Explainable Machine Learning for Bag of Words-based Phishing Detection

Maria Carla Calzarossa[1][0000−0003−1015−3142], Paolo Giudici[2][0000−0002−4198−0127], and Rasha Zieni[1][0000−0002−5383−2738]

[1] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy
[2] Department of Economics and Management, University of Pavia, Pavia, Italy

Abstract. Phishing is a fraudulent practice aimed at convincing individuals to reveal sensitive information, such as account credentials or credit card details, by clicking the links of malicious websites. To reduce the impacts of phishing, the timely identification of these websites is essential. For this purpose, machine learning models are often devised. In this paper, we address the problem of website phishing detection by proposing an explainable machine learning model based on bag of words features extracted from the content of the webpages. To select the most important features to be used in the model, we propose to employ the Lorenz Zonoid, the multidimensional generalization of the Gini coefficient. The resulting model is characterized by a good accuracy and it provides explanations of which words are most likely associated with phishing websites. In addition, the number of features retained is significantly reduced, thus making the model parsimonious and easier to interpret.

Keywords: Explainable machine learning · Phishing detection · Lorenz Zonoid.

1 Introduction

Phishing is a fraudulent practice that has been around for many years because of its straightforward implementation and of its large potential financial benefits.

The main goal of this practice is to manipulate individuals and convince them to reveal various types of sensitive data, such as bank account credentials, credit card details or other important financial or personal information. As a consequence, individuals and companies might suffer monetary losses, identity thefts and reputation damages. In fact, attackers often leverage the hijacked accounts for performing illegal online transactions or simply sell the collected data on the marketplace.

The implementation of phishing campaigns requires some simple operations, such as:

– Creation of websites that look similar to the legitimate counterparts attackers are trying to impersonate;

- Dissemination of the links of the malicious websites using spoofed email messages or other communication media;
- Creation of a sense of urgency in the individuals targeted by the attacks.

Phishing is a very active phenomenon as indicated in the Phishing Activity Trends Reports[3] periodically published by Anti-Phishing Working Group (APWG). For example, a total of 1,350,037 phishing attacks has been observed in the fourth quarter of 2022. This was up slightly from the third quarter, which, with 1,270,883 attacks, represents a new record and the worst quarter for phishing observed by APWG. These numbers are partly due to the large number of attacks from persistent phishers against several specific targets.

To protect individuals from the risks deriving from phishing attacks, the detection of phishing websites is of paramount importance (see, e.g., Zieni et al. [22] for a detailed review). Approaches based on machine learning are particularly suitable for classifying sites as either phishing or legitimate. Features are typically extracted from the links used to reach the websites, i.e., the Uniform Resource Locators (URLs), and from the page source codes. These features have to take into account the characteristics that differentiate the two classes of websites as well as the strategies implemented by attackers to deceive individuals.

In the context of machine learning, explainability plays a fundamental role to understand the decisions taken by the learning algorithms. In this paper, we address this issue by proposing a methodology and an application that allow us to identify the most important features able to explain a machine learning model for bag of words-based phishing detection, thus extending the recent work of Calzarossa et al. [4] which was based on structured data.

The rest of the paper is organized as follows. After this introduction, the state of the art is analyzed in Section 2. Section 3 describes the considered unstructured textual data and feature extraction. Section 4 focuses on the methodology proposed to make machine learning models explainable, based on Lorenz Zonoids, while the experimental results are presented in Section 5. Finally, after a summary of the main findings given in Section 6, some concluding remarks are offered in Section 7.

2 Related work

As already mentioned, approaches based on machine learning have been extensively investigated in the literature for detecting phishing websites. In fact, machine learning models are particularly effective since they cope well with zero-hour attacks and allow on-the-fly detection of phishing webpages. In what follows, we review the state of the art in the contexts of phishing detection and of explainable Artificial Intelligence.

[3] https://docs.apwg.org/reports/apwg_trends_report_q4_2022.pdf

2.1 Phishing detection

The machine learning approaches applied in the context of phishing detection mainly differ in terms of the features chosen to describe the properties of the websites and of the learning algorithms applied for the classification.

Features can be extracted from the page source codes or from the page URL. For example, features can be generated by analyzing the textual and visual properties of the source codes, the relationships among their components; and the visual appearance of the page (see, e.g., [5, 9, 15]).

