Information Extraction from Security-Related Datasets

S. Seljan*, N. Tolj* and I. Dunđer*
* Faculty of Humanities and Social Sciences, University of Zagreb, Department of Information and Communication Sciences, Zagreb, Croatia
sanja.seljan@ffzg.hr, ntolj@ffzg.hr, ivandunder@gmail.com

***Abstract* - There are various approaches to executing security breaches which are nowadays massively occurring in electronic communication environments, and phishing attacks are one of the most applied ones. A vast majority of phishing attacks are initiated using electronic messages, which attackers utilize to direct users to harmful or fake websites, to infect computers or to obtain personal or sensitive data for malicious purposes. Consequently, it is necessary to identify phishing messages in order to provide suitable user protection. Research and numerous studies have included machine learning algorithms and techniques from the field of artificial intelligence which predominantly depend on language-specific datasets and characteristics of phishing messages, and which have demonstrated to be effective for extracting critical information and for data-driven decision making. However, phishing datasets exist mainly for the English language. The aim of this paper is to present an information extraction pipeline that encompasses phases, such as corpus pre-processing, generating predictions of phishing messages using selected machine learning algorithms, along with a basic analysis, confusion matrices and evaluation scores for Croatian phishing messages. This type of key information can be used for teaching in higher education, e.g. in security-related courses or subjects that deal with artificial intelligence, machine learning, big data analysis, computational linguistics etc. This is essential as it can provide deeper insights into phishing attack strategies and potential countermeasures.**

Keywords - information extraction; machine learning; corpus analysis; security datasets; information security; information and communication sciences

I. INTRODUCTION

In a phishing attack, which is a form of social engineering [1], an attacker assumes the identity of a trustworthy party in order to trick a victim into exposing sensitive data, private or financial information [2].

The effort to identify and prevent phishing attacks on individuals and companies is known as phishing detection. Phishing detection is a vital component of information security, and without adequate identification and prevention procedures, a phishing attack might result in data breaches, financial loss and reputational harm [3].

Protecting against these attacks requires setting up strong security measures and training people in how to spot and deal with phishing. However, identifying phishing attacks is not easy, since there are numerous challenges brought on by linguistic complexity and diversity, as well as technical and technological issues.

Moreover, to make phishing messages seem very authentic, attackers also known as "phishers" [4] frequently use sophisticated terminology and techniques, such as misusing official branding and logos of reputable companies and other organizations [5]. In order to further boost the persuasiveness of their messages, attackers also tend to imitate the language, style and tone of a known sender.

This makes it harder for automated security systems to identify attackers and to discern between phishing and legitimate messages [6], especially if they lack the linguistic knowledge or contextual awareness needed in order to properly examine the messages' content.

When it comes to countermeasures, machine learning can be an effective tool for detecting phishing messages, as it enables automatic identification and classification [7]. Machine learning algorithms can accurately determine the difference between phishing and authentic messages [4], especially when trained with domain-specific data and after applying model fine-tuning.

In order to assess and prevent phishing attacks, special datasets, so-called digital corpora, are needed for training machine learning models [8]. The quality and volume of datasets have a significant impact on the performance of security systems that are based on such models.

Furthermore, a dataset that truly reflects the variety of popular phishing methods and strategies is essential for phishing detection [9]. That means that such datasets should contain large amounts of phishing e-mails, texts, and web pages, along with examples of authentic messages to serve as a comparison. It is important that the dataset accurately reflects real examples, so that the model has enough information and context in which it will be applied [10].

Nevertheless, machine learning is not an infallible method for detecting phishing. Machine learning models are susceptible to errors [11], particularly when the training dataset is not diverse enough or when the model is not adjusted correctly.

Therefore, it is crucial to have a dataset that is fairly large for the model to generalize well to observations that were not seen during model training. In order to prevent or reduce bias toward one output class (also known as label)

during model training, the dataset should be balanced, meaning that it should contain an equal number of phishing and non-phishing messages. Additionally, it should be tagged by domain specialists or security experts who are knowledgeable about phishing strategies. This will increase the validity of the labels and, consequently, the quality of the corpus.

