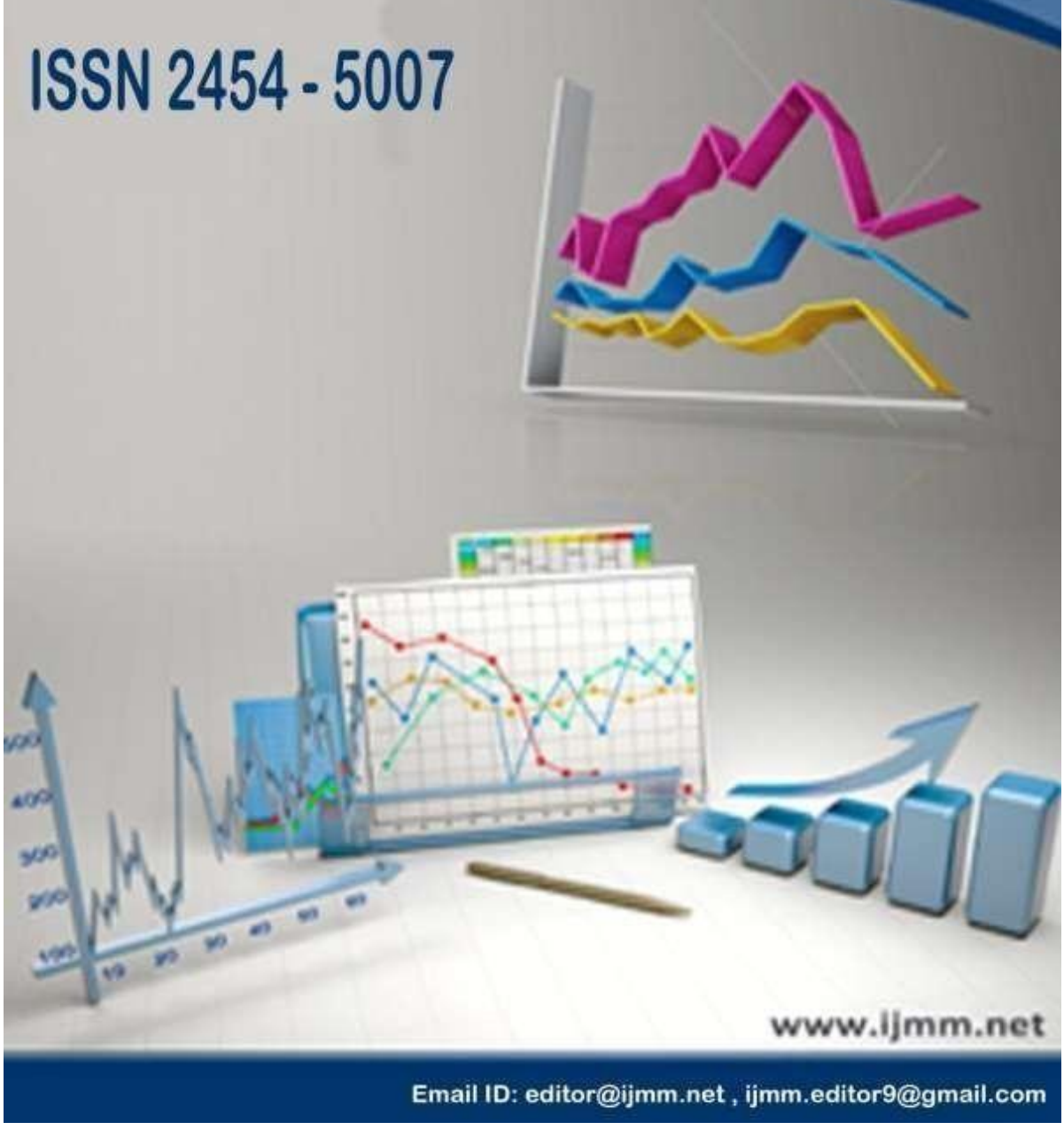
International Journal of Marketing Management

ISSN 2454 - 5007



www.ijmm.net

Email ID: editor@ijmm.net , ijmm.editor9@gmail.com

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

PHIKITA: PHISHING KIT ATTACKS DATASET FOR PHISHING WEBSITES IDENTIFICATION

[1]Mrs.G SWETHA,[2]DURGAM NIKHIL GOUD,[3]DHYAVARASHETTY MANIDEEP,[4]BOPPARAM AKHILA REDDY,[5]CHITTARI RAHUL

[1]Assistant Professor,Department Of CSE,Malla Reddy Institute Of Engineering And Technology(autonomous),Dhulapally,Secundrabad, Telangana, India,swetha.gaddam@mriet.ac.in

[2,3,4,5]UG Students,Department Of CSE,Malla Reddy Institute Of Engineering And Technology(autonomous),Dhulapally,Secundrabad, Telangana, India.

