Poster: PhishLex: A Proactive Zero-Day Phishing Defence Mechanism using URL Lexical Features

Matheesha Fernando, Abdun Naser Mahmood, Mohammad Jabed Morshed Chowdhury
Department of Computer Science and Information Technology
La Trobe University
Australia

***Abstract*—Google reports that 68% of all Phishing URLs that are blocked by them are zero-day phishing attacks that remain undetected using traditional blacklist-based approaches. Machine learning-based (ML) techniques can improve the accuracy of detecting zero-day attacks. However, a key limitation of current ML-based approaches is the lack of quality datasets to train the ML models. Existing publicly available phishing datasets are outdated, limited in size and often depend on third-party services. The latency in third-party look-ups and delay in registering potential phishing URLs in blacklist databases are prohibitive for anti-phishing solutions to be used in standalone or real-time detection scenarios. To address these issues, we have designed new lexical features, created a new dataset using the latest Phishing URLs, and trained a predictive model (PhishLex). Experimental evaluation demonstrates that PhishLex outperforms the state-of-the-art techniques by achieving higher accuracy (97%) and lower false negative rate (0.27%). Furthermore, we have tested PhishLex on zero-day phishing attacks with rolling validations against Google Safe Browsing. Our experiments show 95% phishing detection rate can be achieved for zero-day phishing. We have published the PhishLexURL phishing dataset with 114 lexical URL features on Github which will help researchers to train their model without relying on third-party look-ups.**

I. Introduction

Phishing is the act of stealing sensitive user data (e.g. username, password, social security number) by disguising as a legitimate entity [1]. Phishers often lure users to click on a link (URL) to a counterfeit website of the targeted organization which asks for user's sensitive information [2]. Despite the increasing preventive measures, phishing threats are rising exponentially and costs billions of dollars every year [10]. All the publicly available phishing prevention methods (alert tools, browser warnings, user awareness programs) are blacklist based [7]. There are a several public blacklisting and reporting sites such as Google Safe Browsing list[1], PhishTank.com[2], Total AV[3] and ScamWatch[4]. However, blacklists are a reactive approach to phishing prevention [3] as users are vulnerable to attacks until the URLs are detected, reported and registered for reference. Furthermore, many malicious sites/URLs are not blacklisted either because they are new, short-lived, never evaluated or were incorrectly evaluated. In the current phishing landscape, there is an average gap of 9 hours between the first victim visit and detection [6]. Researchers also have identified that there is an average 7 hours lapse between detection and peak mitigation by browser-based warnings, which gives an average of 16 hours for phishers to achieve their goals [6]. Even after mitigation, Phishers can still continue by changing the phishing URL with a simple character in either the sub-domain, path or query.

There is an extensive amount of research conducted in machine learning domain to detect phishing. However, research suggests that the automatic classification of phishing web pages is limited to experimental systems and not in active use [10][7]. We have identified few factors that make these research outcomes less reliable in zero-day phishing detection. Firstly, most of the researches have used either a self-collected small dataset of phishing and legitimate URLs or a previously collected and outdated dataset (not updated since published) to train their models. Secondly, researchers often rely on third-party services and database look-ups (e.g. ASN, Geolocation, Google Page Quality Score, Google page ranking, Alexa ranking, URL reputation checks, WHOIS look-ups and DNS history look-up) that introduce latency and require Internet access. Thirdly, researchers mostly use a combination of surface features from multiple sources such as URL, domain, host, page content and metadata which can be easily replicated by phishers.

Therefore, we are motivated to find a proactive, zero-day and standalone phishing detection approach using the lexical features from the URL. By proactive, we mean, users/crawlers do not need to click/visit the URL for phishing detection. Secondly, by zero-day, we mean, our mechanism can detect phishing even if that URL signature is not previously flagged as phishing. Thirdly, by standalone, we mean, our method does not rely on third-party calls, so it can provide protection in real-time without any latency. The main contributions of the paper are as follows:

- Proposed new lexical features and modified existing lexical features (114 features) to be able to detect new generation of phishing attacks with unknown signature.
- Created PhishLexURL2021 dataset, which is a contemporary dataset with 106,750 unique URLs using proposed 114 feature-set for phishing detection.
- Developed PhishLex using the proposed feature set which outperforms the existing lexical based proactive phishing detection.
- PhishLex can predict zero-day phishing urls with average 95% accuracy when we compare against Google Safe Browsing blacklist[5] which takes on average 24-48hrs to confirm a zero-day Phishing URL. In other words,

[1]https://safebrowsing.google.com
[2]https://phishtank.com
[3]https://www.totalav.com/features
[4]https://www.scamwatch.gov.au/report-a-scam
[5]https://safebrowsing.google.com/

PhishLex can accurately predict an unknown URL is phishing or not in first encounter way before it gets added into Google's Safe Browsing blacklist.

