Fraudulent e-commerce websites detection through machine learning*

Manuel Sánchez-Paniagua[1,2], Eduardo Fidalgo[1,2], Enrique Alegre[1,2], and Francisco Jáñez-Martino[1,2]

[1] Department of Electrical, Systems and Automation, University of León, León, Spain
[2] Researcher at INCIBE (Spanish National Cybersecurity Institute), León, Spain
manuel.sanchez,eduardo.fidalgo,enrique.alegre,francisco.janez@unileon.es

Abstract. With the emergence of e-commerce, many users are exposed to fraudulent websites, where attackers sell counterfeit products or goods that never arrive. These websites take money from users, but also they can stole their identity or credit card information. Current applications for user protection are based on blacklists and rules that turn out into a high false-positive rate and need a continuously updating. In this work, we built and make publicly available a suspicious of being fraudulent website dataset based on distinctive features, including seven novel features, to identify these domains based on recently published approaches and current web page properties. Our model obtained up to 75% F1-Score using Random Forest algorithm and 11 hand-crafted features, on a 282 samples dataset.

Keywords: E-commerce · Fraud detection · Machine Learning · Cybersecurity

1 Introduction

Over the past few years, retail companies have started the digital transformation to provide their services and products through the Internet [1]. Most physical and new companies are migrating to e-commerce websites to reach customers and display the available goods [25]. Those websites often have the same structure to make them accessible and intuitive to anybody interested in the brand or the company. Fraudsters take advantage of the similarity between websites and create e-commerce online stores to place counterfeit products into the market or to take customer's money in return for no items.

Due to this change in business model, more and more people are entering e-commerce services to obtain products. Statista, a global business data platform, stated that user penetration will be 50.8% in 2021 and is expected to hit 63.1% by 2025. These predictions promote an increase in the online market, which means that more users will be exposed to e-commerce fraud. The Organisation

* Supported by INCIBE.

for Economic Co-operation and Development (OECD) found in 2016, that up to 3.3% of the global trade are counterfeit products [19]. Furthermore, the European Commission found that 62% of surveyed customers have suffered a buying scam[3] [6]. These surveys and reports highlighted the impact of these scams on the final user and the targeted companies.

Current protection for users depends on its knowledge and experience to prevent these attacks. There are also online detection tools, like ScamAdviser[4], ScamFoo[5] or Scamner[6], where users can check the confidence of a certain domain. These services use information systems and rules to collect and analyze data about the domain. However, many users may not known these services or may directly trust in the websites. Thereby, an automatic system to speed up the early detection of fraudulent pages is required without depending on human participation.

The main problem of rule-based systems is that new legitimate websites obtain low confidence scores due to their similarities with fraud ones. Registration date and volume of users for a new domain are the main reason for the low confidence score in many of the aforementioned tools. For both reasons, companies and cybersecurity experts look for implementing automatic systems based on Artificial Intelligence to speed up the early detection of fraudulent websites [27,4,14]. Recent works retrieve features from HTML, text, images and 3^{rd} party services and use traditional machine learning algorithms to detect fraudulent websites [27].

In this paper, we propose a pipeline capable of detecting suspicious of being fake online shops using data from the actual page and 3^{rd} party services. Throughout this manuscript, we use *fraudulent websites* term to refer to *suspicious of being fraudulent websites*. We adapt the set of features of Wu et al. [27] and add novel features from Secure Sockets Layer (SSL) certificate and TrustPilot service information, as well as policies pages, e-commerce development technologies and social media links. We look for exploring the challenge of detecting fraudulent websites to set a baseline for this research line with a novel feature vector. First, we collect a set of e-commerce websites to build a tailored dataset to use for our proposal, called Features from Fraudulent Websites 282 (FFW-282). Second, we evaluate the use of a feature vector based on sample analysis to determine the features to be extracted. Finally, we assess five different machine learning algorithms to state the best performance model.

Using machine learning instead of traditional rules provide our work with a holistic point of view. Some rules may categorize a website by its age [27], therefore if the website is one month old, it is directly identified as fraud, generating a false positive since old domains have their legitimacy proved over time.

