A Trustworthy Swift Weapon to Detect the Phishing URLs by Machine Learning Approaches

Suseta Datta [a], Shibaprasad Sen [a] and Pritam Kundu [a]

[a] *University of Engineering & Management, Kolkata, Newtown, India*

Abstract

Now a days Phishing is a mundane attack on gullible people by making them to disclose their personal information utilizing counterfeit URLs. The main purpose of Phishing URLs Detection by Machine Learning approaches is to give security and safety to the unique information like passwords of personal portals, personal information's and online transactions. In Machine Learning techniques, various approaches are the puissant implements that have been used to grapple against phishing attacks. This paper consists of various Ma-chine Learning approaches which have been utilized for detecting phishing URLs. The best fitted approach has been derived and modified using another ML approach which is giving almost 97% testing accuracy. This paper has shown that the precision, recall, f1-score and training-testing accuracy have been calculated based on the confusion matrix for each applied approach. An interactive and responsive web frame work has been designed for making this project user-friendly. Here, phishing domain characteristics have been explained in details and the features which distinguish these domains from anti-phishing domains. The phishing URLs within the body of the inputs are designed to make it appear that they go to the defraud organization utilizing that organization's logos and other legitimate contents.

Keywords

Natural Language Processing, Security, Interactive and Responsive Web Framework.

CEUR-WS.org/Vol-3244/PAPER_04.pdf

1. Introduction

Security researchers are now concerned about phishing primarily due to the ease with which an authentically fraudulent URL can be forged that resembles legitimate URLs [1]. Phishers utilized the URLs which are visually homogeneous to authentic URLs. Phishing assailments are becoming prosperous because lack of utilizer cognizance [2]. To eschew blacklists assailers uses ingenious techniques to illude users by modifying the URL to appear legitimate [3]. Malefactors are endeavoring to convince online users to reveal passwords, account numbers, convivial security numbers or other personal information [5]. The URLs and their all features will be analyzed for detecting the phishing URLs. To evade extensive losses different authors had proposed to determine characteristic features of phishing emails. These features accommodate as inputs to statistical relegation techniques, which are then trained for identifying phishing URLs [7]. H. Huang et al. [2] suggested structure that determine the phishing use of page section similarity that fails macrocosmic assets spotter token to engender forecast CSS vogue is usually kept as objective pages by phishing pages [10]. This technique is proposed by S. Marchal et al. [2] to separate. On the inspection of genuine site server branch erudition phishing URL is dependent Mustafa Aydin et al. [2] proposed a relegation method to detect the phishing URLs and its URL features and survey subset predicted feature cull methods. Phish storm is a robotic system to detect that can examine in authentic time any URL [8]. Muhemmet Baykara et al. [2] Nominate an application which is kenned as not phishing clone it gives details about the spotting quandary of phishing and the way of spotting phishing URLs [6].

WCNC-2022: Workshop on Computer Networks and Communications, April 22 – 24, 2022, Chennai, India.
EMAIL: shibubiet@gmail.com (Shibaprasad Sen)
ORCID: 0000-0003-4815-6621 (Shibaprasad Sen)
© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

2. Problem Statement and Proposed Solution

As technology perpetuates to grow, phishing techniques commenced to progress expeditiously to be averted by utilizing anti-phishing mechanisms to detect phishing [9]. Phishing becomes a main area of concern for security researchers because it is not arduous to engender the unauthentically spurious URL which looks so proximate to legitimate URL [16]. Major drawback of previous technique is that it can't deter-mine 'zero-hour' phishing URLs attack. The most prevalent technique has utilized and this paper has shown that the best fitted algorithm has predicted the result correctly with the best accuracy. To develop best-fit model, programs are divided into their felicitous domains and subsequently categorized as phishing or legitimate. A classifier and a regressor method have been used. The regressor model has been giving the best accuracy than the classifier. Confusion matrix has been calculated for determining the best fitted algorithm based on precision, recall and f1-score. Page content inspection had been utilized by some strategies to surmount the erroneous negative quandaries and complement the susceptibilities of the stale lists. A toolkit has been developed to utilize as a platform for all the users. It will be acclimated to detect a given URL either phishing or not. The URL is engendered for all users; hence it must be facile to operate with and no utilizer should face any arduousness while making its use. The different features-based dataset makes up to be taken in the meantime of determining a URL as phishing [19]. The features for detecting and relegating of phishing URLs are as follows: Hypertext Markup Language and JavaScript based, Abnormal based, Address bar based, Domain based [9]. Machine Learning strategies, Natural Language processes, and other applied approaches are de-scribed further down in this paper.