Table I presents distinctive features of PhishLex against the state-of-the art, lexical feature based phishing detection approaches in literature.

Characteristic	PhishLex	[4]	[8]	[11]	[3]	[5]
3rd party independence	True	True	False	False	False	False
Doesn't require Internet	True	True	False	False	False	False
Run-time efficiency	High	High	High	Low	Low	Low
Zero-day detection Test	True	False	False	False	False	False
Low false negative rate	True	True	-	-	False	False
Use of URL lexical features	True	True	True	False	True	True
Use of content/host based features	False	False	False	True	True	True
Use of contemporary dataset	True	False	True	False	True	True

TABLE I. CHARACTERISTIC COMPARISON WITH RELATED WORK

II. METHODOLOGY

The latency in third-party look-ups and delay in registering potential phishing URLs in blacklist databases are prohibitive for anti-phishing solutions to be used in standalone or real-time detection scenarios. To address this issue, we propose a set of new lexical features, and generate a dataset using the latest Phishing URLs in order to train a predictive model called PhishLex. Figure 1 presents the proposed approach for the phishing detection system, PhishLex.

For URL lexical features, we identified that, some features yield different values based on the component of the URL they are belong to. Therefore, we considered the URL component locality based feature extraction process. We collected a large contemporary URL dataset with both phishing and benign URLs (106750 URLs) and extracted the lexical features. Phishing URLs were collected from two sources; PhishTank.com [6] and openphish.com [7] using scheduled script to download latest phishing dataset every 24 hrs. Alexa.com top domains[8] and CommonCrawl[9] dataset was used to compose our benign dataset. We used this dataset of new features identify the best algorithm with 12 classifier algorithms to train and test a ML model for zero-day phishing detection. Next we evaluated the proposed PhishLex ML model against three benchmark lexical feature-based techniques[4][8][9]. Finally, we evaluated the prediction performance of PhishLex against Google Safe Browsing blacklist for zero-day phishing detection.

III. CONCLUSION

We proposed a novel approach to unleash the full potential of URL lexical features for proactive phishing detection. We described the feature extraction methods for collecting URLs and generating 114 URL features which resulted in a new dataset containing over 100K phishing URLs. We have published this dataset for the machine learning community. This paper also presented the results of our experiments which shows the potential of a proactive lexical feature based phishing detection technique compared to other techniques.

[6]https://phishtank.com/developerinfo.php
[7]https://openphish.com
[8]https://www.alexa.com/topsites
[9]https://registry.opendata.aws/commoncrawl/

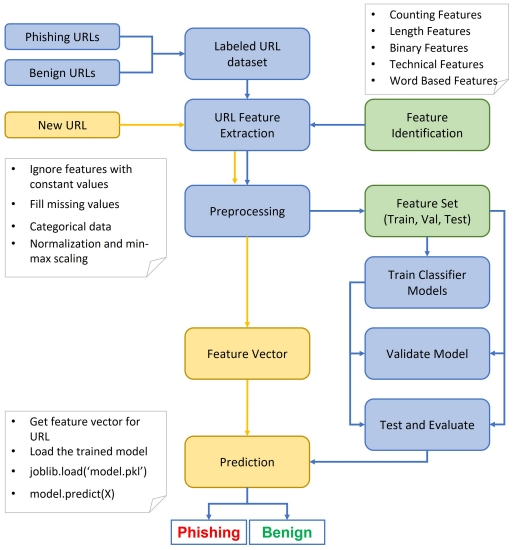


Fig. 1. Proposed Approach for PhishLex zero-day phishing detector

REFERENCES

[1] APWG. Phishing Activity Trends Report. Technical report, APWG, 2021. https://apwg.org/trendsreports/.

[2] Matheesha Fernando and Nalin Asanka Gamagedara Arachchilage. Why Johnny cant rely on anti-phishing educational interventions. In *ACIS 2019 Proceedings*, page 11. ACIS, 2019. https://acis2019.io/paper/83/.

[3] Sujata Garera, Niels Provos, Monica Chew, and Aviel D. Rubin. A framework for detection and measurement of phishing attacks. In *ACM workshop on Recurring malcode (WORM '07)*, page 1, 2007.