[3] a buying scam includes: fake goods, undelivered goods or services, fake invoices and unwanted monthly subscriptions.
[4] https://www.scamadviser.com/
[5] https://www.scamfoo.com/
[6] https://www.scamner.com/

To address this issue, we count on different features to correctly identify these threats and prevent users from getting their money stolen.

This paper is structured as follows. Section 2 presents a review of the literature and related work. Section 3 explains the methodology, the features proposed and the metrics used. Section 4 present the results of the different experiments, Section 5 contains the conclusion along with the limitations and future work.

2 Related work

2.1 Literature Review

The detection of fraudulent e-commerce websites is an emerging challenge for cybersecurity agencies due to their fast growth and the potential harm to people. Many authors have been developed machine learning systems to deal with fraudulent websites [25,27,4,14,18,16,24]. Due to its continuous evolution and because of being a dynamic environment, recent works focused their proposals on different technologies to retrieve enough information.

On the one hand, some authors focused their research on the features extraction from the websites [24,18,14,27]. Wadleigh et al. [24] based their feature set on URL, web content and WHOIS properties. Mostard et al. [18] extracted features from HTML code which include web-page length, emails, phone number or payment methods. Khoo et al. [14] also retrieved text and images from the pages to detect fraudulent websites, keeping HTML features. Wu et al. [27] widened the number of features and used a feature vector that includes information from URL, social media and email addresses, payment forms or phone number appearance, WHOIS and content structure. However, previous works have not explored features such as policies pages, Trustpilot data, e-commerce technologies and Secure Sockets Layer (SSL) certificate, that may contain information more in line with current frauds.

On the other hand, other works considered natural language processing techniques to find out similarities among web pages, rather than retrieving features [4,16]. Beltzung et al. [4] used similarity between source code from websites by analyzing its HTML, Javascript, CSS, among others. Maktabar et al. [16] applied techniques of text classification introducing a sentiment analysis model on the textual content of the web-page.

2.2 Online Fraudulent Tools

In this subsection, we briefly describe online tools that are available and help users to detect fraudulent websites.

ScamAdviser relies on an automated algorithm that checks different sources to retrieve information about the domain. They rely on the registration and expiration date from WHOIS, the ranking of the website at Alexa, the reviews about the domain, a verification of the SSL certificate, and the e-commerce technology used by the online shop. By applying these rules, the website generates a confidence number from zero to 100 and left the user obtain its conclusions.

ScamFoo also has a set of rules to check website trustworthiness. It recollects information from external services to generate a confidence score. It implements similar services as ScamAdviser and also checks the domain on different blacklists looking for previous reports.

Scamner has a similar structure as previous tools. Additionally, it uses MozRank to provide more information about the ranking based on the number of websites linking to a target domain. It also displays the social media interactions for users to obtain their conclusions.

Finally, **Web Of Trust** is another tool that states to use community ratings, reviews and machine learning algorithms to obtain their rankings but does not provide further information about their method.

3 Methodology

In this study, (1) we propose a set of feature vectors to describe the most relevant information of a website to be considered as a fraudulent one, (2) we build a novel dataset based on the selected features, and (3) we evaluate five traditional machine learning algorithms. We show the entire assessment process in the Fig. 1.

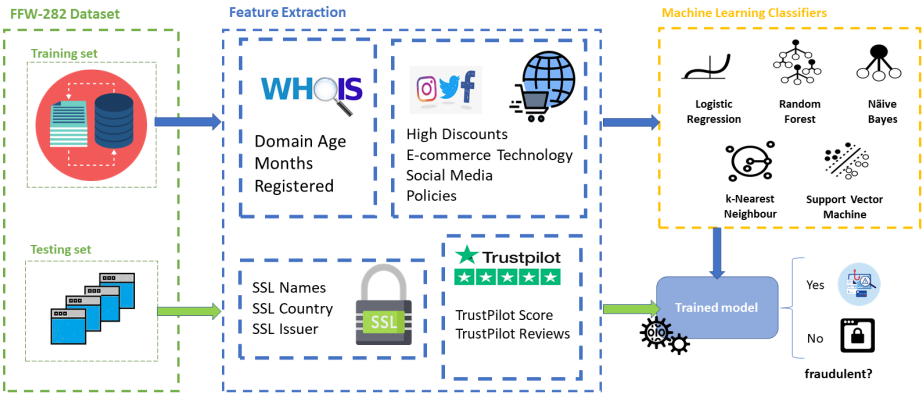

Fig. 1. Graphical Abstract of the evaluation process: (1) we split our dataset in training and testing sets, then (2) we extracted the proposed features from each domain and, finally, (3) we trained and tested five machine learning models to detect fraudulent websites.