3. Dataset

The dataset for detecting phishing sites has been taken from https://www.kaggle.com/taruntiwarihp/phishing-site-urls. The raw dataset contains 5,49,346 samples where 72% is for legitimate URLs and 28% for phishing URLs. The dataset consists of 5,07,195 unique samples out of total. Legitimate URLs have been labelled as 'good' and phishing URLs have been assigned as 'bad'. In all, 2 instances have been utilized and there is none of which have a null value. The dataset embedding procedure has been used according to the natural language processing methods. Vectorization process has been used to transform the stemmed words into a vector form. This vector form has been engendered from the tokenized and stemmed dataset and the name of the URLs has been utilized as input after developing the model. 75% of the dataset has been utilized for training, while the remaining 25% has been utilized pristinely for testing.

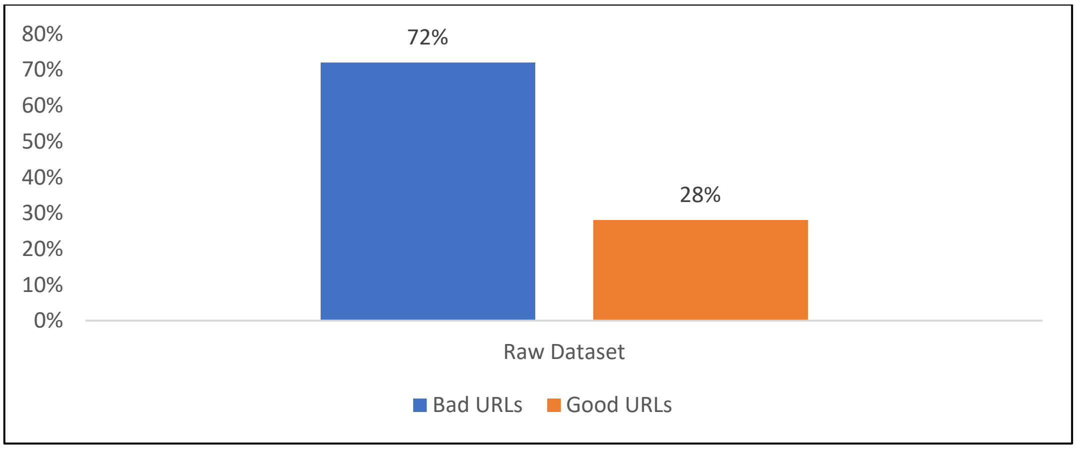

Figure 1: This figure consists of the numbers of some good and bad URLs which are used in the dataset

4. Methodology

4.1. Tokenization, Stemming and Joining Root Words

Tokenization is the process of turning a paramount piece of data into a desultory string of characters that has no consequential value. The tokenization process has been utilized to break a URL which is given as a string and then the given URL has been broken into some tokens. These tokens have been assigned as consequential value. The stemming process has been used for engendering morphological variants of these generated tokens. After generating the root words, 3 instances have been created to store the tokens and the root words in the raw dataset. These 3 instances are 'Tokenized_text', 'Stemmed_text', and 'Sent_text'. In the below table, the processing time of tokenizing, stemming and storing have shown. Please, check Table 1.

Table 1: Processing time of Tokenizing, Stemming and Embedding

Tokenizing Time	Stemming Time	Joining Time
4.609520400000001 sec	96.10180020000001 sec	0.4269779999999912 sec

