E004 Diksha Maheshwari

E012 Shaheen Mondal

E016 Nilima Pai

Prof. Ameyaa Biwalkar

Mukesh Patel School of Technology Management and Engineering, NMIMS

MACHINE LEARNING

Detection of Phishing Websites using Machine Learning

27th August, 2019

OUTLINE

- Introduction and Problem definition
- Literature review
- UML Diagrams
- Algorithms
- Implementation Plan and Datasets
- Action plan for rest of the Project
- Conclusion & Future work
- References

INTRODUCTION

- The word ‘phishing’ is a variation on the word ‘fishing’. The idea is that bait (email, phone call or text message) is thrown out hoping that a user will grab it just like the fish.
- Targets are lured by attackers into providing sensitive data such as personally identifiable information, banking and credit card details, and passwords.
- The personal information obtained by the phisher is used to access important accounts and can result in identity theft and financial loss.
- According to the reports released by Anti-Phishing Working Group, the number of unique phishing sites reported till September 2018 were 647,592.
- Our goal is to find out drawbacks in existing phishing detection tools and improve upon them.

PROBLEM DEFINITION

- Phishing is an example of a social engineering technique being used to deceive users by exploiting human psychology.
- Phishing websites have certain attributes that can be used as features for effective means of fraudulent website detection.
- Our approach is to identify phishing websites by analyzing various criteria like the URL, website favicon, layout, and other features from existing datasets.
- Most existing approaches make use of classification algorithms; we plan to use clustering techniques instead because of some inherent benefits that clustering provides for larger datasets.

LITERATURE REVIEW (1 OF 3)

Name	Description and Algorithms used	Accuracy	Features and Drawbacks
Phishing Alarm [1]	It describes an algorithm to quantify the suspiciousness ratings of Web pages based on their similarity. Their approach uses CSS as the basis to accurately quantify the visual similarity of each page element based on a rating method of weighted page-component similarity.	99.74 %	It is language independent. Consists of CSS pre-processing, causing significant overhead. Same final layout can be achieved by using different CSS properties, or even by using inline or internal CSS.
AntiPhishing Framework [9]	They have compared 5 approaches of detection. They have mentioned some features which are available in datasets like '@' symbol, right click disabling and suspicious link redirection which are not valid today.	99.60 %	Will not work with a newly added spam feature, since it works on predefined thresholds.
Machine learning approach [8]	Three approaches to detect phishing websites- by URL, by checking the legitimacy of website and by using visual appearance based analysis.	96.58 %	Hybrid model combining all the three approaches Random Forest gives the most accuracy.