Concerning URL features, these mainly refer to the lexical properties of the URL strings as well as to the patterns existing in these strings considered either at the character or at the word level (see, e.g., [1, 11, 13, 14, 20, 21]). For example, for analyzing the composition of URL strings, Ma et al. [12] consider the words URLs consist of and extract features according to the bag of words representation. In detail, a bag of words representation is derived for each of the identified URL components, namely, hostname, second level domain, top level domain, pathname and file extension. As a results, tens of thousands of features are generated.

We recall that a bag of words is a representation of text that describes and maps the occurrence of words within a document into a vector. This model, often used for extracting features from unstructured textual data, is motivated by the intuition that documents containing similar words have similar meanings.

Despite the good predictive power of bag of words features, we outline that these representations have not been used frequently in the framework of phishing website detection because of two main limitations, namely, the high dimensionality of the feature vectors combined with their sparsity.

The methodology proposed in this paper overcomes these limitations since it significantly reduces the number of features without significantly affecting the performance of the machine learning model. In addition, our approach makes the model explainable, that is, it clearly identifies the words most likely associated with phishing websites.

2.2 Explainable AI

Machine learning models are boosting Artificial Intelligence (AI) applications in many domains. Although complex machine learning models may reach high predictive performance, their predictions are not explainable: they cannot be understood and overseen by humans in terms of their drivers (see, e.g., [2]).

The requirement of explainability is fulfilled "by design" through classic statistical models, such as logistic and linear regression. However, in complex data analysis problems, classical statistical models may have a limited predictive accuracy, in comparison with "black-box" machine learning models, such as neural networks and Random Forests.

This suggests to empower machine learning models with post-modelling tools that can "explain" them. Recent attempts in this direction, based on the cooperative game theory work by Shapley [17], have led to promising applications of

explainable AI methods (see, e.g., [2] and [3]). Shapley values have the advantage of being agnostic: independent on the underlying model with which predictions are computed, although they have the disadvantage of not being normalised and, therefore, they make it difficult the interpretation and comparison between different models.

A recent work [8] has shown how to overcome this issue by means of Shapley-Lorenz values, which combine Shapley values with Lorenz Zonoids [7], obtaining a global measure of the contribution of each explanatory variable to the predictive accuracy of the response, rather than to the value of the predictions, as is the case for Shapley values.

The Shapley Lorenz metrics has so far been applied only to structured data. With this paper we contribute to the state of the art on explainable AI by proposing and implementing a Shapley Lorenz based explainable AI method for textual data.

Numerous attempts at explaining bagging ensembles (by model reduction, model simplification, etc) have been proposed, see e.g., [16]. Most of these models rely on explainability notions that are model dependent. A noticeable example is the feature importance plot, which can be used for tree models, but cannot be employed for other types of machine learning models. In general, to achieve a full explainability, agnostic tools, independent on the underlying model, are necessary.

In the context of phishing detection, explainable AI is very important for ensuring transparency, providing trust in the decisions made by machine learning models and enhancing cybersecurity practices. Nevertheless, the research on model agnostic explainable AI models in this field is rather limited (see, e.g., [4, 6]). In particular, in [6] two explanatory techniques, namely, Local Interpretable Model-agnostic Explanations and Explainable Boosting Machine, are applied for explaining the models developed for detecting phishing URLs, whereas in [4], the Lorenz Zonoid approach is applied to detect phishing websites described by structured data.

3 Data collection and feature extraction

The dataset used in this study is part of a publicly available dataset consisting of about 1.6 million observations of legitimate and phishing webpages [18, 19]. One target response variable and ten explanatory attributes referring to various properties of the page URL and source code are associated with each observation.

To keep our dataset manageable, we extract from it 70,000 observations, each described by one attribute, namely, the raw content of filtered text and JavaScript code, and the target variable. We remark that these observations are equally distributed between legitimate and phishing webpages, thus, unlike the original dataset, our dataset is balanced.

To generate the features, we represent the raw content of each webpage using the bag of words model. In detail, from the content we extract words, i.e., tokens

delimited by spaces, and we associate with each word its frequency: the number of times a word appears in a webpage.

A preliminary analysis of the raw content of each webpage shows that some words (e.g., those related to the JavaScript code) do not clearly differentiate phishing and legitimate pages, thus, it is better to discard them. To this aim, we apply multiple filters created using regular expressions and we remove the tokens referring to JavaScripts function names (e.g., `eval(), find(), lastIndex()`) and to English stopwords. In addition, we discard one character words as well as the numbers and any character that does not belong to the English alphabet. Finally, all words are transformed into lower cases using case folding process.