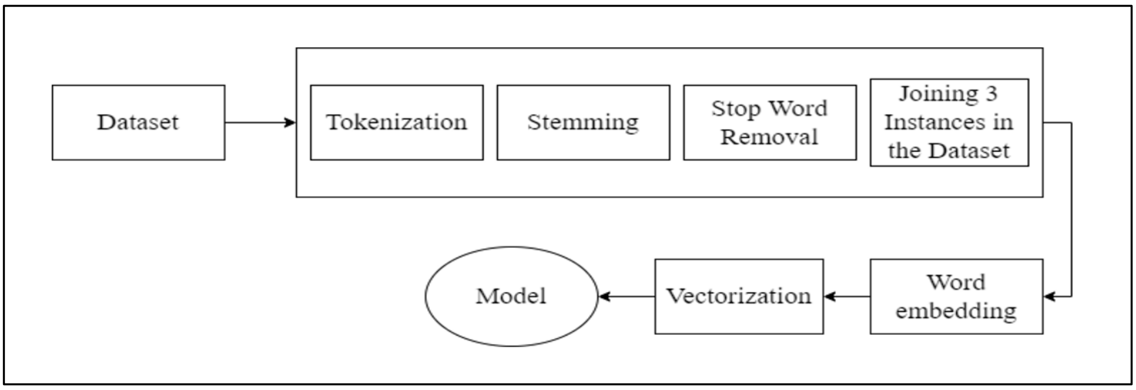

Figure 2: The process of Tokenization, Stemming and Root Words Extraction

4.2. Embedding Root Words

The extracted root words of both good and bad URLs have been embedded through word cloud. This word cloud visualization has been utilized for showing the root words. In the below figures, the output of both good and bad URLs has been shown.

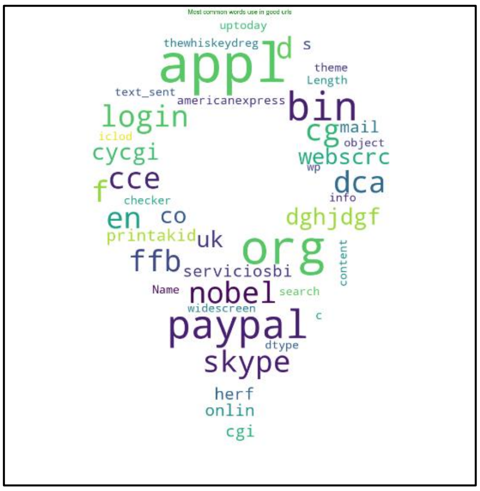

Figure 3: This figure consists of root words of Good URLs

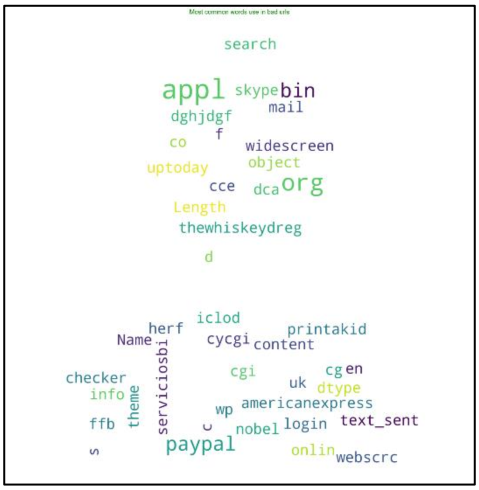

Figure 4: This figure consists of root words of Bad URLs

4.3. Web Driver Automation for Hyperlink Extraction

Web driver automation tool has been used for mechanical testing of a sample of phishing URL. Instead of web browser, the web driver has been used for automatic testing. The relevant hyperlinks of this tested phishing URL has been extracted and plotted into a frame by feeding.

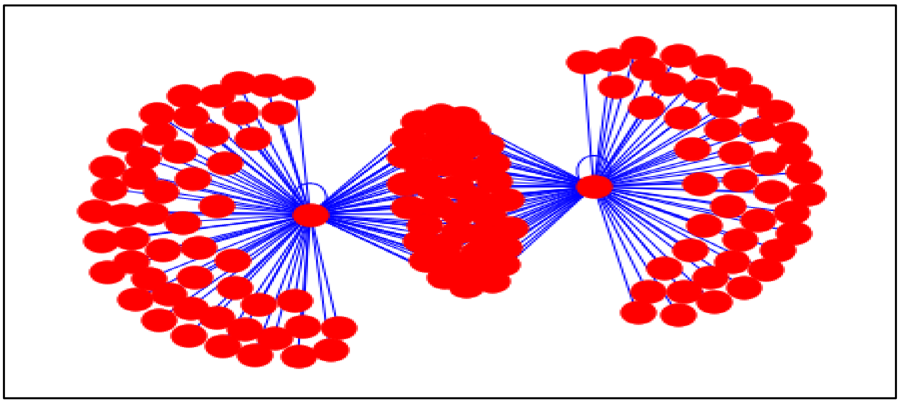

Figure 5: This figure consists of the Data Frame of Internal Links for a phishing sample

4.4. Vectorization

Vectorization process has been utilized to transform a collection of text to a vector of token counts. This allows to conduct the cross validations for training and testing sets. Now the label and the vector form have been used in the process of splitting. These two parameters have been used for the model creation. Mainly two algorithms have used for creating the model. Further discussion has discussed in the result analysis part.