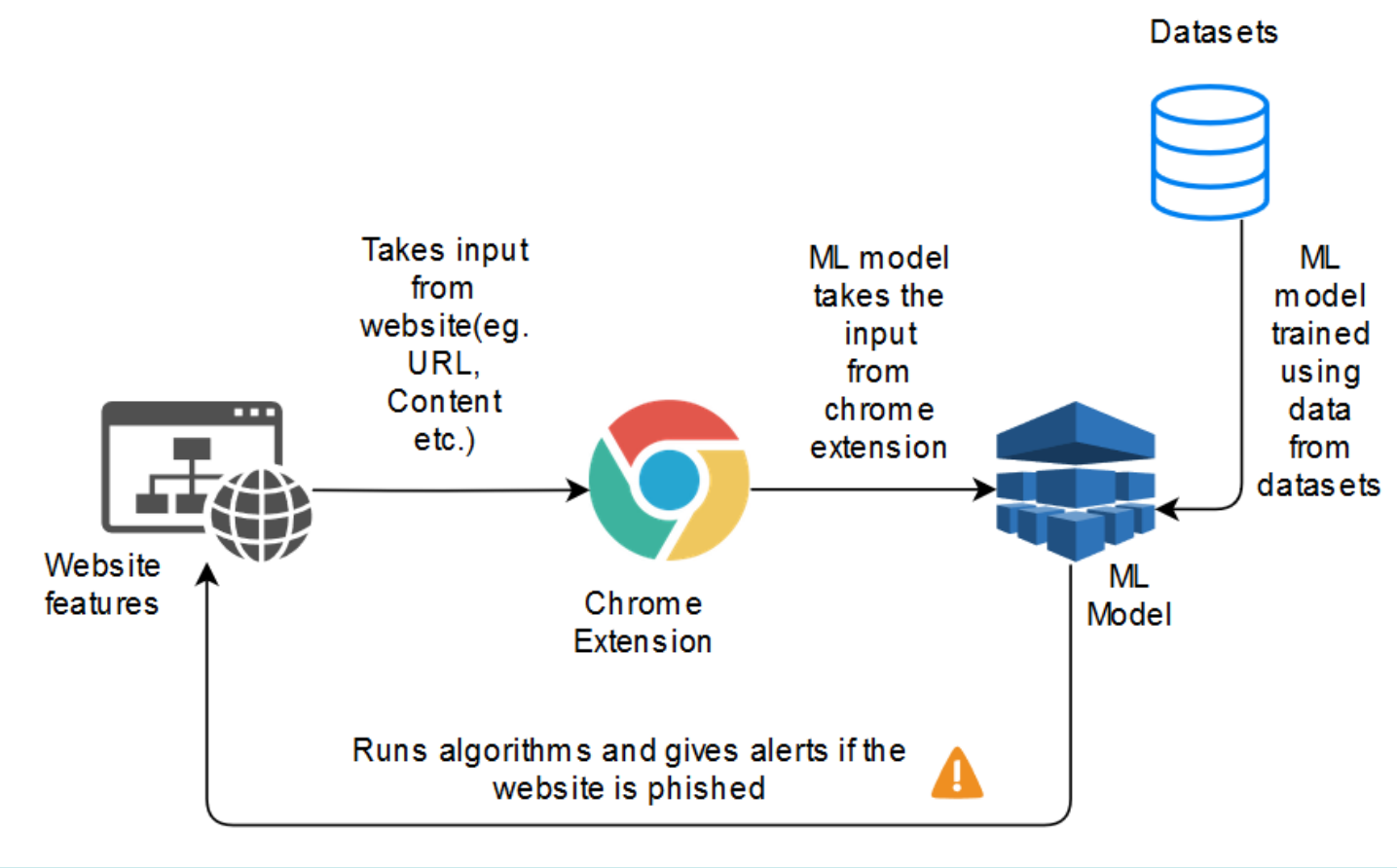
LITERATURE REVIEW (2 OF 3)

Name	Description /Algorithms used	Accuracy	Features and Drawbacks
Feature Selection and Dimensionality Reduction [11]	4 pre-processing techniques (CFS, IG, Consistency subset, PCA) were used along with 5 classification algorithms(RF, J48,NB,SVM,AdaBoost). Best combination was selected(RF and Consistency subset)	97.47 %	Feature selection algorithms identified 30 features. They grouped the best features among them.
Favicon [3]	Focuses on a tiny but powerful visual element– favicon, which is widely used by phishers but ignored by anti-phishing researchers.	99.5 %	Locates the suspicious brand sites, including legitimate and fake brands sites, and then PageRank and DNS filtering algorithm discriminates the sites with branding rights from fake brands sites.
Detection of Phishing Attacks [6]	Proposes a system that detects old as well as new phishing URLs using Data Mining.	97.2592 %	Classification model was used to extract attributes from URL to be used as input data. RF came out to be the best.

LITERATURE REVIEW (3 OF 3)

Name	Description /Algorithms used	Accuracy	Features and Drawbacks
Fuzzy data mining [10]	The solution suggested by the paper has three parts: blacklist, URL heuristics and CSS similarity; and uses classification algorithms. The heuristics for the phished websites have been derived by a detailed analysis of hyphen counts, length, etc.		It works even when the website characteristics are not definite.
Sender-Centric Approach [13]	A sender-centric approach to detecting phishing emails is used here. This approach was developed based on the observation that phishers can't completely conceal the sender information of a phishing message. Such sender information is often inconsistent with the target institution of the phishing email.	98.94 %	It has only focused on physical messages for banks. Cannot be used to clarify commerce sites.