The paper is organized in the following way. In the Introduction section the authors present a vital topic in information security, i.e. phishing and potential countermeasures. The second section states the motivation for conducting this research, whereas the third section presents related work and relevant research on datasets and machine learning algorithms for detecting phishing messages. The fourth section deals with ensemble learning methods, which is followed by the research section which presents all results and discusses important findings. Finally, in the last section conclusions are given with regard to research outcomes.

II. MOTIVATION

Information extraction deals with applying natural language processing methods to automatically extract essential details from text documents [12]. These methods are language dependent and focused on the domain of the text. Machine learning, on the other hand, is fast and language independent, and can been used in order to facilitate the process of information extraction. Machine learning algorithms have proven to be effective for extracting critical information and making data-driven decisions.

Nevertheless, they primarily depend on language-specific and high-quality datasets, and characteristics of phishing messages that need to be representative, diverse and labeled in a correct way. Model performance directly depends on the features and properties of the dataset.

However, Croatian-English phishing datasets and recent analyses are lacking, and this very absence of studies was the main motivation behind this research. New insights could be incorporated into various courses in higher education, such as natural language processing, machine learning, artificial intelligence, data analysis, information security, computational linguistics, statistics etc. This is crucial since it can offer more in-depth information about phishing attack techniques and potential responses.

Depending on the perspective of the problem, students could be taught how to evaluate attackers and assess their profiles, how to distinguish different phishing strategies, and how to deal with the principles of phishing techniques and potential countermeasures.

III. RELATED WORK

One research dealt with detecting phishing websites by employing a novel dataset, which contains 5,000 legitimate and 5,000 phishing websites. In order to obtain the best results, various machine learning algorithms, such as Random Forest algorithm and Multilayer Perceptron were explored [13].

Ref. [14] proposed a system that uses the Support Vector Machine algorithm and Naïve Bayes to train predictive models on a 15-dimensional feature set. The experiments were based on datasets consisting of 2541 phishing instances and 2500 benign instances. Using 10-fold cross-validation, experimental results showed 0.04% false positives and 99.96% accuracy for both predictive models.

Another paper also used Random Forest classification and the Support Vector Machine algorithm in order to achieve a 99.87% accuracy score on a dataset that was comprised of 1605 e-mails (1191 phishing and 414 safe e-mails) [15].

In a recent research various types of phishing attacks, such as e-mail phishing, instant messaging phishing, smishing, bulk phishing, spear phishing, whaling, vishing and pharming were examined [4]. For instance, phishing in general is considered as a type of cyberattack in which a hacker tries to trick people into disclosing private or important information by impersonating a trustworthy organization or user, whereas, e.g. spear phishing is a highly specialized form of phishing in which a malicious e-mail is specifically intended for a particular person, business, or group of people. This paper also proposed machine learning algorithms suitable for phishing detection, such as Decision Trees, Random Forest, Support Vector Machine, k-Nearest Neighbors etc.

Another study concentrated on the appropriateness of online services for machine learning of models for predicting classification outcomes, and provided a method that could be utilized for identifying phishing messages [16].

One paper discussed the absence of relevant datasets for phishing detection. Therefore, a specific dataset was constructed by following a set of proposed guidelines, and was used afterwards to assess the level of effectiveness in phishing detection systems based on the Random Forest classifier algorithm [17].

Ref. [18] proposed a method that concentrates on ULRs, extracts features, lowers the dimensionality of the problem, and, when combined with a Support Vector Machine classifier, showed high efficacy in the detection of phishing.

IV. ENSEMBLE LEARNING METHODS

Ensemble learning methods are used in machine learning to combine predictions from multiple models in order to improve predictive performance. There are three main types of ensemble learning methods: Bagging (Bootstrap Aggregation), Stacking and Boosting [19, 20].

Bagging or Bootstrap Aggregation uses multiple Decision Trees on different samples of the same dataset, and then averages predictions [21]. Bagging employs different members of an ensemble group by using variations of training data. It uses replacement, meaning that once the instance (row) is selected, it is returned, and can be selected again. This technique of bootstrap resampling is often used in statistics on small datasets, where many training datasets can be prepared to achieve an overall better estimation [21].

An example of Bagging is the popular Random Forest algorithm, which is used for solving classification and regression problems. This algorithm contains numerous Decision Trees on different subsets of a dataset, and therefore "takes the average to increase the predictive accuracy of that dataset" [22]. It makes predictions depending on each tree, and makes a final decision based on the maximum votes of predictions, which is especially useful when individual trees are not correlated among themselves. By using multiple Decision Trees, it overcomes the problem of overfitting. On the other hand, the accuracy of Random Forest depends on the larger variety of trees [23].