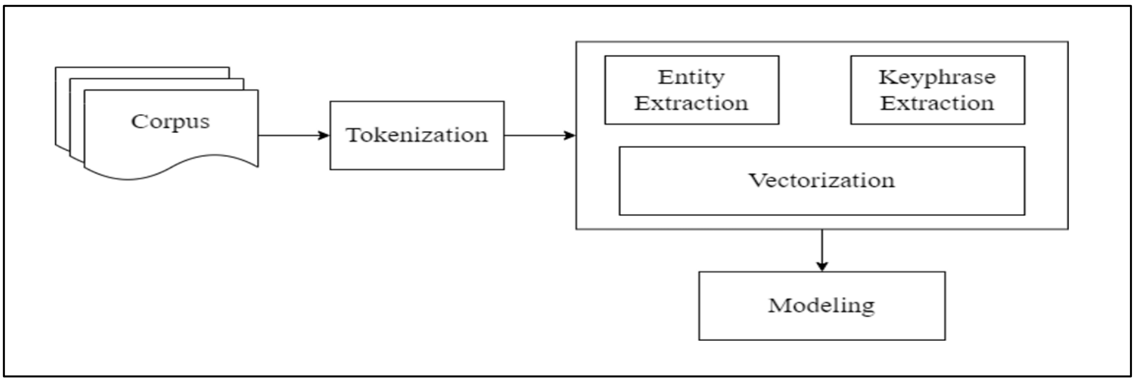

Figure 6: This figure shows the process of Vectorization

5. Result Analysis

Logistic Regression and Multinomial Naïve algorithm are the probabilistic learning techniques that is mostly utilized in Natural language processing. The LR and MNB algorithms have been applied on the selected training dataset. 75% of total dataset has been already assigned for training purpose in dataset pre-processing. The training dataset has been fitted to both classifier and regression. The previous natural language processing methods are carried out with the avail of their respective classifier class and regression class. To expect the test state result, a confusion matrix has been plotted for each algorithm. To evaluate the accuracy, the confusion matrix (please, check Table 1.) has been utilized. The logistic regression (LR) has been giving 96.35% testing accuracy and the multinomial naïve bayes (MNB) algorithm has been giving 95.79% testing accuracy. Following that, the values for Actual Good - Predicted Good (True Positive), Actual Bad – Predicted Good (False Positive), Actual Good - Predicted Bad (False Negative), Actual Bad - Predicted Bad (True Negative) have been measured. Based on the value of these parameters, precision, recall, f1-score and training-testing accuracy have been calculated accordingly for each algorithm (please, check Table 1.). Scikit-Learn pipelining algorithm is applied on the best fitted algorithm. Now again the parameters of splitting dataset have been changed. The vectorized instance has been changed and the name of the URLs instance has been placed over it. The train set has been fitted as per the previous process. This technique has been giving 96.58% testing accuracy and the confusion matrix has been created according to those predicted and actual parameters. The final model has been dumped into a pickle file and load-ed it in the responsive and interactive web frame work. This model will work as a product key in the background of interface.

Finally, the study has been committed the concept for further detection of phishing URLs strategies. The Logistic Regression has performed based in the terms of ACC, TPR/FPR, PPV and F1-Score when applying machine learning approaches to identify given URLs correctly. Multinomial Naïve Bayes algorithm has performed very well but given a lower accuracy than LR based on the confusion matrix. The Logistic Regression algorithm has provided the best accuracy during fitting the training dataset. However, the Logistic Regression has proved to be the most accurate in the end with 96.35% testing accuracy, the accuracy of pipeline is 96.58%. This most appropriate approach is pretty equal to this most exact value of LR.