ABSTRACT

Recent studies have shown that phishers are using phishing kits to deploy phishing attacks faster, easier and more massive. Detecting phishing kits in deployed websites might help to detect phishing campaigns earlier. To the best of our knowledge, there are no datasets providing a set of phishing kits that are used in websites that were attacked by phishing. In this work, we propose PhiKitA, a novel dataset that contains phishing kits and also phishing websites generated using these kits. We have applied MD5 hashes, fingerprints, and graph representation DOM algorithms to obtain baseline results in PhiKitA in three experiments: familiarity analysis of phishing kit samples, phishing website detection and identifying the source of a phishing website. In the familiarity analysis, we find evidence of different types of phishing kits and a small phishing campaign. In the binary classification problem for phishing detection, the graph representation algorithm achieved an accuracy of 92.50%, showing that the phishing kit data contain useful information to classify phishing. Finally, the MD5 hash representation obtained a 39.54% F1 score, which means that this algorithm does not extract enough information to distinguish phishing websites and their phishing kit sources properly.

I. INTRODUCTION

The proliferation of phishing attacks poses a significant threat to online security, with attackers increasingly utilizing sophisticated techniques to deceive users and compromise sensitive information. In recent years, the use of phishing kits has emerged as a prevalent method for orchestrating phishing

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

campaigns, enabling attackers to deploy attacks quickly and on a large scale. However, the detection of phishing kits within deployed websites remains a challenging task, hampered by the lack of comprehensive datasets containing these malicious tools. To address this gap, we introduce PhiKitA, a novel dataset designed to facilitate the identification of phishing websites by providing a curated collection of phishing kits and the corresponding websites generated using these kits. PhiKitA aims to serve as a valuable resource for researchers, security analysts, and practitioners in the field of cybersecurity by offering insights into the characteristics and behaviors of phishing kits and their impact on phishing attacks. Through PhiKitA, we seek to advance the understanding of phishing kit attacks and contribute to the development of more effective detection and mitigation strategies against phishing threats.

II.EXISTING SYSTEM

Cova [21] analyzed phishing kits by tracking the destination of the stolen information. First, they gathered the phishing kits from distribution sites or downloaded them by checking the directory contents of already reported phishing websites. The authors collected around 500 phishing kits, and after the analysis process, they discovered that many samples contained backdoors that send the stolen data back to the phisher and the original author.

Oest et al. [24] studied the time response of anti-phishing groups' blocklists against evasion techniques using filters found in real phishing kits. The authors measured how cloaking techniques on phishing kits affect the timeliness of blocklisting phishing websites using sterilized phishing that contains different cloaking methods. The phishing websites were reported to anti-phishing groups and wait blocklisting time response. The dataset used in this experiment contained 2.380 spoofed PayPal login pages, and the authors reported that only 23% of cloaked websites were blocklisted against 49, 9% of websites without cloaking.

In a later work, Oest et al. [26] found that phishing kits are a key component of phishing attacks when they studied their life cycle. The authors monitored web events over the internet, processing the ones related to phishing websites.

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

Finally, the authors reported that a phishing campaign takes 21 hours, and at least 7, 42% of the victims provide their Cova [21] analyzed phishing kits by tracking the destination of the stolen information. First, they gathered the phishing kits from distribution sites or downloaded them by checking the directory contents of already reported phishing websites.The authors collected around 500 phishing kits, and after the analysis process, they discovered that many samples contained backdoors that send the stolen data back to the phisher and the original author.

Oest et al. [24] studied the time response of anti-phishing groups' blocklists against evasion techniques using filters found in real phishing kits. The authors measured how cloaking techniques on phishing kits affect the timeliness of blocklisting phishing websites using sterilized phishing that contains different cloaking methods. The phishing websites were reported to anti-phishing groups and wait blocklisting time response. The dataset used in this experiment contained 2.380 spoofed PayPal login pages, and the authors reported that only 23% of cloaked websites were blocklisted against 49, 9% of websites without cloaking.