According to [24], advantages of the Random Forest algorithm are automatization of lost values in data and, overcoming overfitting issues with efficiency in handling large datasets. However, disadvantages are in the context of more computing and more resources that are needed for efficient results. Random Forest requires more time for training as well since it relies on many Decision Trees.

Stacking uses different models (e.g. Linear Regression for regression tasks, or Logistic Regression for binary classification) on the same dataset and another model to learn how to make predictions. Ensemble members are referred to as level-0 model, and the model that is used to combine predictions as level-1 model [19].

Boosting is an ensemble learning method that aims to change the training data in order to focus on examples that were erroneously predicted during previous fitting of models on the training dataset [19]. Is focused on instances that were misclassified by previous classifiers. It is called Adaptive Boosting since weights are re-assigned to each instance, i.e. each row, having higher weights assigned to incorrectly classified instances. Instances are weighted to indicate the needed amount of focus on the dataset during training. The output is given as a weighted average of predictions.

Popular examples of the boosting method are Adaptive Boosting (AdaBoost), Stochastic Gradient Boosting (XGBoost and similar) and Gradient Boosting Machines [19]. AdaBoost is an adaptive machine learning algorithm where weights are reassigned to each instance, i.e. each row which is misclassified. Incorrectly classified instances have higher weights, and are widely used for imbalanced data. While Decision Trees in the Random Forest algorithm have equal contributions, in AdaBoost they have different contributions.

V. Research, Results and Discussion

The objective of this paper is to present a pipeline for information extraction that includes steps like corpus pre-processing, generating predictions for phishing messages using specific machine learning algorithms, performing an accuracy-based analysis, creating confusion matrices and evaluating scores for Croatian phishing messages.

A. Dataset Characteristics

The dataset used in this research consists of 550 phishing e-mails that arrived at private e-mail addresses during 2022. The entire dataset was pre-processed with special focus on lowercasing, removing accents and numbers, text filtering with help of a specially crafted stop word list, discarding e-mail attachments etc. All e-mails were manually checked and annotated. The dataset was then divided into two subsets:

- **Set A**: 275 e-mails originally written in Croatian,
- **Set B**: 275 e-mails originally written in English that were automatically translated into Croatian by the web service Google Translate.

Examples of phishing e-mails from both sets are given below:

Set A:

- "Obavijest: Vaša kompenzacijska bankovna kartica na bankomatu u vrijednosti od 1.500.000,00 dolara registrirana je kod voditelja kurira DHL-a g. Marka Adjovija. Radi trenutne isporuke kontaktirajte ga putem e-pošte (dhlcourierexpressbj1@outlook.com) za više informacija o tome kako ćete je zatražiti."
- „Draga moja, jesi li primila poruku koju sam ti poslao? Pozdrav, Jerry Ngessan"

Set B:

- "Poštovani, Vaša verzija e-pošte je zastarjela, nenadogradnja na najnoviju verziju ffzg.hr 7.1 sada će dovesti do trajnog zatvaranja računa. Sukladno odredbi 17.9 Uvjeta, ffzg.hr može u bilo kojem trenutku prekinuti svoje usluge za račune. Za nadogradnju kliknite ovdje i odmah ponovno potvrdite svoj račun. Hvala, ffzg.hr Mail Team"
- „Sada sam spreman za poziv, mogu li vas nazvati? Odgovorite na -> angelodicosta@gmail.com. Možemo li se naći u kafiću?"

Each phishing e-mail was manually marked with one of the following categories (also known as labels or classes):

- Finances – commercial phishing e-mails with the aim to scam users for monetary benefit,
- Health – offers and promises life changing products that do not exist, in order to acquire valuable personal information from victims,
- Adult content – e-mails containing erotic and raunchy content, or having allusions to pornographic topics, nudity, explicit sexual material or violence,
- Short communication – short e-mails containing brief and generic greetings to encourage victims to continue a conversation in which they provide sensitive information to the attacker.