Table 2: Parameters of Confusion Matrix

Actual-Predicted	Predicted Good	Predicted Bad
Actual Good	True Positive	False Negative
Actual Bad	False Positive	True Negative

Table 3: Final Result of Precision, Recall and F1-Score for Both Algorithms

Metrics	Formula	Result of Algorithms		
		LR	MNB	Pipelining with LR
Precision	PPV = TP / (TP+FP)	98.78%	97.53%	98.77%
Recall	TPR = TP / (TP + FN)	96.25%	96.65%	96.54%
F1-Score	F1 = 2TP / (2TP + FP + FN)	97.50%	97.09%	97.64%

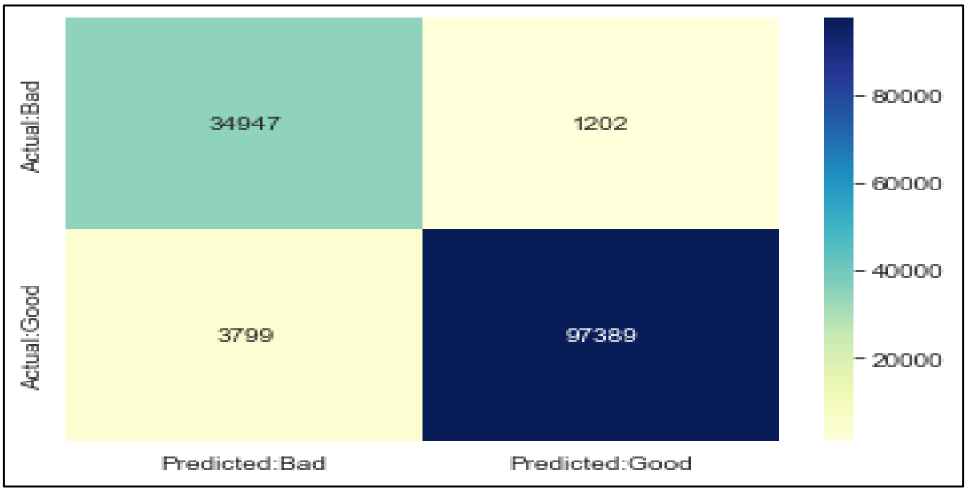

Figure 7: Confusion Matrix of Logistic Regression Algorithm

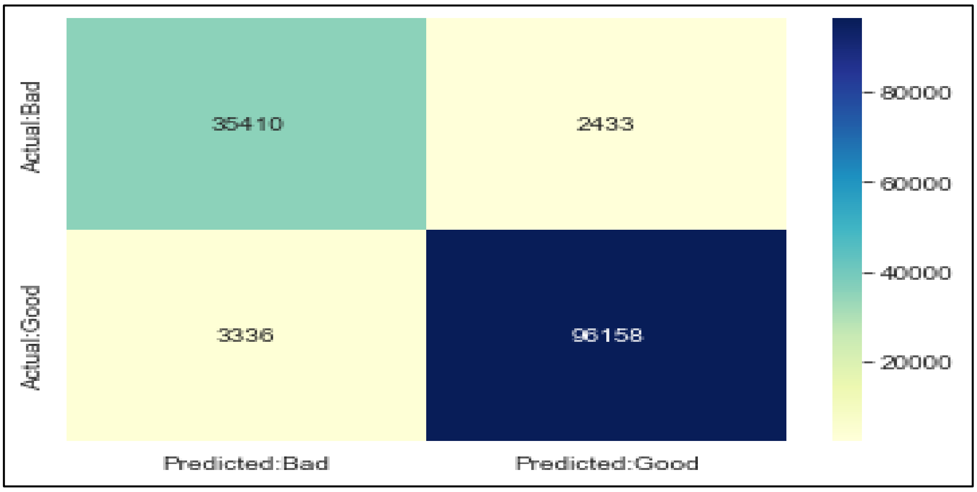

Figure 8: Confusion Matrix of Multinomial Naïve Bayes Algorithm

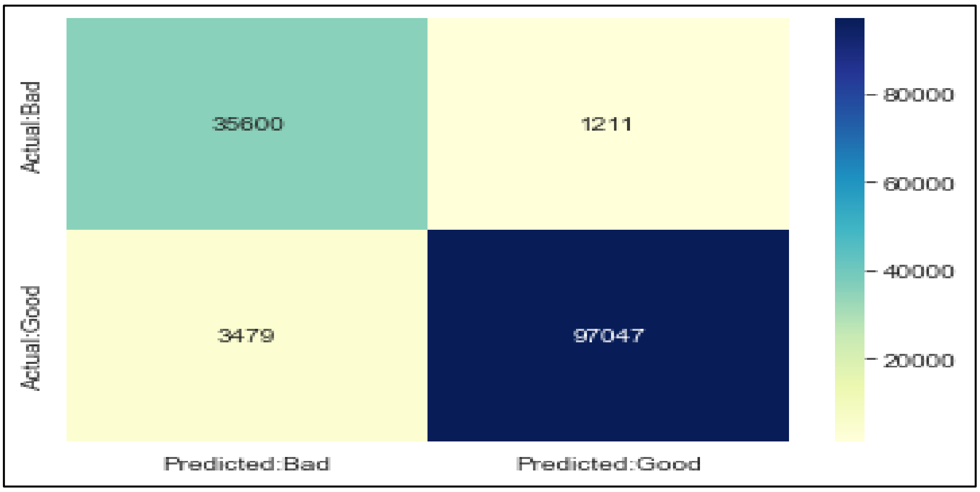

Figure 9: Confusion Matrix of Scikit-Learn Pipelining Algorithm

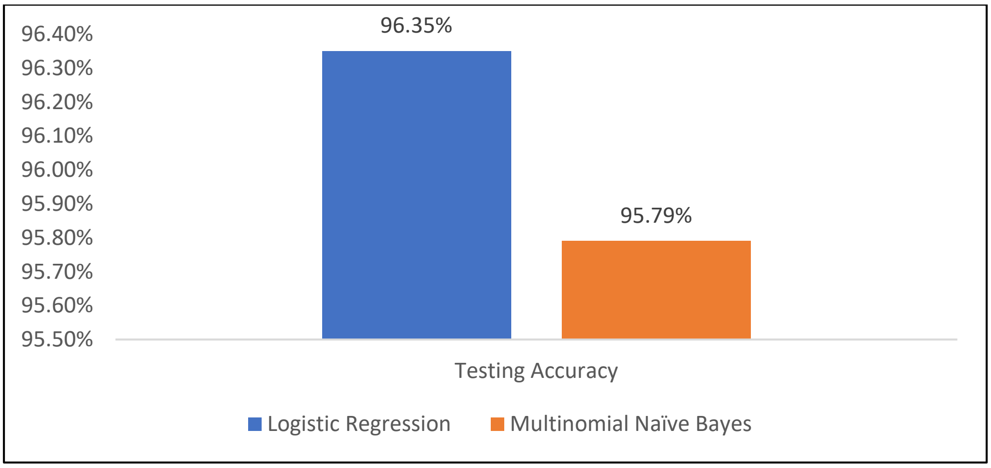

Figure 10: Comparison of Testing Accuracy for Both Approaches

6. Conclusion and Future Scope

Thus, to summarize, the model had been visually perceived how phishing is a sizably voluminous threat to the security and safety of the web and how phishing detection is a paramount quandary domain. The model had been tested two Machine Learning approaches on the Phishing URLs Dataset and calculated their results. Then the model had been culled the best algorithm predicated on its performance and built a Chrome Driver extension for detecting phishing URLs. The model had been detected phishing URLs utilizing Logistic Regression with and precision of almost ~97%.

This paper aims to enhance detection technique to detect phishing URLs utilizing Machine Learning technology. In future, the model had been intended with Random Forest algorithm and black list method to build the phishing detection system as a scalable web accommodation which will incorporate online learning so that incipient phishing attack patterns can facilely be learned and amend the precision of our models.

7. Acknowledgements

The co-author would relish to express his gratitude to Dr. Shibaprasad Sen, Department of Computer Science and Technology and Suseta Datta, Department of Computer Science and Engineering, University of Engineering & Management, Kolkata, for their insightful comments and for providing me with all pertinent details and information about the paper in a timely manner.

8. References

[1] Buber, Ebubekir, Banu Diri, and Ozgur Koray Sahingoz. "NLP based phishing attack detection from URLs." International Conference on Intelligent Systems Design and Applications. Springer, Cham, 2017.

[2] Sahoo, Doyen, Chenghao Liu, and Steven CH Hoi. "Malicious URL detection using machine learning: A survey." arXiv preprint arXiv:1701.07179 (2017).