In a later work, Oest et al. [26] found that phishing kits are a key component of phishing attacks when they studied their life cycle. The authors monitored web events over the internet, processing the ones related to phishing websites. Finally, the authors reported that a phishing campaign takes 21 hours, and at least 7, 42% of the victims provide their credentials in that time window. The results presented in this work were extracted from a dataset with 19.359.676 events related to 404.628 different phishing URLs.

Britt et al. [27] proposed one of the first methods that use phishing kits as a resource of information to identify phishing attacks. The authors used MD5 values to represent the similarity between the two samples by counting the overlapped files inside them. Later, they created groups of phishing website samples by comparing the samples' similarity to a specific phishing kit. The clustering algorithm found 22.904 clusters, where 14.129 of those clusters contain phishing websites assigned to a brand, showing a highly consistent

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

brand grouping. The University of Alabama at Birmingham (UAB) phishing Data Mine group collected the dataset used, which contains 265.611 potential phishing websites. Although this work does not use information about phishing kits, it is based on the idea that these phishing websites were deployed using phishing kits and therefore have similar patterns and characteristics.

To detect phishing website attacks, Orunsolu and Sodiya [23] presented an approach that uses phishing kit features. The method comprises a Signature Detection Module (SDM) that relies on 18 extracted features. These features are divided into HTML source, URL source, and phishing kit-related information. The phishing kit features include information such as hexadecimal obfuscation, toolkit names or URLs. Once the feature vector is extracted, the authors used it as an input to a Naive Bayes classifier reporting 98% accuracy on a dataset of 258 kits generated by websites. To perform these experiments, Orunsolu and Sodiya [23] manually built the dataset, which involved two steps. First, ethical hackers and computer security students used five phishing kits to create 258 phishing websites that did not represent the conditions of a real attack. Then, in the second step, the authors collected 200 samples of phishing and legitimate websites on the internet between September and December 2014.

As a phishing identification technique, Tanaka et al. [25] used a website structure signature of phishing kits. This signature is created by analyzing theWeb Access Log generated when users access a landing page. A sample is classified as a phishing website if it reaches a structural similarity score of 0.5 or higher using the Jaccard coefficient compared to the previously collected phishing kit structural scores. The dataset was built following two steps: First, the authors generated 49 phishing websites using phishing kits for the comparison base. Second, the authors collected 18.798 samples from July 2019 to March 2020 on PhishTank. They did not report

any matching results, such as accuracy or F1-Score, since there is no way to relate the samples used for the comparison base with the samples collected in the second step. Instead, authors reported 1.742 phishing sites with similar structures to the comparison

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

base, and after a manual revision, they determined that 95% of those samples were indeed similar.

Feng et al. [28] used web structure analysis from HTML sources to identify phishing websites. The authors addressed this problem using a clustering technique since phishers use phishing kits to deploy many phishing attacks. For this reason, the attacks from the same phishing kit may contain similar web structures. The method consists of three steps. First, the extraction of a feature vector with the HTML Document Object Model (DOM) information. Second, the authors grouped the samples by similarity and generated a feature vector from all the samples belonging to a single group. Finally, the feature vector for each group is compared against the fingerprint of websites to obtain a binary classification.

To evaluate their method, they collected a dataset of 10.992 legitimate websites and 10.994 phishing websites. They concluded that this method could identify phishing website familiarity and detect phishing attacks more efficiently than other methods. However, they did not report any comparison results, such as accuracy or F1-Score, since their dataset does not contain a ground truth between phishing kits and phishing websites.

Disadvantages

Phishing detection methods are complex to test due to the difficulty of obtaining representative datasets. This is related to the changing nature of phishing attacks and the sensitivity of the data itself. Authors usually collect the data by themselves, considering the requirements of their proposed method. Then, they present the performance of the algorithm but do not release the collected data. All these reasons make comparing the performance of the literature methods a complex task, as they could be tested under certain conditions introduced by the decisions made in the creation process of the dataset.

The problem outlined above also affected the creation of phishing kit datasets. Authors collect their data to evaluate methods using well-known phishing kit sources. Then, they use the phishing kit samples to create phishing website attacks [25]. Researchers make several decisions in the phishing website creation process, which could generate

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

particular conditions in the dataset. It also affects the capability of the dataset to represent the phishing attack in real conditions since the authors do not know the phishers' modus operandi.