Distribution of categories in the dataset is presented in Table I, showing the proportions of different classes (labels) of e-mails. In both subsets, the test set equals to 10% of the training set. The most represented categories in the training and test sets are "Adult content" and "Finances". In almost all cases, the least represented category is "Health".

Although divided into four different categories, all phishing e-mails are related to financial matters, which might contribute to later-stage misclassification. E-mails that do not deal with financial topics are aimed at collecting data through a friendly conversation, with the goal of obtaining financial benefit through fraud.

TABLE I. DATASET CHARACTERISTICS

	Set A		**Set B**	
	Training set	**Test set**	**Training set**	**Test set**
No. e-mails	250	25	250	25
Finances	89	11	90	10
Health	8	2	17	5
Adult content	94	8	103	9
Short communication	59	4	40	1

A large amount of phishing e-mails focuses on the financial aspect of victims, asks for valuable data or directly for money, employs blackmailing strategies, presents inappropriate services, asks for charity donations or promises lottery wins. Such e-mails frequently offer victims money in exchange for their data.

B. Prediction of Categories

In order to perform prediction of all categories, both subsets were split in the ratio of 90:10 (proportion of training and test set). Training was performed using the following classifier algorithms [4, 19]: Multivariable Logistic Regression (LR), which is suitable for multiple variables (categories), Random Forest (RF), k-Nearest Neighbors (kNN), Naïve Bayes (NB) and AdaBoost (AB).

These algorithms were chosen on purpose in order to detect the algorithm and approach with the highest efficacy, i.e. percentage of correct predictions for this task.

Table II shows predictions achieved by all algorithms, and across all four categories. When comparing the two subsets, Set A, originally written in Croatian, has slightly better predicted scores than Set B, which was automatically translated from English into Croatian.

TABLE II. PREDICTED ACCURACY RESULTS ACROSS ALL CATEGORIES

	Set A		**Set B**	
Algorithm	Accuracy	F1	Accuracy	F1
LR	0.684	0.686	0.680	0.675
RF	**0.776**	**0.767**	**0.732**	**0.711**
kNN	0.704	0.691	0.648	0.613
NB	0.460	0.454	0.448	0.453
AB	**0.748**	**0.743**	**0.760**	**0.753**

Remark: more is better.

When comparing prediction outcomes, accuracy and F1 measures achieved similar results. The highest predictions were achieved by the algorithms Random Forest and AdaBoost, ranging from ca. 71% up to ca. 77% for accuracy and F1 scores.

The F1 score is based on the balance between the metrics of precision and recall, i.e. it is the harmonic mean of precision and recall [25], usually used for imbalanced datasets.

Precision identifies all correctly identified cases, divided by all positives (true and false positives). Recall is the measure that counts true positives divided by true positives and false negatives.

F1 and classification accuracy as prediction measures obtained by Random Forest and AdaBoost are compared with manual testing results.

C. Confusion Matrix

The confusion matrix is a predictive analytics tool used widely in machine learning and in various classification tasks [26]. It is a summarized *NxN* table that presents the number of correct and incorrect predictions in a classification task, and therefore estimates the performance of a classification algorithm.

The confusion matrices determine the differences between the actual and the predicted accuracy across all four categories. The confusion matrix takes the total number of all training datasets in all categories and predicts accuracy.

Fig. 1 and Fig 2. show the confusion matrices for the Random Forest and AdaBoost algorithms for each of the four categories. Both algorithms predicted similar order of categories in both datasets.

Interestingly, both algorithms predicted that “Health” in Set A will gain a 100% score, followed by close scores for the categories “Adult content” and “Finances”. For Set B, both algorithms predicted the same ranking of categories: “Adult content”, followed by “Finances”, then “Short communication”, and finally “Health”. This ranking is contrary to the rankings in Set A.

