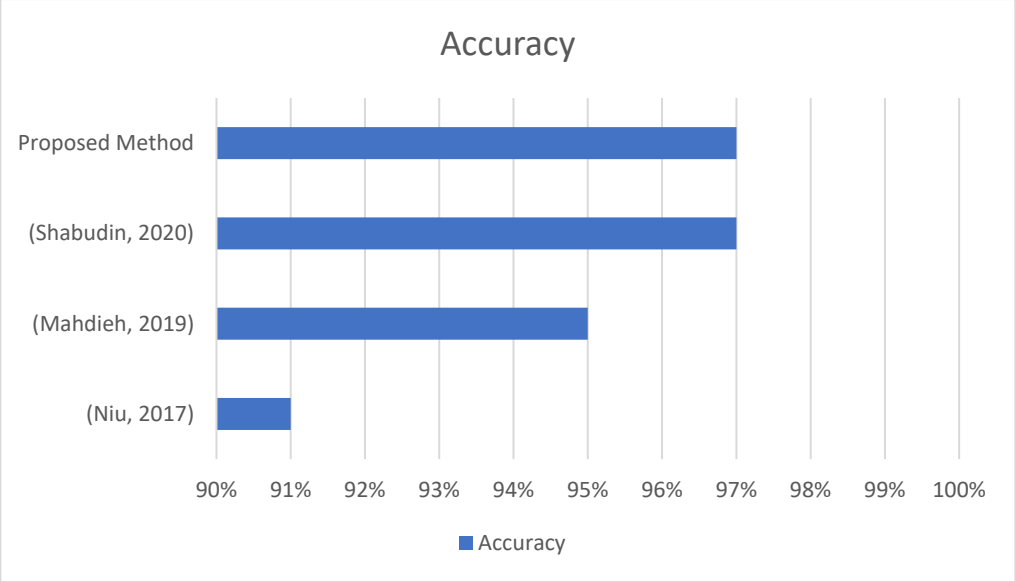


FIGURE 4. Result Comparison

Figure 4 compares the accuracy of this experiment with several previous studies related to phishing using the URL feature. Compared to (Niu, 2017), who extracted 23 features, the proposed method proved to produce better accuracy. The proposed method shows promising results even though there is only one feature category, namely the URL feature selected and used to improve the classification of phishing emails. 97% accuracy is obtained with SVM as the classifier, which is the same result as (Shabudin, 2020) which uses a different dataset, namely Phishing Websites. The results of this experiment also show that the effect of the URL feature has a significant impact on classification in the phishing area, especially phishing email.

CONCLUSION

Selecting a set of features from datasets is an effective way to improve the classification performance or reduce the processing time. In this experiment, the URL feature significantly impacts obtaining good accuracy in classifying phishing emails. The proposed method shows the performance of the URL feature with machine learning with a variety of email data, namely 500, 1000, and 1500. The accuracy results obtained are directly proportional to the large amount of data used; this experiment proved to be very promising, which resulted in an accuracy value of 97% by SVM. These results are better than previous experiments using the same approach. For future work, adding different features and combining them with URL features may improve the result for phishing email classification. The integration between email and human features in phishing email classification has not been fully integrated. This feature integration is expected to improve classification performance by analyzing what features are optimal for integration with URL features. For the classification technique, a different approach can be implemented to get different results that can be analyzed further, such as using a deep learning approach with a combination feature or embedding for feature representation. Therefore, this is a challenge in integrating features selected to improve the performance and the technique used in the phishing email classification.
This study has several limitations; this paper focuses only on identifying the effect of the URL features for phishing email classification using machine learning approaches. The scope can be extended in future reviews.

ACKNOWLEDGEMENT

This research was supported by Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia.

REFERENCES

Bagui, S., D. Nandi, & S. Bagui. 2019. Classifying Phishing Email Using Machine Learning and Deep Learning. Conference: 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). 2019.

Sumathi, K. & Sujatha V., 2019. Deep Learning Based-Phishing Attack Detection. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019.

Kumar, A., Chatterjee, J. M., & Díaz, V. G. "A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing," Int. J. Electr. Comput. Eng., 2020, DOI: 10.11591/ijece.v10i1.pp486-493.