Proposed DATA FLOW

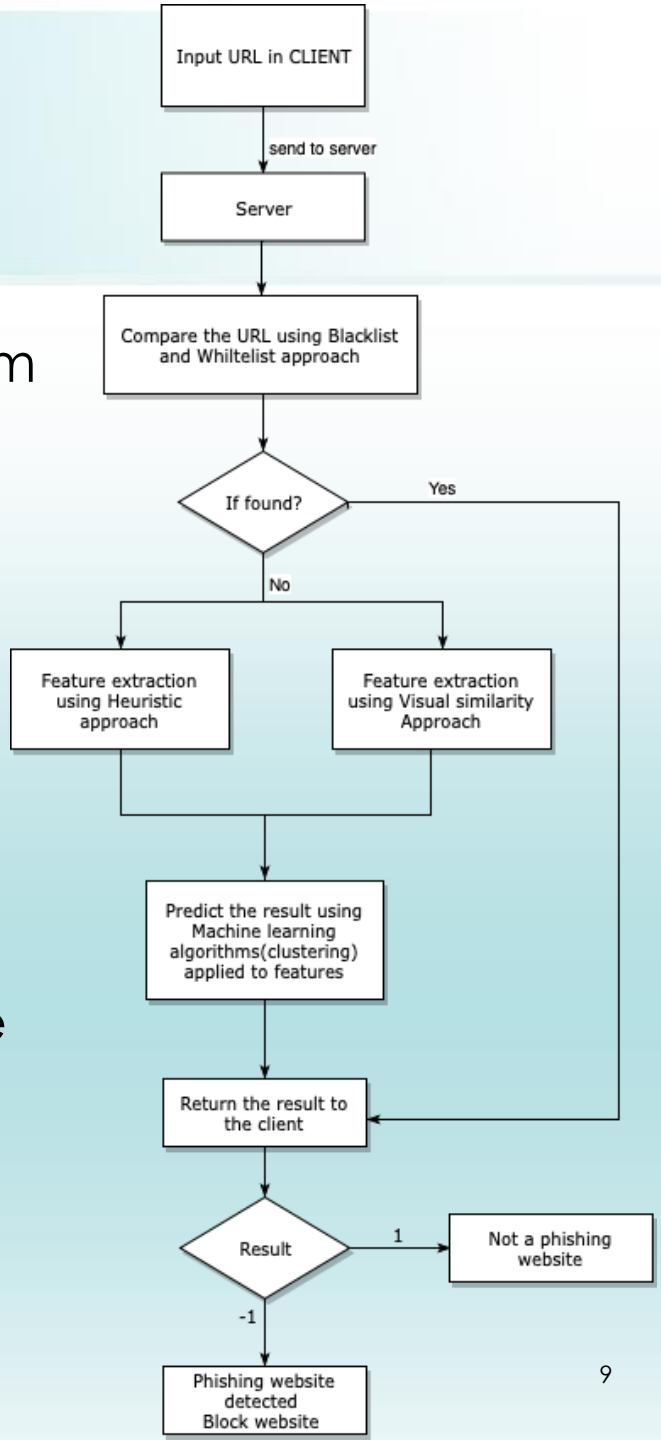


ALGORITHMS

1. Monitor all “http” traffic of end-user system by creating a browser extension.

2. Compare domain of each URL with the white-list of trusted domains and also the black-list of illegitimate domains.

3. Now the analysis will be done by using various features. Multiple features used together will give higher accuracy for the system.



ALGORITHMS

4. Attackers make the fake pages as similar to the original ones as possible to deceive the users of its authenticity . To counter this, we will extract and compare CSS of suspicious URL and compare it with the CSS of each of the legitimate domains in queue.

5. The machine learning clustering algorithms will be applied to the collected data and a score is generated.

6. If the score is above threshold, we mark it as Phished.

CLUSTERING ALGORITHM

- **Fuzzy C-means (FCM) Algorithm**

In this algorithm, each data point in a cluster has the probability of belonging to the other. Therefore, the data point does not have an absolute membership over a particular cluster. This is the reason the algorithm is named 'fuzzy'.

- **Hierarchical Clustering Algorithms**

These algorithms have clusters sorted in an order based on the hierarchy in data similarity observations. Hierarchical clustering is categorized into two types, divisive (top-down) clustering and agglomerative (bottom-up) clustering.

CLUSTERING ALGORITHM

- **Density-based spatial clustering of applications with noise (DBSCAN)**

Clusters are dense regions in the data space, separated by regions of the lower density of points. The DBSCAN algorithm is based on this intuitive notion of "clusters" and "noise". The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

IMPLEMENTATION PLAN (1 of 2)

- **Variable ranking**
  - Our dataset consists of 30 variables, of which some are not valid today [11] or not of much value in decision making.
  - These can be discarded using different variable ranking algorithms (Consistency Subset has highest accuracy [11]).

- **Implement multiple clustering algorithms, association rules**
  - Grouping the websites in such a way that objects in the same cluster are more similar to each other than to those in other clusters.