As a result of these pre-processing steps, we obtain a bag of words consisting of 21,156 words, which will be the candidate explanatory features.

4 Methods

In this section we present our proposed methodology: an explainable machine learning procedure based on Lorenz Zonoids for bag of words feature selection.

It is well known that non linear machine learning models, such as ensemble trees and neural networks, can lead to highly accurate predictions, typically better than those obtained with linear models.

Bag of words phishing detection is a classification problem, for which a Random Forest model could be an appropriate model to consider. A Random Forest model averages the classifications obtained from a set of tree models, each of which is based on a boostrap sample of training data and of feature variables. In a tree, the feature variables determine the splitting rules, and a statistical measure of variability (such as the one dimensional Gini measure of variability) determines when to stop splitting.

A Random Forest model increases the predictive accuracy of the trees of which it is the average, at the expense of explainability. To overcome this weakness, a variable importance plot can be used. However, such a plot is not fully agnostic, as it cannot be applied for models differerent from ensemble models. For this reason, explainable Artificial Intelligence methods need to be employed, such as methods based on Shapley values [8, 17].

Variable importance plots associate with each predictor the corresponding reduction in the Gini index, averaged over all tree models. Although useful from a descriptive viewpoint, the variable importance plot has the additional disadvantage of not choosing the most significant predictors. In this paper, We fill this gap by proposing a variable selection procedure based on Lorenz Zonoids, which extends the Gini coefficient.

Lorenz Zonoids were proposed in [10] as a multidimensional generalization of the ROC curve. In the one-dimensional case, the Lorenz Zonoid coincides with the Gini coefficient, a well known measure in the study of income inequality or wealth inequality within a nation or a social group. The Lorenz Zonoid measures statistical dispersion in terms of mutual variability among the observations, a metric that is more robust to anomalous and extreme data, with respect to the

variance, which measures statistical dispersion in terms of variability from the mean.

For a given set of n observations of a response variable Y to be predicted, the Lorenz Zonoid can be defined as the area between the Lorenz and the dual Lorenz curves.

More formally, given a response variable Y, the Lorenz Zonoid L_Y is obtained ordering the Y values in a non-decreasing sense: joining the points with coordinates $(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y}))$, for $i = 1, \ldots, n$, where r_j and $\bar{y}$ indicate the (non-decreasing) ranks of Y and the Y mean value, respectively. Similarly, the dual Lorenz curve $L_Y^{'}$ is obtained ordering the Y values in a non-increasing sense: joining the points with coordinates $(i/n, \sum_{j=1}^{i} y_{d_j}/(n\bar{y}))$, for $i = 1, \ldots, n$, where d_j indicates the (non-increasing) ranks of Y. The area lying between the L_Y and $L_Y^{'}$ curves is the Lorenz Zonoid.

Giudici and Raffinetti [7] introduced Lorenz Zonoids in the field of machine learning, for model comparison and selection of a parsimonious model. They showed that, given a set of K explanatory variables, and letting $\hat{Y}_{X' \cup X_k}$ and $\hat{Y}_{X'}$ be, respectively, the predicted values obtained from a model – which includes a covariate X_k – and the predicted values obtained from a model – which excludes covariate X_k – the additional contribution of a covariate X_k can be obtained as:

$$\frac{LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'})}{LZ(Y) - LZ(\hat{Y}_{X'})}, \tag{1}$$

where $LZ(\hat{Y}_{X' \cup X_k})$, $LZ(\hat{Y}_{X'})$ and $LZ(Y)$ are, respectively: the Lorenz Zonoids computed on the predictions obtained – including covariate X_k; the Lorenz Zonoids computed on the predictions obtained excluding covariate X_k; and the Lorenz Zonoids computed on the Y response variable values.

Model comparison can then be implemented associating with each additional variable (feature) the term $LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'})$, which measures its relative importance in terms of its additional contribution to the predictive accuracy of a model.

To implement model comparison, it remains to choose in which order inserting the variables in a model. To save computational time, we suggest to follow a forward stepwise procedure, which starts from the null model (containing no variables) and proceeds inserting one variable at a time, following the ranking determined by the variable importance plot.