3.1 Proposed features

In this section, we describe the selected features for this task. After reviewing other works [4,27,24], we also consider features used in these works but using them from a novel perspective, such as high discounts, social media, domain age and months registered. To improve the performance of fraudulent detectors against current websites properties, we incorporate seven novel features as are

SSL names, country and issuer, Trustpilot score and review, e-commerce technologies and policies. The complete list of features used are the following ones.

High discounts: Users are susceptible of end up tricked when they have the opportunity to obtain a valuable product with a great deal or discount [11]. Fraudsters use large discounts ($> 70\%$) to persuade users to buy a bargain. Usually, these offers come with a countdown which is an urgent appeal to increase even more users' susceptibility [26]. Furthermore, some works [24] calculate the average discount for the displayed items and used it as an input feature. For this reason, we look over the HTML code for high discounts, and we include in the feature vector how many discount banners we find.

SSL names: A single Secure Sockets Layer (SSL) certificate can protect multiple domain names. In this way, brands and companies use the same certificate for their online shop on different servers and countries. On the opposite side, attackers do not count on big infrastructures to serve their website in different countries. Therefore, if they have an SSL certificate, it may have only one name registered. We check the number of names registered as a feature for the model.

SSL country: There is a set of banned countries that cannot obtain an SSL certificate due to restrictions from the Certificate Authority (CA) in verification task for organization and domains [5]. We verify if the certificate has any of the banned or risk country codes included in Table 1.

Table 1. Banned and risk countries on SSL certificates.

Country	Country Code	Status
Cuba	CU	Banned
Iran	IR	Banned
North Korea	NP	Banned
Sudan	SD	Banned
Syria	SY	Banned
Eritrea	ER	Risk
Guinea	GN	Risk
Iraq	IQ	Risk
Lebanon	LB	Risk
Pakistan	PK	Risk
Rwanda	RW	Risk
Sierra Leona	SL	Risk
Zimbabwe	ZQ	Risk

SSL issuer: Most legitimate websites use common SSL issuers. After an analysis, we found 16 companies in charge of generating SSL certificates: Let's Encrypt, Symantec, Geotrust, Comodo, DigiCert, Thawte, Network Solutions, Rapid SSL, Entrust Datacard, SSL.com, Sectigo, Cloudflare, GoDaddy, Google Trust Services, Amazon and CPanel. Let's Encrypt is one of the most common due to their free service [3], which is the main option for fraudsters. However, we should pay attention because it is also used by small companies that are looking for getting into the digital market. Furthermore, CPanel is not a certificate issuer

but a service provider. Since there are a great number of websites using CPanel, we also added it to the list.

Trustpilot score: Trustpilot[7] is a website where users provide opinions about websites, products and services based on their experiences. Reviews go along with a rating score ranging from one to five and they are important for online shops to increase user confidence [21]. Trustpilot calculates the website score with the mean of all rates provided by users. We obtained this score and used it as a legitimate feature.

Trustpilot reviews: Another important parameter from Trustpilot is the number of reviews. Aged online shops have a great number of reviews but with a mid-range score due to the polarity of customers reviews [22]. By adding the number of reviews, we provide a holistic view of the legitimacy of a web page.

E-commerce technologies: Most legitimate online shops watch over their design, accessibility and functionality to set an initial trust with the customer [13]. To do that, developers tend to use tools and frameworks with these capabilities to provide the best user experience. Unlike legitimate sites, fraudsters display raw websites, sometimes, implemented from a simple HTML template. We propose to detect the different technologies used by the website since poor and fast designs are commonly related to fraud sites. Wappalyzer allowed us to retrieve the technologies used by the website through the fingerprint exposed in the HTML code. In this feature, we counted the number of e-commerce frameworks or tools were used in the website development, like *Shopfy*, *WooCommerce* or *Zen Cart*.