- **Comparison of classification and clustering accuracies**

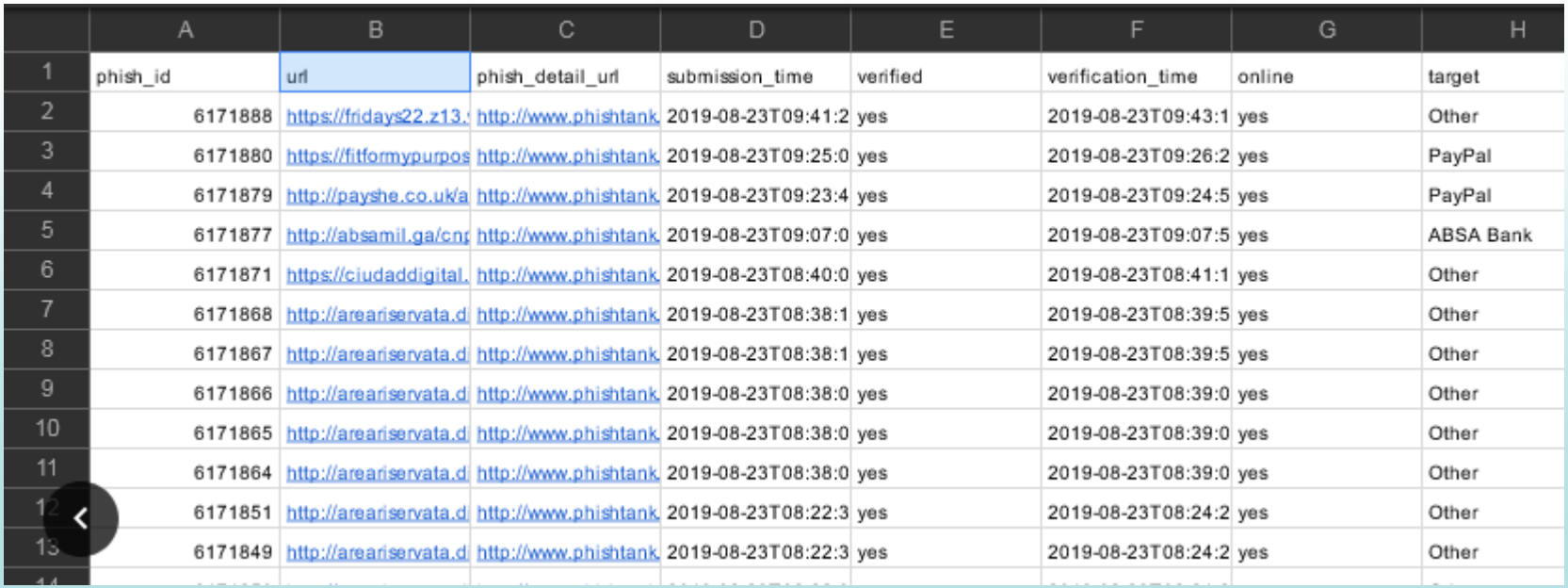
IMPLEMENTATION PLAN (2 of 2)

- **Select the best features from both**
  - Combine classification and clustering algorithms at different stages and utilize their features to further increase accuracy

- **Build interface**
  - Build a Chrome browser extension to enable user to interact with the application, and to send the necessary website details to the model, by monitoring the HTTP traffic.

DATASETS (1 of 2)

- Verified blacklisted websites



	A	B	C	D	E	F	G	H
1	phish_id	url	phish_detail_url	submission_time	verified	verification_time	online	target
2	6171888	https://fridays22.z13.	http://www.phishtank	2019-08-23T09:41:2	yes	2019-08-23T09:43:1	yes	Other
3	6171880	https://fitformypurpos	http://www.phishtank	2019-08-23T09:25:0	yes	2019-08-23T09:26:2	yes	PayPal
4	6171879	http://payshe.co.uk/a	http://www.phishtank	2019-08-23T09:23:4	yes	2019-08-23T09:24:5	yes	PayPal
5	6171877	http://absamil.ga/cnf	http://www.phishtank	2019-08-23T09:07:0	yes	2019-08-23T09:07:5	yes	ABSA Bank
6	6171871	https://ciudaddigital.	http://www.phishtank	2019-08-23T08:40:0	yes	2019-08-23T08:41:1	yes	Other
7	6171868	http://areariservata.d	http://www.phishtank	2019-08-23T08:38:1	yes	2019-08-23T08:39:5	yes	Other
8	6171867	http://areariservata.d	http://www.phishtank	2019-08-23T08:38:1	yes	2019-08-23T08:39:5	yes	Other
9	6171866	http://areariservata.d	http://www.phishtank	2019-08-23T08:38:0	yes	2019-08-23T08:39:0	yes	Other
10	6171865	http://areariservata.d	http://www.phishtank	2019-08-23T08:38:0	yes	2019-08-23T08:39:0	yes	Other
11	6171864	http://areariservata.d	http://www.phishtank	2019-08-23T08:38:0	yes	2019-08-23T08:39:0	yes	Other
12	6171851	http://areariservata.d	http://www.phishtank	2019-08-23T08:22:3	yes	2019-08-23T08:24:2	yes	Other
13	6171849	http://areariservata.d	http://www.phishtank	2019-08-23T08:22:3	yes	2019-08-23T08:24:2	yes	Other