We remark that the variable importance plot is a well known explainability tool employed in the context of ensemble tree models, based on the notion of reduction in Gini variability, strictly related to Lorenz Zonoid. We cannot rely exclusively on the variable importance plot, for explainability purposes, as we cannot use it for models different from ensemble trees (such as neural networks or regression models). However, given its relationship with Lorenz Zonoids, it is worth employing its results as a pre-processing step to Lorenz Zonoid model comparison, to determine the order in which to insert variables. This can substantially help in reducing computational complexity.

5 Experimental results

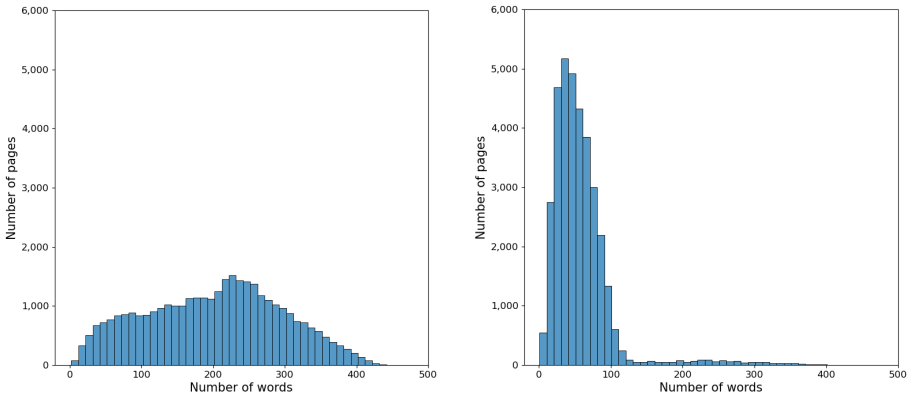
This section presents the experimental results obtained by applying the proposed methodology to the bag of words representation of phishing and legitimate webpages described in Section 3.

An exploratory analysis of the words extracted from each webpage provides some preliminary insights on the variables of interest. A first interesting result refers to the composition of phishing and legitimate pages in terms of the number of words they include, reported in Table 1.

Webpage	Mean	Std dev	Max	1st Quartile	Median	3rd Quartile	Number of pages
Phishing	201	94.9	458	126	207	270	35,000
Legitimate	58	45.3	431	32	49	71	35,000
All	129	103.2	458	45	84	214	70,000

Table 1. Statistics of the composition of the webpages expressed in terms of number of words they include.

From the table we observe that phishing pages differ significantly, and contain more words with respect to legitimate ones. On average, a phishing page includes 201 words compared to 58 words of a legitimate page. The frequency distributions of the words, in both phishing and legitimate pages, are shown in Figure 1.



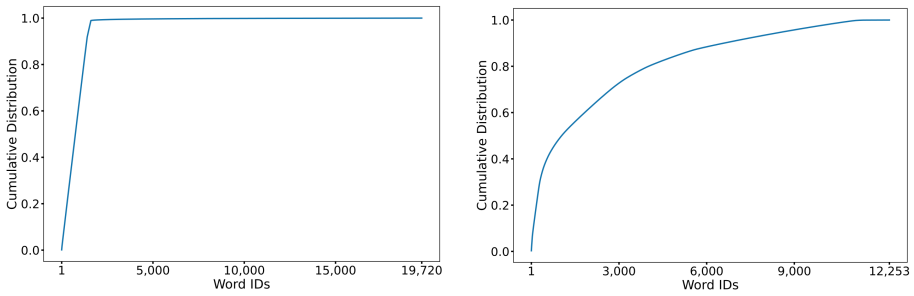
(a) Phishing webpages (b) Legitimate webpages

Fig. 1. Distributions of frequency of the words appearing in phishing (a) and legitimate (b) webpages

The figure confirms the difference in word frequency between phishing and legitimate pages: the former not only contains, on average, more words, but it is also much more variable. From an interpretational viewpoint, the identified differences could be seen as a specific strategy of the attackers, who tend to insert longer texts to confuse users.

A second result that can be obtained from the exploratory analysis is that the number of unique words appearing in phishing and legitimate pages also differs. As expected, while many unique words are used in phishing pages (i.e., 19,720 out of the 21,156 words extracted from the dataset), the text of legitimate pages is not as rich, as only 12,253 unique words appear in their content.