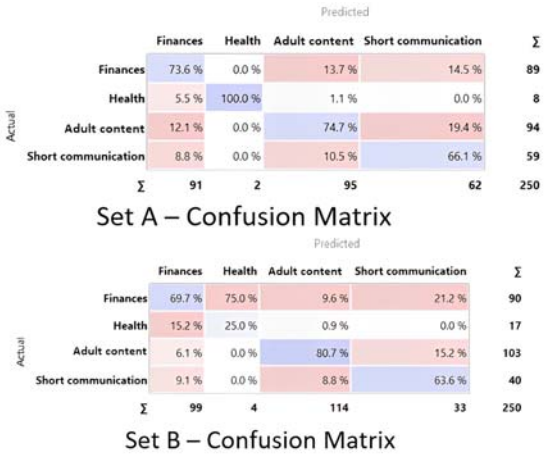

Figure 1. Confusion Matrices per Category for Random Forest

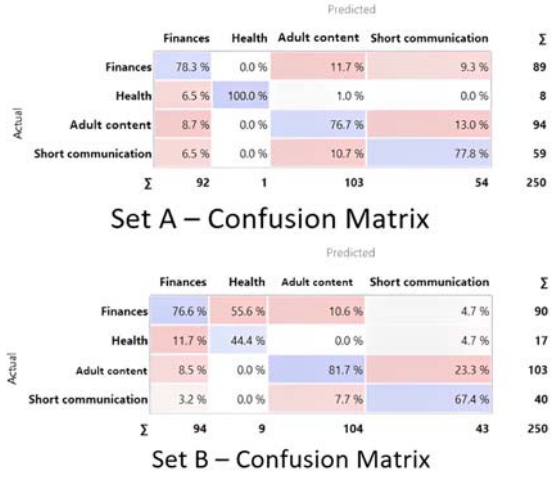

Figure 2. Confusion Matrices per Category for AdaBoost

Most represented categories in datasets are predicted more consistently: "Adult content" and "Finances" with predictions ranging from 70-80% for both datasets and both algorithms, whereas less represented categories demonstrated significant variations.

D. Results of Predicted and Tested Accuracies

Predicted scores are based on the complete training set, and were obtained by each algorithm across all four categories. Tested scores were obtained through test sets (25 phishing e-mails) which are used to test the accuracy for prediction of categories by each algorithm.

Fig. 3 presents results of tested and predicted scores across all four categories. Accuracy of tested scores is lower than accuracy of predicted scores, generally by 2-6%.

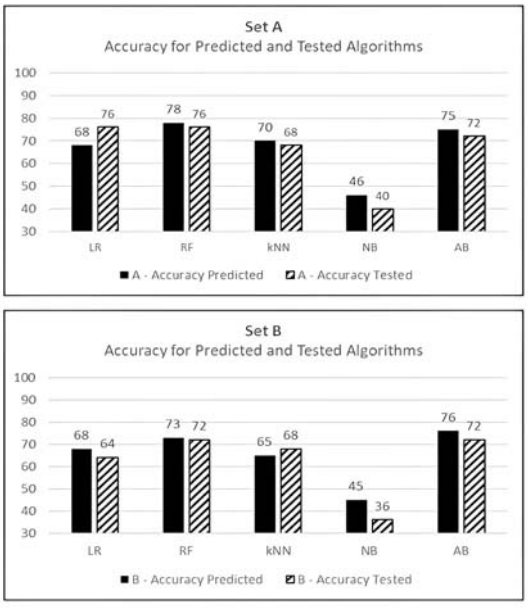

Figure 3. Tested and predicted accuracies by algorithm

Fig. 4 presents results for predicted and tested average scores per category, showing significant variations among categories.

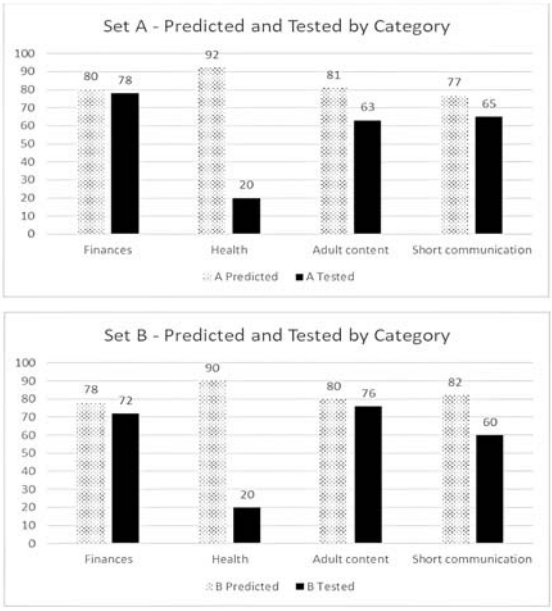

Figure 4. Predicted and tested accuracies by category

Results show a decline of tested accuracy scores in all categories, but the smallest decrease is in the "Finances" category, followed by the category "Adult content", which are the most represented categories in the training sets. The most significant decline was in the "Health" category (72% in Set A and 70% in Set B), which had the smallest amount of training data.