Social media: A company developing its digital commerce usually implies consumer brand engagement (CBE) campaigns on social media by creating accounts on different platforms such as Twitter, Instagram or Facebook [12]. We have noticed that some fake websites provide empty social media links of redirections to blank pages. We extracted the social media links from the HTML code and verified two statements: first, if there is a link to any of the three platforms we have mentioned and second, if those links are connected to the actual company profile. We count how many of the three social media accounts are linked to the online shop website.

Policies: Legitimate e-commerce websites clarifies their conditions, terms and policies related to refunds, user's data, shipments and consumer contracts. This information improves the relationship between the user and the online shop [17]. On the other hand, fake shops do not usually provide any of those statements or links to them are blank. Therefore, we check how many links related to policies are responding correctly.

Domain age: WHOIS information has been used in other fields such as phishing detection to identify short aged servers [2]. Fraud websites also have a short life span since authorities try to take them down once a user reports them. Besides, legitimate websites domains have been registered since their beginning and it is likely that to have a longer life span than fraud web pages. However, this feature can discriminate against recent legitimate websites. We count the

[7] `https://www.trustpilot.com/` Retrieved July 2021

number of months from the registration date on WHOIS information for the target domain.

Months registered: Following the previous feature, attackers do not register their domain for a long time since the attack does not last for so long. We count the number of months from the registration date to the domain expiration date on WHOIS information, i.e., the life period.

3.2 Dataset Creation

The objective of this work is to detect fraudulent websites among e-commerce sites. Furthermore, we need information from the WHOIS service and technology analysis, so we decided to build a dataset that complies with this task. Figure 2 displays the process to collect the dataset used in this work.

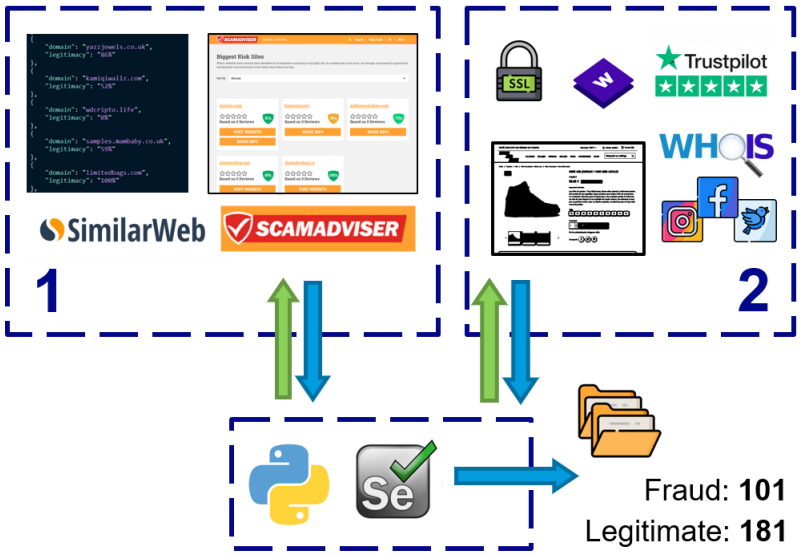

Fig. 2. Dataset recollection process. (1) We obtain the domains from the sources and (2) the script visits the websites and collect the data from it and the 3^{rd} party services

The first step was to identify the sources that provide the target domains. Starting on the fraudulent class, we used user reports on ScamAdviser. These reports contain the domain name submitted by the user and the confidence score calculated using the predefined rules of ScamAdviser. We collected 197 total domains from ScamAdviser from February to June 2020. For those domains, a low confidence score is an indicator of a suspicious website, while higher scores identify legitimate e-commerce websites, which were the most common type among the obtained reports. Since domain names were submitted by users and evaluated with fixed rules, we performed a manual analysis over the collected samples to determine the best threshold to treat a website as suspicious or not.

After the examination, we observed that all domains with a score higher than 75% were legitimate, while most of the rest were suspicious. Therefore, legitimate domains with a score under 75% were reviewed and manually labelled to avoid bias in the dataset.

For the legitimate class, we obtained the most visited online stores. We obtained the top 50 worldwide e-commerce domains from SimilarWeb[8]. We also introduced 34 well-known domains that did not appear in the list. The final dataset, Features from Fraudulent Websites 282 (FFW-282), is composed of 181 legitimate e-commerce domains and 101 fraudulent domains and it is publicly available for research purposes[9].