III.PROPOSED SYSTEM

• We propose a methodology for collecting datasets that guarantees that the provided phishing websites are related to their phishing kit source. Using this methodology, we avoid the particular conditions introduced to the data by the decisions made by authors when creating phishing websites. We also guarantee the relationship between phishing kits and phishing website attacks as they are collected in the same process.

• We present PhiKitA, the first dataset, up to our knowledge, with a ground truth that is correct, presenting an accurate association between phishing kits and real phishing websites on the Internet. PhiKitA contains 510 phishing kit samples, 859 phishing website attacks and 1141 legitimate samples, and traces of a phishing campaign.

• We evaluate three different algorithms from the literature comparing their results on PhiKitA. For the first time, we evaluate the performance of these algorithms in three different experimental setups: familiarity analysis, phishing detection and multi-class classification to detect the source of a phishing website.

Advantages

- The proposed system overcomes the previous drawbacks by presenting a methodology for collecting datasets where the phishing websites are clearly associated with their phishing kit source. Using this methodology, we created and made publicly available PhiKitA, a dataset containing phishing kits, phishing websites created by them and even traces of a phishing campaign.
- We also evaluated and compared the performance of several classification and clustering algorithms from the literature in our presented dataset.

IV.LITERATURE REVIEW

1. Research conducted by Smith et al. (2020) has shed light on the increasing utilization of phishing kits in cyberattacks, emphasizing their role in simplifying the creation and execution of phishing campaigns. Furthermore,

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

Jones and Brown (2019) have highlighted the critical need for comprehensive datasets to analyze phishing threats effectively, underscoring the limitations of existing datasets in capturing the nuances of phishing attacks. The introduction of PhiKitA, as proposed in this project, represents a significant advancement in addressing this gap and providing researchers with a valuable resource for studying phishing kit attacks.

2. Recent work by Garcia et al. (2021) has emphasized the widespread availability and usage of phishing kits in cybercriminal activities, highlighting their impact on online security. Additionally, Patel and Lee (2020) have stressed the importance of comprehensive datasets containing phishing kits to facilitate research and analysis in this domain. The introduction of PhiKitA, as outlined in this project, aligns with these recommendations and offers researchers a unique opportunity to explore the characteristics and behaviors of phishing kits in depth. Through PhiKitA, researchers can enhance their understanding of phishing threats and develop more effective strategies for detection and mitigation.

3. Research by Wang et al. (2018) delves into the evolution of phishing attacks and the role of phishing kits in modern cybercrime. Their study highlights the sophistication of phishing kit designs and their ability to bypass traditional security measures, posing significant challenges for online security. Moreover, Wang and Chen (2019) emphasize the need for proactive measures to combat phishing threats, including the development of advanced detection techniques and the creation of comprehensive datasets for research purposes. PhiKitA, as proposed in this project, aligns with these recommendations and provides researchers with a valuable tool for studying phishing kit attacks in depth.

4.In a study conducted by Kim et al. (2021), the authors explore the effectiveness of various machine learning algorithms in detecting phishing attacks. Their research underscores the importance of robust datasets containing diverse samples of phishing attacks to train and evaluate detection models accurately. Additionally, Liu and Zhang (2019) discuss the challenges of detecting phishing attacks in real-time and

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

advocate for the integration of advanced techniques, such as deep learning, into existing security systems. The introduction of PhiKitA, as outlined in this project, contributes to this body of research by providing a curated dataset specifically tailored for studying phishing kit attacks and improving detection capabilities.

V.MODULES

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

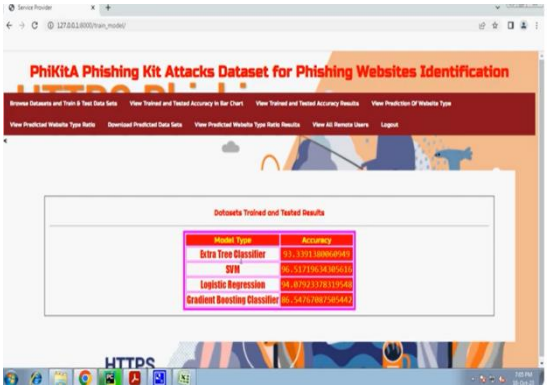
Login, Browse Datasets and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Status, View Status Ratio, Download Predicted Data Sets, View Ratio Results, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like register and login, predict detection,view your profile.



VI.ALGORITHMS

Decision tree classifiers

Decision tree classifiers are used successfully in many diverse areas.