VI. CONCLUSION

The main goal of this research was to demonstrate a corpus-based information extraction pipeline that is language independent and built on machine learning techniques for predicting categories of phishing e-mails.

The research was performed on two sets, differentiated by language originality (Croatian or automatically translated from English into Croatian). Training and test sets were imbalanced in terms of categories. The most represented categories in training and test sets were "Adult content" and "Finances". The category "Health" was the least represented category, and achieved the lowest accuracy scores. However, a significant number of phishing e-mails contain mixed content, i.e. all of them deal more or less with financial issues, and therefore can belong to multiple categories. This was confirmed to be the main limitation of this research.

Results are compared between predicted accuracy scores and tested accuracy scores, with regard to the chosen machine learning algorithm and with regard to the predicted category. When comparing algorithms with manual testing results, the best predictions were obtained by Random Forest and AdaBoost, while Naïve Bayes performed worst.

Predicted results obtained by classification accuracy by applying Random Forest and AdaBoost are compared with manual testing results for each of the four categories. Both algorithms predicted most consistently categories of "Adult content" and "Finances", which were the most represented categories in the training datasets, with scores of 70-80% for both subsets (A and B). This result shows the immense importance of data quantity in the training set. The biggest variations in prediction scores and when comparing predicted and tested score values are for the category "Health", which is the least represented category in the training dataset, exposing the weakness of machine learning algorithms on small datasets.

Results in this paper confirm that predicted accuracy increases with data quantity. The best predictions are obtained for categories that were most represented in datasets, whereas worst results were obtained for the least represented category "Health". Overall, algorithms that exhibited best prediction results were ensemble machine learning algorithms – Random Forest and AdaBoost.

In order to improve accuracy results, the authors plan to increase sample size equalize the number of phishing e-mails that are used for in training and test sets, and harmonize the length of phishing messages.

ACKNOWLEDGEMENT

This research is funded by institutional research projects "61-920-2553" and "11-933-1053" at the Faculty of Humanities and Social Sciences, University of Zagreb.

REFERENCES

[1] M. Dadkhah, S. Shamshirband, and A. Wahab, "A hybrid approach for phishing web site detection", The Electronic Library, vol. 34, no. 6, pp. 927–944, 2016.

[2] Australian Cyber Security Centre, "ACSC Annual Cyber Threat Report: July 2019 to June 2020", https://www.cyber.gov.au/sites/default/files/2020-09/ACSC-Annual-Cyber-Threat-Report-2019-20.pdf

[3] M. Rosenthal, "Must-Know Phishing Statistics: Updated 2022", Tessian, https://www.tessian.com/blog/phishing-statistics-2020/

[4] A. Kovač, I. Dunđer, and S. Seljan, "An overview of machine learning algorithms for detecting phishing attacks on electronic messaging services", 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2022, pp. 954–961, DOI: 10.23919/MIPRO55190.2022.9803517.

[5] D. Aljeaid, A. Alzhrani, M. Alrougi, and O. Almalki, "Assessment of End-User Susceptibility to Cybersecurity Threats in Saudi Arabia by Simulating Phishing Attacks", Information, vol. 11, no. 12, DOI: https://doi.org/10.3390/info11120547, 2020.

[6] F. Carroll, J. A. Adejobi, and R. Montasari, "How Good Are We at Detecting a Phishing Attack? Investigating the Evolving Phishing Attack Email and Why It Continues to Successfully Deceive Society", SN Computer science, vol. 3, DOI: https://doi.org/10.1007/s42979-022-01069-1, 2022.

[7] S. Pandiyan, P. Selvaraj, V. K. Burugari, J. P. Benadit, and P. Kanmani, "Phishing attack detection using Machine Learning, Measurement", Sensors, vol. 24, ISSN: 2665-9174, DOI: https://doi.org/10.1016/j.measen.2022.100476, 2022.

[8] O. Yavanoglu, and M. Aydos, "A review on cyber security datasets for machine learning algorithms", IEEE International Conference on Big Data, Boston, USA, 2017, pp. 2186–2193, DOI: 10.1109/BigData.2017.8258167.