Once we have the domains, we used Selenium Webdriver and Python3 to visit and recollect the features from each domain. First, we obtain the HTML content and use a regex to retrieve if high discounts are in the text. Then, using the SSL module, we call *getpeercert* to retrieve SSL certificate information. To obtain the e-commerce technologies used in the Wappalyzer library, which identifies them by the code fingerprints in the website. After collecting the offline features, we call the external services for further information. First, we introduce the domain name in Trustpilot and collect its score and number of opinions registered on the website. Second, we use the WHOIS library to collect the information related to the actual domain. Finally, we search in the HTML code for social media links, specifically Facebook, Instagram and Twitter. As soon as we have those links, we check if the link corresponds to a valid user or not. All this information is stored in a JSON file that we use later for creating the feature vectors.

3.3 Classifiers

We trained five classifiers to obtain fraudulent prediction models, all of them widely used in the literature [9,10] and very different from each other: Random Forest (RF) [23], Support Vector Machine (SVM) [7], k-Nearest Neighbour (kNN) [20], Logistic Regression (LR) [8], and Naïve Bayes (NB) [15].

4 Experiments and results

4.1 Experimental setup

Experiments are executed on an Intel Core i3 8100 at 3.6Ghz and 16GB of DDR4 RAM. We used scikit-learn and Python 3. Due to the small dataset size and likely bias, we averaged the output of a 5-Fold Cross Validation.

We have tested different settings to select the best combination of the main parameters for each classifier. For the rest of the parameters, we have used scikit-learn default values since we found no difference in their tuning. In the

[8] `https://www.similarweb.com/top-websites/category/e-commerce-and-shopping/` Retrieved July 2021

[9] `http://gvis.unileon.es/dataset/features-from-fraudulent-websites-282/` Retrieved July 2021

case of Random Forest, we obtained the best result using *n_estimators = 10* and $max\_features = auto$. SVM obtained its best results using *C = 1.0* and Radial Basis Function (RBF) kernel. A high value of C parameter looks for a lower margin of hyperplane separation. Optimal parameters for kNN were 4 neighbours using *manhattan* metric. Logistic Regression was set with *l2* penalty, $C = 1$ and *Limited-memory BFGS* solver. Finally, Naïve Bayes obtained its best results with the Bernoulli algorithm.

4.2 Performance metrics

To assess the performance of the proposed classification models, we employed accuracy, precision, recall and F-score. We denoted the fraud class as the positive class and the legitimate class as the negative one.

We used F1-Score as the main metric for evaluation purposes since our dataset is slightly unbalanced. It can be computed as shown in Eq. 1.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{1}$$

Precision is also a relevant metric in this field and it is defined as the fraction of correctly classified fraud samples over the number of items classified as fraud, as indicated by Eq. 2.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

True Positive (TP) indicates the number of fraud samples correctly classified and False Positive (FP) depicts the number of legitimate samples wrongly classified as fraud.

Recall refers to the fraction of correctly classified fraud samples over the total number of fraud instances as indicated by Eq. 3.

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

True Negative (TN) refers to the number of legitimate samples correctly identified as legitimate and False Negative (FN) denotes the number of fraud samples improperly classified as legitimate.

Finally, the accuracy represents the number of samples that were correctly classified and it is calculated as shown in Eq. 4

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

4.3 Evaluation of machine learning algorithms

In this experiment, we compare the proposed algorithms with the best parameters for this task. Results were obtained by calculating the mean value between the 5 folds.

Table 2. Results of the main machine learning algorithms

Algorithm	Precision	Recall	F1-Score	Accuracy
RF	80.00	70.59	**75.00**	**85.96**
kNN	60.87	82.35	70.00	78.95
SVM	61.90	76.47	68.42	78.95
LR	55.56	88.24	68.18	75.44
NB	53.33	94.12	68.09	73.68

Based on the results in Table 2, Random Forest obtained the best results among other classifiers with a 75% F1-Score. Random Forest was the most balanced algorithm and the only one with higher precision than recall, a conservative standpoint where fraud predictions are certain. However, the model misses an important rate of fraud samples (low recall). The opposite happens to the rest of the algorithms. According to the overall precision (80.00%) and recall (70.59%), most of the fraud websites are detected (high recall) but with a higher false-positive rate. Therefore, more legitimate samples are predicted as fraud while they are not. In LR and NB, almost half of the legitimate samples were misclassified as fraud (low precision). Since the increasingly risk of fraudulent websites, like scams or leaked data, for users, we recommend Random Forest.