Another important aspect that can be explored is related to the concentration of the words. For this purpose, Figure 2 plots the cumulative distributions of the frequency of the words used in phishing and legitimate pages. We can easily notice that the two distributions are quite different. For the phishing pages a



(a) Phishing webpages (b) Legitimate webpages

Fig. 2. Cumulative distributions of frequency of the words appearing in phishing (a) and legitimate (b) webpages

very small fraction of the words, that is, 1,375 words, accounts for about 90% of the overall frequency of all words, and a large fraction of all words (about 51%) appears only once. On the contrary, for legitimate pages, 6,700 words are required to account for 90% of the overall frequency, whereas very few words (838) appear only once. From an interpretational result, the previous findings demonstrate that attackers tend to use the same words, leading to a highly concentrated distribution.

We now move to the application of machine learning models. To build a predictive model able to recognize webpages either as phishing or legitimate, a Random Forest classifier is applied to our data, namely, considering 21,156 candidate explanatory features for the binary phishing/legitimate response. The performance of the classifier, using all features, as well as the first 10,000 and 5,000 most important features (plus a case with nine features that will be explained later) is illustrated in Table 2.

Features	Accuracy	Precision	Recall	F1-score
21,156 (all)	0.9506	0.95	0.95	0.95
10,000	0.9510	0.95	0.95	0.95
5,000	0.9512	0.95	0.95	0.95
9	0.8904	0.90	0.89	0.89

Table 2. Random Forest performance as a function of the number of features used by the classifier.

The table clearly shows that the predictive performance of the classifier with all features is very good: the accuracy is about 95%. Similarly, precision, recall and F1-score are also rather good. Reducing the number of features to 10,000 and 5,000 does not significantly reduce such performance.

Figure 3 shows a snapshot of the Random Forest feature importance plot, in which the first 32 most important features are displayed. These features are the words that strongly differentiate phishing and legitimate webpages. Note that in the figure we do not explicitly name the words because they are generally dirty words. In fact, our conjecture is that most phishing pages refer to fake adult websites.

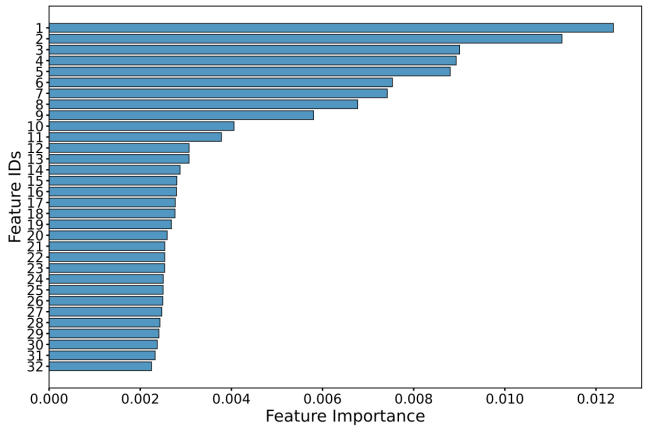


Fig. 3. Random Forest feature importance plot

The figure shows that much of feature importance is captured by the first five words, suggesting that the Random Forest model may be simplified without losing too much accuracy. Indeed, this figure is the basis for selecting the most important features to be retained, while making the model explainable, by means of our proposed Lorenz Zonoid feature selection procedure.

To this end, the Lorenz Zonoid is calculated for the most important features by adding one of them at a time, following the order described by the feature importance plot: starting from the feature with the highest importance, then with that with the second highest importance, and so on.

In more detail, the Random Forest model is first trained using only the most important word. On the basis of such a model, predictive scores of phishing are calculated for all observations in the validation set. The Lorenz Zonoid is applied to the obtained predictive scores, leading to a first accuracy measure, based on a model with only one feature. The procedure is repeated using both the first and the second most explainable variables, leading to a second Lorenz Zonoid accuracy measure which is likely to have a higher value than the first one, because of the inclusion property of Lorenz Zonoid. The process of adding features, in the order indicated by the feature importance plot, is then repeated, leading to progressively greater values of the Lorenz Zonoids, until a stopping point is reached.

As a stopping point we suggest to follow the "elbow rule" employed in determining the optimal number of principal components in factor analysis: stop when the second derivative of the plot of the Lorenz Zonoid measures changes sign.