- Contains 12269 records, to be used in first stage of detection

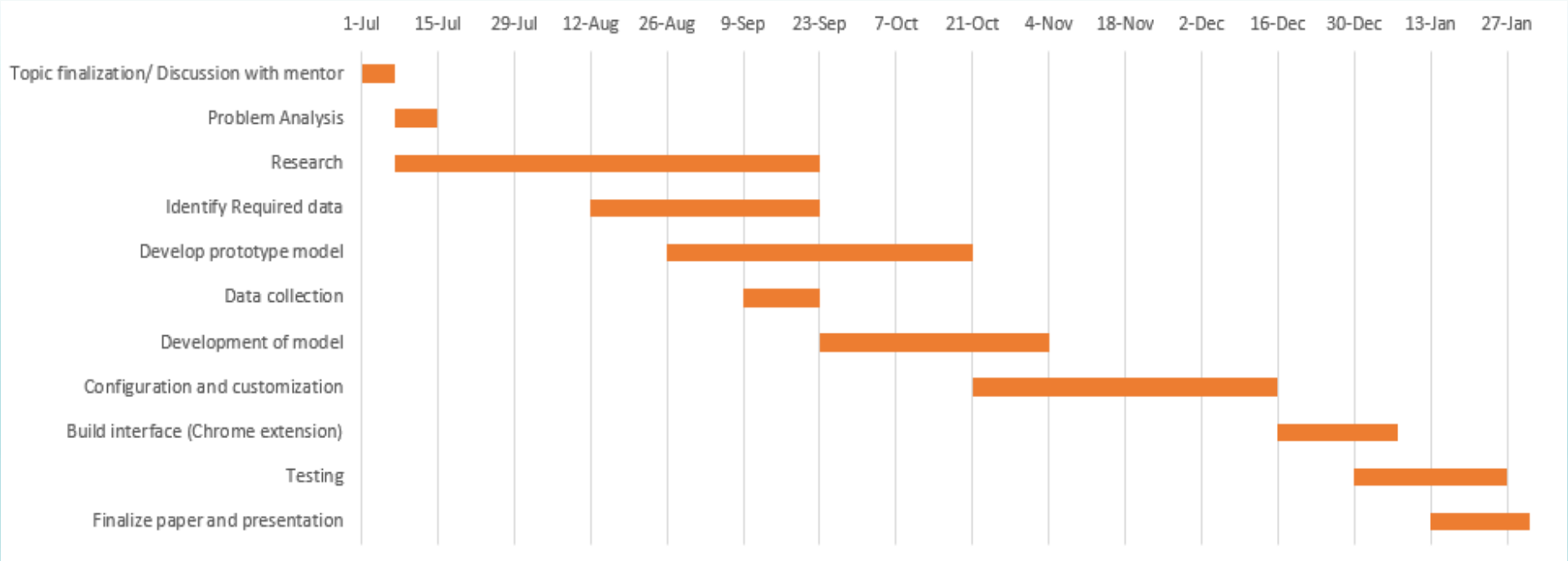
DATASETS (2 of 2)