5 Conclusion and future work

In this paper, we have proposed a model for fraudulent e-commerce website detection. We have presented a collection of features, including seven novel ones like SSL properties, Trustpilot metrics, policies analysis and e-commerce technologies, that can help in this task. Those are based on sample inspection and fraudsters techniques that differentiate a legitimate website from a scam one. Using these features, we have created a model to detect these websites with a 75% F1-Score. Results suggest that Random Forest is the algorithm with best performance. Although precision is higher than recall, we recommend its usage for this task. We consider this work as a good baseline to improve the detection of fraudulent websites from different perspectives but also presenting several contributions. First, we introduce and made publicly available our dataset FFW-282, although it contains a small number of samples which may not be enough to train a complex machine learning model.Since ScamAdviser advised that its information should consider as a recommendation rather than a ground truth, we manually inspected the websites to fix a score threshold above 75% labelling its e-commerce websites as legitimate ones, the rest ones were considered as suspicious of fraudulent activities. Therefore, we consider that building a larger dataset could be a high priority task, since there is no data available that can be used for detecting fraudulent e-commerce website. However, building a proper dataset has limitations since the sources of the most visited online shops requires a payment.

Second, the proposed feature vector could be improved to achieve a more descriptive descriptor of each website. We found on fraudulent websites specific

identifiers that indicated a high probability of fraud, and they were the core of the proposed features set in this paper. Hence, with the enlargement of the dataset and a profound analysis of the websites and their components, we could look for more valuable features that may improve the performance of the model.

References

1. Adoption rate of emerging technologies in organizations worldwide as of 2020. https://www.statista.com/statistics/661164/worldwide-cio-survey-operational-priorities/, accessed: 2021-02-25
2. Adebowale, M., Lwin, K., Sánchez, E., Hossain, M.: Intelligent web-phishing detection and protection scheme using integrated features of images, frames and text. Expert Systems with Applications **115**, 300–313 (2019). https://doi.org/10.1016/j.eswa.2018.07.067
3. Aertsen, M., Korczyński, M., Moura, G., Tajalizadehkhoob, S., Van Den Berg, J.: No domain left behind: Is Let's Encrypt democratizing encryption? In: ANRW 2017 - Proceedings of the Applied Networking Research Workshop, Part of IETF-99 Meeting. pp. 48–57 (2017). https://doi.org/10.1145/3106328.3106338
4. Beltzung, L., Lindley, A., Dinica, O., Hermann, N., Lindner, R.: Real-Time Detection of Fake-Shops through Machine Learning. In: Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020. pp. 2254–2263 (2020). https://doi.org/10.1109/BigData50022.2020.9378204
5. Bottarini, J.: List Of Countries Banned & Restricted From Obtaining SSL Certificates (2016), https://www.wiyre.com/list-of-countries-banned-restricted-from-obtaining-ssl-certificates/, accessed: 2021-04-16
6. Commission, E.: Survey on scam and fraud experienced by consumers (2020), https://ec.europa.eu/info/sites/info/files/aid_development_cooperation_fundamental_rights/ensuring_aid_effectiveness/documents/survey_on_scams_and_fraud_experienced_by_consumers_-_final_report.pdf, accessed: 2021-02-25
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995). https://doi.org/https://doi.org/10.1007/BF00994018
8. Cox, D.R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) **20**(2), 215–232 (1958). https://doi.org/doi.org/10.1111/j.2517-6161.1958.tb00292.x
9. Das, M., Saraswathi, S., Panda, R., Mishra, A.K., Tripathy, A.K.: Exquisite Analysis of Popular Machine Learning–Based Phishing Detection Techniques for Cyber Systems. Journal of Applied Security Research **0**(0), 1–25 (2020). https://doi.org/10.1080/19361610.2020.1816440
10. Dou, Z., Khalil, I., Khreishah, A., Al-Fuqaha, A., Guizani, M.: Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection. IEEE Communications Surveys Tutorials **19**(4), 2797–2819 (2017). https://doi.org/10.1109/COMST.2017.2752087
11. Goel, S., Williams, K., Dincelli, E.: Got phished? Internet security and human vulnerability. Journal of the Association for Information Systems **18**(1), 22–44 (2017). https://doi.org/10.17705/1jais.00447
12. Hollebeek, L., Glynn, M., Brodie, R.: Consumer brand engagement in social media: Conceptualization, scale development and validation. Journal of Interactive Marketing **28**(2), 149–165 (2014). https://doi.org/10.1016/j.intmar.2013.12.002