[4] Brij B. Gupta, Krishna Yadav, Imran Razzak, Konstantinos Psannis, Arcangelo Castiglione, and Xiaojun Chang. A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. *Computer Communications*, 175:47–57, July 2021. https://www.sciencedirect.com/science/article/pii/S0140366421001675.

[5] Sophie Le Page, Guy-Vincent Jourdan, Gregor V. Bochmann, Iosif-Viorel Onut, and Jason Flood. Domain Classifier: Compromised Machines Versus Malicious Registrations. In Maxim Bakaev, Flavius Frasincar, and In-Young Ko, editors, *Web Engineering*, Lecture Notes in Computer Science, pages 265–279, Cham, 2019. Springer International Publishing.

[6] Adam Oest, Penghui Zhang, Brad Wardman, Eric Nunes, Jakub Burgis, Ali Zand, Kurt Thomas, Adam Doupé, and Gail-Joon Ahn. Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. pages 361–377.

[7] Issa Qabajeh, Fadi Thabtah, and Francisco Chiclana. *A recent review of conventional vs. automated cybersecurity anti-phishing techniques*, volume 29. Elsevier, August 2018. https://www.sciencedirect.com/science/article/pii/S1574013717302010.

[8] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, and Banu Diri. Machine learning based phishing detection from URLs. *Expert Systems with Applications*, 117:345–357, March 2019.

[9] Martyn Weedon, Dimitris Tsaptsinos, and James Denholm-Price. Random forest explorations for URL classification. In *2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, pages 1–4. IEEE, June 2017. http://ieeexplore.ieee.org/document/8073403/.

[10] Colin Whittaker, Brian Ryner, and Marria Nazif. Large-Scale Automatic Classification of Phishing Pages. In *NDSS '10*, 2010.

[11] Guang Xiang, Jason Hong, Carolyn P. Rose, and Lorrie Cranor. CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites. *ACM Transactions on Information and System Security (TISSEC)*, 14(2):21:1–21:28, September 2011. https://doi.org/10.1145/2019599.2019606.

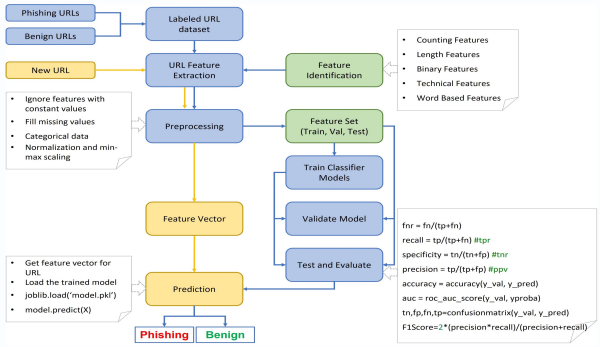
Poster: PhishLex: A Proactive Zero-Day Phishing Defence Mechanism using URL Lexical Features

Matheesha Fernando, Abdun Naser Mahmood, Mohammad Jabed Morshed Chowdhury
Department of Computer Science and Information Technology
La Trobe University, Australia

PROBLEM INTRODUCTION

- Google reports that 68% of all Phishing URLs are zero-day phishing attacks [1], that remain undetected using traditional blacklist-based approaches.
- Currently there is an average gap of 9 hours between the first victim visit and phishing detection, while there is an average 7 hours lapse between detection and peak mitigation by browser-based warnings, which gives an average of 16 hours for phishers to achieve their goals [2].
- ML-based approaches to detect zero-day attacks is constrained by the existing phishing datasets, which are outdated, limited in size and often depend on third-party services.

ARCHITECTURE



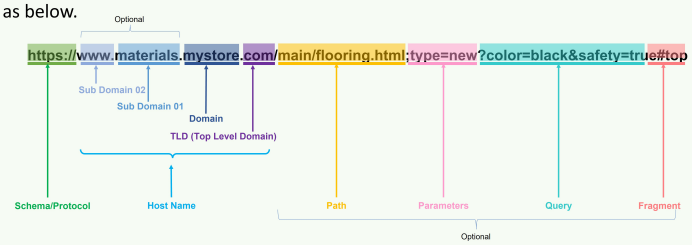
PHISHLEX URL DATASET

We created a contemporary dataset of 106750 URLs (both phishing and benign) using the 3rd party independent features we identified.