[9] H. F. Atlam, and O. Oluwatimilehin, "Business Email Compromise Phishing Detection Based on Machine Learning: A Systematic Literature Review", Electronics, vol. 12, no. 42, DOI: https://doi.org/10.3390/electronics12010042, 2023.

[10] H. Hindy et al., "A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems", IEEE Access, vol. 8, pp. 104650–104675, DOI: 10.1109/ACCESS.2020.3000179, 2020.

[11] S. Doroudi, "The Bias-Variance Tradeoff: How Data Science Can Inform Educational Debates", AERA Open, vol. 6, no. 4, DOI: https://doi.org/10.1177/2332858420977208, 2020.

[12] A. Téllez-Valero, M. Montes, and L. Villaseñor-Pineda, "A Machine Learning Approach to Information Extraction", Lecture Notes in Computer Science, vol. 3406, pp. 539–547. DOI: 10.1007/978-3-540-30586-6_58, 2005.

[13] M. Almseidin, A. Abu Zuraiq, M. Al-kasassbeh Mouhammd, and N. Alnidami, "Phishing Detection Based on Machine Learning and Feature Selection Methods", International Journal of Interactive Mobile Technologies (iJIM), vol. 13, no. 12, DOI: doi:10.3991/ijim.v13i12.11411, 2019.

[14] A. A. Orunsolu, A. S. Sodiya, and A. T. Akinwale, "A predictive model for phishing detection", Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 2, pp. 232–247, DOI: https://doi.org/10.1016/j.jksuci.2019.12.005, 2022.

[15] S. Rawal, B. Rawal, A. Shaheen, and M. Shubham, "Phishing Detection in E-mails using Machine Learning", International Journal of Applied Information Systems, vol. 12, no. 7, pp. 21–24, 2017.

[16] I. Zatezalo, and I. Dunđer, "Online service for accessible machine learning of prediction models", Zbornik radova Međimurskog veleučilišta u Čakovcu, vol. 12, no. 2, 2021.

[17] A. Hannousse, and S. Yahiouche, "Towards benchmark datasets for machine learning based website phishing detection: An experimental study", Engineering Applications of Artificial Intelligence, vol. 104, ISSN: 0952-1976, DOI: https://doi.org/10.1016/j.engappai.2021.104347, 2021.

[18] J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, "Phishing Detection Using Machine Learning Technique", First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Riyadh, Saudi Arabia, 2020, pp. 43–46, DOI: 10.1109/SMART-TECH49988.2020.00026.

[19] J. Brownlee, "A Gentle Introduction to Ensemble Learning Algorithms", 2021, https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/

[20] Y. Zhang, J. Liu, and W. Shen, "A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications", Applied Sciences, vol. 12, DOI: 10.3390/app12178654, 2022.

[21] R. Odegua, "An Empirical Study of Ensemble Techniques (Bagging, Boosting and Stacking)", Deep Learning IndabaX Conference, Ibadan, Nigeria, 2019, DOI: 10.13140/RG.2.2.35180.10882.

[22] K. Nikhil, D. S. Rajesh, and D. Raghavan, "Phishing Website Detection Using ML", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), vol. 7, no. 4, pp. 194–198, 2021.

[23] H. Kumar, A. Prasad, N. Rane, N. Tamane, and A. Yeole, "Dr. Phish: Phishing Website Detector", International Advanced Research Journal in Science, Engineering and Technology (IARJSET), vol. 8, no. 1, pp. 176–182, 2021.

[24] M. Gupta, and S. D. Pandya, "A Comparative Study on Supervised Machine Learning Algorithm", International Journal for Research in Applied Science and Engineering Technology (IJRASET), vol. 10, no. 1, pp. 1023–1028, 2022.

[25] B. E. Boukari, A. Ravi, and M. Msahli, "Machine Learning Detection for SMiShing Frauds", IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, USA, 2021, DOI: 10.1109/CCNC49032.2021.9369640.

[26] K. M. Ting, "Confusion Matrix", in: C. Sammut, and G. I. Webb (eds), Encyclopedia of Machine Learning. Springer, DOI: https://doi.org/10.1007/978-0-387-30164-8_157