The Lorenz Zonoids selection procedure, applied to our data, is shown in Figure 4. As can be seen, the Lorenz Zonoid measure, which is approximately

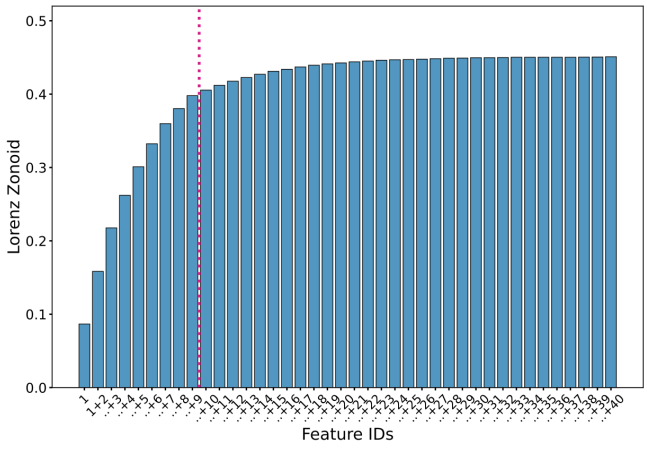


Fig. 4. Lorenz Zonoid for the most important features. The plot reports the Lorenz Zonoid values for models of increasing size, obtained adding one feature at a time, in the order described by the feature importance plot.

equal to 10% using only the first word feature, rapidly increases until the first nine words are considered, although with a declining acceleration. The elbow

rule, applied to Figure 4, suggests that a good stopping point is to consider nine features.

6 Discussion of findings

The findings outlined in this section explain why we have inserted in Table 2 also a model with nine features: it is exactly the model selected by our Lorenz Zonoid procedure.

Indeed, from Table 2 we can observe that, if we drastically reduce the number of words necessary for the predictive classifiers, from as many as 21,156 to only 9, the accuracy only reduces by 6%. A similar behavior is obtained for the other performance measures: the precision reduces only by 5 basis points, from 0.95 to 0.90; the recall only by 6 basis points, from 0.95 to 0.89, and similarly the F1-score.

Altogether, the accuracy measures indicate that a model with 9 features is only slightly less accurate than a model with 21,156 features. On the other hand, a model with only 9 features is much better explained than a model with 21,156 features. It is also very likely that a model with 9 variables is more robust to data variations with respect to a model with 21,156 variables.

We finally remark that a simpler alternative to the Shapley Lorenz method is the variable importance plot, often employed to achieve explainability in ensemble tree models. While variable importance plots are useful, and we make use of them in the paper, they are not fully agnostic as they cannot be employed to other types of learning models, thereby limiting model comparison.

7 Conclusion

In this paper we deal with the problem of building an explainable machine learning model, that is able to correctly classify, in advance, whether a certain website is phishing or legitimate. We have shown how to tackle the problem in the realistic case of data consisting of unstructured documents.

To build a machine learning model, we have first proposed a bag of words representation of the data. We have then given some intuition on how phishing operates, by means of exploratory data analysis.

We have then applied a Random Forest model to a set of 21,156 candidate explanatory feature words. Given the high complexity of the full model, we have shown that our proposed Lorenz Zonoid selection procedure can lead to a drastic reduction of complexity, without significantly affecting the model accuracy: we are able to reduce the important features with 9 feature words, losing only 6% of the predictive accuracy.

Our contribution to the explainable AI field is twofold. From a methodological viewpoint, we have proposed a feature selection method, based on Lorenz Zonoid comparisons, which can drastically simplify a bag of Words model, without losing much predictive accuracy. From an applied viewpoint, we have shown how our

proposed methods can lead to a model explained by a limited set of features, that can be oversight to monitor and detect on time possible phishing attacks.

Our results indicate that bag of words phishing detection could lead to very useful insights on the nature of phishing websites, and can help not only to identify a priori suspicious websites, but also to understand the words that identify and explain them. These are important results for end users of websites and also for the authorities aimed at monitoring cyber attacks and implementing cyber security measures.

Future research is needed in the extension of what presented here to similar data problems. Indeed, our approach is very general and can be applied to any machine learning model based on unstructured data analysis. The advantage of our proposal will be particularly evident, in terms of a gain in explainability and a reduction in computational costs, in problems in which the number of available feature variables is large.