- Heuristics dataset

T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
edirect	on_mouseover	RightClick	popUpWidnow	Iframe	age_of_domain	DNSRecord	web_traffic	Page_Rank	Google_Index	Links_pointing_to_pa	Statistical_report	Result
0	1	1	1	1	-1	-1	-1	-1	1	1	-1	-1
0	1	1	1	1	-1	-1	0	-1	1	1	1	-1
0	1	1	1	1	1	-1	1	-1	1	0	-1	-1
0	1	1	1	1	-1	-1	1	-1	1	-1	1	-1
0	-1	1	-1	1	-1	-1	0	-1	1	1	1	1
0	1	1	1	1	1	1	1	-1	1	-1	-1	1
0	1	1	1	1	1	-1	-1	-1	1	0	-1	-1
0	1	1	1	1	-1	-1	0	-1	1	0	1	-1
0	1	1	1	1	1	-1	1	1	1	0	1	1
0	1	1	1	1	1	-1	0	-1	1	0	1	-1
0	1	1	1	1	-1	1	1	1	1	-1	-1	-1
0	1	1	1	1	-1	-1	-1	-1	1	0	-1	-1
0	-1	1	-1	1	1	-1	-1	-1	1	0	1	-1
0	1	1	1	1	-1	-1	0	-1	1	1	1	-1
0	1	1	1	1	1	-1	1	-1	1	-1	1	1
0	1	1	1	1	1	-1	-1	-1	1	0	1	-1
0	1	1	1	1	1	-1	0	-1	1	1	-1	-1
0	1	1	1	1	-1	1	1	-1	1	1	-1	-1
0	1	1	1	1	1	-1	-1	1	1	-1	-1	1
0	-1	-1	-1	-1	1	-1	0	-1	1	0	-1	1
0	-1	1	-1	1	-1	1	1	-1	1	-1	-1	1
0	1	1	1	1	-1	1	-1	-1	1	0	-1	1
0	1	1	1	1	1	1	0	-1	1	-1	1	1

- 30 variables (to be reduced by variable ranking), 11055 records

ACTION PLAN FOR THE REST OF THE PROJECT



CONCLUSION AND FUTURE WORK

- Today, every country is aiming for cashless transactions, online business, paperless tickets, etc. to upgrade with the world. Phishers are targeting payment industry and cloud services the most. But phishing is becoming an obstacle.
- We reviewed various anti-phishing approaches. All methods are discussed to give a clear idea of existing techniques, their limitations and possible improvements.
- We plan to describe the most important steps to build an efficient anti-phishing model with the help of the algorithm diagram.
- We will compare the models using all the 5 types of approaches based on the number of features used, accuracy and size of dataset, and build a final model.

REFERENCES

1) J. Mao, W. Tian, P. Li, T. Wei and Z. Liang, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity," in IEEE Access, vol. 5, pp. 17020-17030, 2017, 24-07-2019
2) R. Atat, L. Liu, J. Wu, G. Li, C. Ye and Y. Yang, "Big Data Meet Cyber-Physical Systems: A Panoramic Survey," in IEEE Access, vol. 6, pp. 73603-73636, 2018, 24-07-2019
3) Guang-Gang Geng, Xiao-Dong Lee, Wei Wang and Shian-Shyong Tseng, "Favicon - a clue to phishing sites detection," 2013 APWG eCrime Researchers Summit, San Francisco, CA, 2013, pp. 1-10, 23-07-2019
4) R. Aravindhan, R. Shanmugalakshmi, K. Ramya and Selvan C., "Certain investigation on web application security: Phishing detection and phishing target discovery," 2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, 2016, pp. 1-10, 24-07-2019
5) A. Madaan, X. Wang, W. Hall and T. Tiropanis, "Observing data in IoT worlds: What and how to observe?," Living in the Internet of Things: Cybersecurity of the IoT - 2018, London, 2018, pp. 1-7, 24-07-2019
6) https://www.cs.nmt.edu/~rbasnet/research/DetectionOfPhishingAttacks.pdf, 24-07-2019
7) M. Thaker, M. Parikh, P. Shetty, V. Neogi and S. Jaswal, "Detecting Phishing Websites using Data Mining," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 1876-1879, 24-07-2019
8) V. Patil, P. Thakkar, C. Shah, T. Bhat and S. P. Godse, "Detection and Prevention of Phishing Websites Using Machine Learning Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, 24-07-2019
9) S. Patil and S. Dhage, "A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 588-593. , 16-08-2019
10) Aburrous, Maher, Hossain, Alamgir, Dahal, Keshav and Thabtah, Fadi (2010) Intelligent phishing detection system for e-banking using fuzzy data mining. Journal of Expert Systems with Applications, 37 (12). pp. 7913-7921. ISSN 0957-4174, 16-08-2019
11) Singh, Pradeep & Jain, Niti & Maini, Ambar. (2015). Investigating the effect of feature selection and dimensionality reduction on phishing website classification problem. 388-393. 10.1109/NGCT.2015.7375147. , 16-08-2019
12) https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5, 16-08-2019
13) F. Sanchez and Z. Duan, "A Sender-Centric Approach to Detecting Phishing Emails," *2012 International Conference on Cyber Security*, Washington, DC, 2012, pp. 32-39.