[3] Lemley, Joe, Shabab Bazrafkan, and Peter Corcoran. "Deep Learning for Consumer Devices and Services: Pushing the limits for machine learning, artificial intelligence, and computer vision." IEEE Consumer Electronics Magazine 6.2 (2017): 48-56.

[4] Ghafir, Ibrahim, et al. "Detection of advanced persistent threat using machine-learning correlation analysis." Future Generation Computer Systems 89 (2018): 349-359.

[5] Gandotra, Ekta, and Deepak Gupta. "An efficient approach for phishing detection using machine learning." Multimedia Security. Springer, Singapore, 2021. 239-253.

[6] Sahingoz, Ozgur Koray, et al. "Machine learning based phishing detection from URLs." Expert Systems with Applications 117 (2019): 345-357.

[7] Gualberto, Eder S., et al. "From feature engineering and topics models to enhanced prediction rates in phishing detection." Ieee Access 8 (2020): 76368-76385.

[8] Lee, Minyoung, and Eunil Park. "Real-time Korean voice phishing detection based on machine learning approaches." Journal of Ambient Intelligence and Humanized Computing (2021): 1-12.

[9] Yadollahi, Mohammad Mehdi, et al. "An adaptive machine learning based approach for phishing detection using hybrid features." 2019 5th International Conference on Web Research (ICWR). IEEE, 2019.

[10] Gupta, Brij B., et al. "A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment." Computer Communications 175 (2021): 47-57.

[11] Kiruthiga, R., and D. Akila. "Phishing websites detection using machine learning." International Journal of Recent Technology and Engineering 8.2 (2019): 111-114.

[12] Chiew, Kang Leng, et al. "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system." Information Sciences 484 (2019): 153-166.

[13] Jain, Ankit Kumar, and Brij B. Gupta. "A machine learning based approach for phishing detection using hyperlinks information." Journal of Ambient Intelligence and Humanized Computing 10.5 (2019): 2015-2028.

[14] Jain, Ankit Kumar, and Brij B. Gupta. "Towards detection of phishing websites on client-side using machine learning based approach." Telecommunication Systems 68.4 (2018): 687-700.

[15] Peng, Tianrui, Ian Harris, and Yuki Sawa. "Detecting phishing attacks using natural language processing and machine learning." 2018 IEEE 12th international conference on semantic computing (icsc). IEEE, 2018.

[16] Abdelhamid, Neda, Fadi Thabtah, and Hussein Abdel-jaber. "Phishing detection: A recent intelligent machine learning comparison based on models content and features." 2017 IEEE international conference on intelligence and security informatics (ISI). IEEE, 2017.

[17] Ubing, Alyssa Anne, et al. "Phishing website detection: an improved accuracy through feature selection and ensemble learning." International Journal of Advanced Computer Science and Applications 10.1 (2019): 252-257.

[18] Kumar, Abhishek, Jyotir Moy Chatterjee, and Vicente García Díaz. "A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing." International Journal of Electrical and Computer Engineering 10.1 (2020): 486.

[19] Zamir, Ammara, et al. "Phishing web site detection using diverse machine learning algorithms." The Electronic Library (2020).

[20] Gharge, Sagar, and Manik Chavan. "An integrated approach for malicious tweets detection using NLP." 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, 2017.

[21] Kulkarni, Arun D., and Leonard L. Brown III. "Phishing websites detection using machine learning." (2019).

[22] Alswailem, Amani, et al. "Detecting phishing websites using machine learning." 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS). IEEE, 2019.

[23] Lakshmanarao, A., P. Surya Prabhakara Rao, and MM Bala Krishna. "Phishing website detection using novel machine learning fusion approach." 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). IEEE, 2021.

[24] Wu, Che-Yu, Cheng-Chung Kuo, and Chu-Sing Yang. "A phishing detection system based on machine learning." 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA). IEEE, 2019.

[25] Alam, Mohammad Nazmul, et al. "Phishing attacks detection using machine learning approach." 2020 third international conference on smart systems and inventive technology (ICSSIT). IEEE, 2020.

[26] Kumar, G. Ravi, S. Gunasekaran, and Vignesh AS. "URL Phishing Data Analysis and Detecting Phishing Attacks using Machine Learning in NLP." International Journal of Engineering Applied Sciences and Technology (IJEAST) 3.8 (2018): 70-75.