Finally, we remark that, although our results clearly demonstrate the effectiveness of our proposed methodology, further comparative analyses should be conducted with existing and novel techniques to provide a more robust assessment.

References

1. Blum, A., Wardman, B., Solorio, T., Warner, G.: Lexical feature based phishing URL detection using online learning. In: Proceedings of the ACM Conference on Computer and Communications Security. pp. 54–60 (2010)
2. Bracke, P., Datta, A., Jung, C., Shayak, S.: Machine learning explainability in finance: an application to default risk analysis. Staff Working Paper, Bank of England (816) (2019)
3. Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable AI in Credit Risk Management. Computational economics **57**(1), 203–216 (2021)
4. Calzarossa, M., Giudici, P., Zieni, R.: Explainable machine learning for phishing feature detection. Quality and Reliability Engineering International (2023)
5. Corona, I., Biggio, B., Contini, M., Piras, L., Corda, R., Mereu, M., Mureddu, G., Ariu, D., Roli, F.: DeltaPhish: Detecting Phishing Webpages in Compromised Websites. In: Foley, S., Gollmann, D., Snekkenes, E. (eds.) Computer Security – ESORICS, Lecture Notes in Computer Science, vol. 10492, pp. 370–388. Springer (2017)
6. Galego Hernandes, P., Floret, C., Cardozo De Almeida, K., Da Silva, V., Papa, J., Pontara Da Costa, K.: Phishing Detection Using URL-based XAI Techniques. Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), IEEE (2021)
7. Giudici, P., Raffinetti, E.: Lorenz model selection. Journal of classification **37**(2), 754–768 (2020)
8. Giudici, P., Raffinetti, E.: Shapley-Lorenz explainable artificial intelligence. Expert Systems with Applications **158**(895), 1–9 (2021)
9. Jain, A., Gupta, B.: A Machine Learning Based Approach for Phishing Detection Using Hyperlinks Information. Journal of Ambient Intelligence and Humanized Computing **10**, 2015–2028 (2019)
10. Koshevoy, G., Mosler, K.: The Lorenz Zonoid of a Multivariate Distribution. Journal of the American Statistical Association **91**(434), 873–882 (1996)
11. Le, A., Markopoulou, A., Faloutsos, M.: PhishDef: URL names say it all. In: Proceedings of the 30th IEEE International Conference on Computer Communications - INFOCOM. pp. 191–195. IEEE (2011)
12. Ma, J., Saul, L., Savage, S., Voelker, G.: Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD. pp. 1245—1254. ACM (2009)
13. Ma, J., Saul, L., Savage, S., Voelker, G.: Learning to detect malicious URLs. ACM Transactions on Intelligent Systems and Technology **2**(3) (2011)
14. Marchal, S., Francois, J., State, R., Engel, T.: PhishStorm: Detecting Phishing with Streaming Analytics. IEEE Transactions on Network and Service Management **11**(4), 458–471 (2014)
15. Rao, R., Pais, A., Anand, P.: A Heuristic Technique to Detect Phishing Websites Using TWSVM Classifier. Neural Computing and Applications **33**(11), 5733–5752 (2021)
16. Sagi, O., Rokach, L.: Explainable decision forest: Transforming a decision forest into an interpretable tree. Information Fusion **61**, 124–138 (2020)
17. Shapley, L.: A value for n-person games. Contributions to the Theory of Games II pp. 307–317 (1953)

18. Singh, A.: Dataset of malicious and benign webpages. Mendeley Data (2020), Available: https://data.mendeley.com/datasets/gdx3pkwp47/2
19. Singh, A.: Malicious and benign webpages dataset. Data in Brief **32**, 106304 (2020)
20. Tupsamudre, H., Singh, A., Lodha, S.: Everything is in the Name – A URL Based Approach for Phishing Detection. In: Dolev, S., Hendler, D., Lodha, S., Yung, M. (eds.) Cyber Security Cryptography and Machine Learning, Lecture Notes in Computer Science, vol. 11527, pp. 231–248 (2019)
21. Verma, R., Dyer, K.: On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers. In: Proceedings of the 5th ACM Conference on Data and Application Security and Privacy - CODASPY. pp. 111–122. ACM (2015)
22. Zieni, R., Massari, L., Calzarossa, M.: Phishing or not phishing? A survey on the detection of phishing websites. IEEE Access **11**, 18499–18519 (2023)