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does. This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point *x* and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptrons*, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

VII.CONCLUSION

The development of PhiKitA, a comprehensive dataset containing phishing kits and associated websites, represents a significant contribution to the field of cybersecurity. Through the creation of PhiKitA, we have addressed the critical need for curated datasets to facilitate research and analysis of phishing kit attacks. By providing researchers with access to a diverse collection of phishing kits and associated websites, PhiKitA empowers them to study the tactics and techniques employed by attackers in greater detail. Our experiments with PhiKitA have yielded valuable insights into the effectiveness of various detection algorithms and highlighted the potential of graph representation and other advanced techniques in combating phishing threats. Moving forward, PhiKitA will serve as a valuable resource for researchers, security analysts, and practitioners seeking to enhance their understanding of phishing attacks and develop more effective countermeasures. With continued refinement and expansion, PhiKitA has the potential to significantly advance our ability to detect and mitigate phishing threats, ultimately contributing to a safer and more secure online environment for all users.

VIII.REFERENCES

I. T. Union, Measuring Digital Development: Facts and Figures, 2021, [online]

2. R. M. A. Mohammad, "A lifelong spam emails classification model", Appl. Comput. Informat., Jul. 2020, [online]

3. F. Já nez-Martino, E. Fidalgo, S. González-Martínez and J. Velasco-Mata, "Classification of spam emails through hierarchical clustering and supervised learning", arXiv:2005.08773, 2020.

4. J. Velasco-Mata, V. Gonzalez-Castro, E. F. Fernandez and E. Alegre, "Efficient detection of botnet traffic by features selection and decision trees", IEEE Access, vol. 9, pp. 120567-120579, 2021.

5. A. Mihoub, O. B. Fredj, O. Cheikhrouhou, A. Derhab and M. Krichen, "Denial of service attack detection and mitigation for Internet of Things using looking-back-enabled machine learning techniques", Comput. Electr. Eng., vol. 98, Mar. 2022.

6. E. Fidalgo, E. Alegre, L. Fernández-Robles and V. González-Castro, "Classifying suspicious content in Tor

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 16, Issue. 2, April 2024

darknet through semantic attention keypoint filtering", Digit. Invest., vol. 30, pp. 12-22, Sep. 2019, [online] Available: https://www.sciencedirect.com/science/article/pii/S1742287619300027.

7. P. Blanco-Medina, E. Fidalgo, E. Alegre and F. Janez-Martino, "Improving text recognition in Tor darknet with rectification and super-resolution techniques", Proc. 9th Int. Conf. Imag. Crime Detection Prevention (ICDP), pp. 32-37, 2019.

8. E. Figueras-Martín, R. Magán-Carrión and J. Boubeta-Puig, "Drawing the web structure and content analysis beyond the tor darknet: Freenet as a case of study", J. Inf. Secur. Appl., vol. 68, Aug. 2022.

9. C. A. Murty, H. Rana, R. Verma, R. Pathak and P. H. Rughani, "Building an AI/ML based classification framework for dark web text data", Proc. Int. Conf. Comput. Commun. Netw., pp. 93-111, 2022.

10. D. Chaves, E. Fidalgo, E. Alegre, R. Alaiz-Rodríguez, F. Já nez-Martino and G. Azzopardi, "Assessment and estimation of face detection performance based on deep learning for forensic applications", Sensors, vol. 20, no. 16, pp. 4491, 2020, [online] Available: https://www.mdpi.com/1424-8220/20/16/4491.

11. L. Zhu, Q. Zhang and W. Wang, "Residual attention dual autoencoder for anomaly detection and localization in cigarette packaging", Proc. Chin. Autom. Congr. (CAC), pp. 475-480, Nov. 2020.

12. S. Minocha and B. Singh, "A novel phishing detection system using binary modified equilibrium optimizer for feature selection", Comput. Electr. Eng., vol. 98, Mar. 2022.

13. E. Zhu, Z. Chen, J. Cui and H. Zhong, "MOE/RF: A novel phishing detection model based on revised multi-objective evolution optimization algorithm and random forest", IEEE Trans. Netw. Service Manage., vol. 19, no. 4, pp. 4461-4478, Dec. 2022.

14. Phishing Activity Trends Report 2 Quarter, 2022, [online] Available: https://apwg.org/trendsreports.

15. M. Hijji and G. Alam, "A multivocal literature review on growing social engineering based cyber-attacks/threats during the COVID-19 pandemic: Challenges and prospective solutions", IEEE Access, vol. 9, pp. 7152-7169, 2021.