13. Karimov, F., Brengman, M., van Hove, L.: The effect of website design dimensions on initial trust: A synthesis of the empirical literature. Journal of Electronic Commerce Research **12**(4), 272–301 (2011)
14. Khoo, E., Zainal, A., Ariffin, N., Kassim, M., Maarof, M., Bakhtiari, M.: Fraudulent e-commerce website detection model using html, text and image features. Advances in Intelligent Systems and Computing **1182 AISC**, 177–186 (2021). https://doi.org/10.1007/978-3-030-49345-5_19
15. Lewis, D.D.: Naive (Bayes) at forty: The independence assumption in information retrieval. In: Nédellec, C., Rouveirol, C. (eds.) Machine Learning: ECML-98. pp. 4–15. Springer Berlin Heidelberg, Berlin, Heidelberg (1998)
16. Maktabar, M., Zainal, A., Maarof, M., Kassim, M.: Content based fraudulent website detection using supervised machine learning techniques. Advances in Intelligent Systems and Computing **734**, 294–304 (2018). https://doi.org/10.1007/978-3-319-76351-4_30
17. McCole, P., Ramsey, E., Williams, J.: Trust considerations on attitudes towards online purchasing: The moderating effect of privacy and security concerns. Journal of Business Research **63**(9-10), 1018–1024 (2010). https://doi.org/10.1016/j.jbusres.2009.02.025
18. Mostard, W., Zijlema, B., Wiering, M.: Combining visual and contextual information for fraudulent online store classification. In: IEEE/WIC/ACM International Conference on Web Intelligence. pp. 84–90 (2019). https://doi.org/10.1145/3350546.3352504
19. OECD, Office, E.U.I.P.: Trends in Trade in Counterfeit and Pirated Goods. (2019). https://doi.org/https://doi.org/10.1787/g2g9f533-en
20. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009). https://doi.org/10.4249/scholarpedia.1883, revision #137311
21. Proserpio, D., Zervas, G.: Online reputation management: Estimating the impact of management responses on consumer reviews. Marketing Science **36**(5), 645–665 (2017). https://doi.org/10.1287/mksc.2017.1043
22. Schoenmueller, V., Netzer, O., Stahl, F.: The Polarity of Online Reviews: Prevalence, Drivers and Implications. Journal of Marketing Research **57**(5), 853–877 (2020). https://doi.org/10.1177/0022243720941832
23. Statistics, L.B., Breiman, L.: Random Forests. In: Machine Learning. pp. 5–32 (2001)
24. Wadleigh, J., Drew, J., Moore, T.: The e-commerce market for "lemons": Identification and analysis of websites selling counterfeit goods. In: WWW 2015 - Proceedings of the 24th International Conference on World Wide Web. pp. 1188–1197 (2015). https://doi.org/10.1145/2736277.2741658
25. Weng, H., Li, Z., Ji, S., Chu, C., Lu, H., Du, T., He, Q.: Online E-Commerce Fraud: A Large-Scale Detection and Analysis. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE). pp. 1435–1440 (2018). https://doi.org/10.1109/ICDE.2018.00162
26. Williams, E., Hinds, J., Joinson, A.: Exploring susceptibility to phishing in the workplace. International Journal of Human Computer Studies **120**, 1–13 (2018). https://doi.org/10.1016/j.ijhcs.2018.06.004
27. Wu, K., Chou, S., Chen, S., Tsai, C., Yuan, S.: Application of machine learning to identify Counterfeit Website. In: iiWAS2018: Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services. pp. 321–324 (2018). https://doi.org/10.1145/3282373.3282407