Kumar, S., Faizan, A., Viinikainen, A., & Hamalainen, T. "MLSPD - machine learning based spam and phishing detection," 2018, DOI: 10.1007/978-3-030-04648-4_43.

Mujtaba, G., Shuib, L., Raj, R. G., Majeed, N. & Al-Garadi, M. A. 2017. Email Classification Research Trends: Review and Open Issues. IEEE Access.

Xiujuan, W., Chenxi, Z., Kangfeng, Z., & Haoyang, T., 2019. Detecting Spear-phishing Emails Based on Authentication. IEEE 4th International Conference on Computer and Communication Systems. 2019.

Fang, Y., Zhang, C., Huang, C., Liu, L., & Yang, Y. "Phishing Email Detection Using Improved RCNN Model with Multilevel Vectors and Attention Mechanism," IEEE Access, 2019, DOI: 10.1109/ACCESS.2019.2913705.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., & Shlomo Levitan, L. "Stylistic text classification using functional lexical features," J. Am. Soc. Inf. Sci. Technol., 2007, DOI: 10.1002/asi.20553.

Kaelbling, L., "Enron Email Dataset," CALO Project. 2011 Available: https://www.cs.cmu.edu/~./enron/

Nazario, J., 2006. Online Phishing Corpus. Available: https://monkey.org/~jose/phishing/

Suhaidi, M., Abdul Kadir, R., & Tiun, S. "A REVIEW OF FEATURE EXTRACTION METHODS ON MACHINE LEARNING," J. Inf. Syst. Technol. Manag., 2021, DOI: 10.35631/jistm.622005.

Abdullah, A. S & Mohd, M. "Spear Phishing Simulation in Critical Sector: Telecommunication and Defense Sub-sector," 2019, DOI: 10.1109/ICoCSec47621.2019.8970803.

Ali M. H & Lailatul Q. Z. "Question Classification Using Support Vector Machine and Pattern Matching," Journal of Theoretical and Applied Information Technology, 20th May 2016. Vol.87. No.2. 2016.

Ahmad F. N., Lailatul Q. Z, Saidah S., "Phishing Emails Classification Research Trends: Datasets, Features and Methods", IJAST, vol. 29, no. 04, pp. 6921 -, Jun. 2020.

Adel A., Omar N., Albared M., & Al-Shabi, A. "Feature selection method based on statistics of compound words for Arabic text classification," Int. Arab J. Inf. Technol., 2019.

Iqbal, F., Khan, L. A., Fung, B. C. M. & Debbabi, M. 2010. Email authorship verification for forensic investigation. Proceedings of the ACM Symposium on Applied Computing.

Tariq, M. 2018. Style, stylistics and stylistic analysis: A re-evaluation of the modern-day rhetorics of literary discourse. International Journal of English Research. ISSN: 2455-2186. Volume 4; Issue 2; March 2018; Page No. 46-50

Gangavarapu, T., Jaidhar, C. D. & Chanduka, B. 2020. Applicability of machine learning in spam and phishing email filtering: review and approaches. Artificial Intelligence Review. doi:10.1007/s10462-020-09814-9

Ahmad F. N., Lailatul Q. Z, Saidah S., "Analyzing Email and Human Behavior Features on Phishing Email Classification", E-Proceedings of Seminar on Information Retrieval and Knowledge Management 2022 (Sirkm'22). 2022.

Shabudin, S., Sani, N. S., Ariffin, K. A. Z., & Aliff, M. “Feature selection for phishing website classification,” Int. J. Adv. Comput. Sci. Appl., 2020, doi: 10.14569/IJACSA.2020.0110477.
Peng, T., Harris, I., & Sawa, Y. “Detecting Phishing Attacks Using Natural Language Processing and Machine Learning,” 2018, doi: 10.1109/ICSC.2018.00056.

Ahmad Fadhil Naswir
Lailatul Qadri Zakaria
Saidah Saad
Fakulty of Information Science and Technology
Universiti Kebangsaan Malaysia
afadhilen@gmail.com, lailatul.qadri@ukm.edu.my, saidah@ukm.edu.my