URL Type	Source	Count	Percentage
Phishing	phishtank.com and OpenPhish	55500	56.4%
Benign	Common Crawl with Alexa.com	51250	43.6%
		106750	100%

URL LEXICAL FEATURES WITH COMPONENT LOCALITY

For URL lexical features, we identified that, some features yield different values based on the component of the URL they are belong to. Therefore, we considered the URL component locality-based feature extraction process. We have extracted 114 features as below.



Url_dot_count	Url_length	path_question_mark_count	filename_hypen_underscore_count	query_double_slash_count	has_file_name
Url_hypen_count	Url_has_username	path_equal_count	query_len	query_semicolon_count	has_punny_code
Url_underscore_count	Url_has_password	path_at_count	query_argument_count	arg_delimeters_count	homoglyphs_count
Url_slash_count	domain_dot_count	path_and_count	query_max_arg_len	having_ip_address	has_suspicious_tld
Url_question_mark_count	domain_hypen_count	path_exclamation_count	query_hypen_count	shortening_service	has_fragment
Url_equal_count	domain_length	path_space_count	query_underscore_count	email_in_url	average_word_length
Url_at_count	domain_digit_count	path_tilde_count	query_plus_count	number_of_domains	longest_word_length
Url_and_count	domain_hex_count	path_comma_count	query_slash_count	common_keyword_count	shortest_word_length
Url_exclamation_count	domain_has_non_std_port	path_plus_count	query_question_mark_count	has_url_redirect	std_word_length
Url_space_count	domain_has_hypen	path_asterisk_count	query_equal_count	subdirectory_count	adjacent_word_count
Url_tilde_count	domain_longest_token_len	path_hashtag_count	query_at_count	has_port	adjacent_avg_len
Url_comma_count	sub-domain_dot_count	path_dollar_count	query_and_count	has_query	random_domain
Url_plus_count	sub-domain_hypen_count	path_percent_count	query_exclamation_count	brandname_in_sub-domain	has_known_tld
Url_asterisk_count	sub-domain_underscore_count	path_double_slash	query_space_count	brandname_in_path	character_repeat
Url_hashtag_count	sub-domain_plus_count	path_semicolon_count	query_tilde_count	brandname_count	
Url_dollar_count	sub-domain_length	path_length	query_comma_count	similar_keyword_count	
Url_percent_count	path_dot_count	path_longest_token	query_asterisk_count	similar_brandname_count	
Url_double_slash	path_hypen_count	path_delimeters_nondot_count	query_hashtag_count	seperated_word_count	
Url_semicolon_count	path_underscore_count	filename_len	query_dollar_count	other_words_count	
Url_digit_count	path_slash_count	filename_dot_count	query_percent_count	random_words_count	

CONCLUSION

PhishLex can predict zero-day phishing URLs with unknown signature with average of 95% accuracy in real-time while Google Safe Browsing blacklist takes on average three days to alert the phishing. We have published a contemporary phishing dataset for the machine learning community with over 100K URLs for training ML models to detect Zero-day phishing attacks with unknow signatures.

MOTIVATION

We are motivated to find a realistic phishing detection approach using the lexical features from the URL with following features:

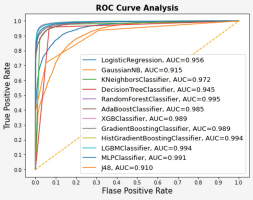
Proactive: users/crawlers do not need to click/visit the URL for phishing detection
Zero-day: can detect phishing even if that URL signature is not previously flagged
Standalone: does not require third party services/lookups or even the Internet

CONTRIBUTIONS

- Proposed 114 lexical feature combination to detect phishing attacks with unknown signature.
- Created PhishLexURL dataset: a contemporary dataset with 106,750 unique URLs to train ML model to detect zero-day phishing
- Developed PhishLex using the proposed feature set which outperforms the existing lexical based proactive phishing detection.

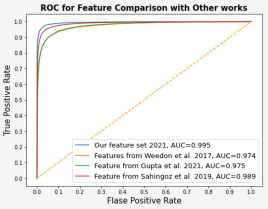
EVALUATION AND RESULTS

- We conducted 10 sets of repeated algorithm evaluations, for 12 machine learning classifiers including all the classifiers used in comparison studies of previous work as Logistic Regression, Gaussian Naive Bayes, K Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, XGB Classifier, Gradient Boosting Classifier, Hist Gradient Boosting Classifier, LGBM Classifier, MLP Classifier, and J48.
- The dataset was divided into 70:20:10 for the experiments
- We conducted a feature-based comparison against the benchmark techniques in literature [3][4][5].



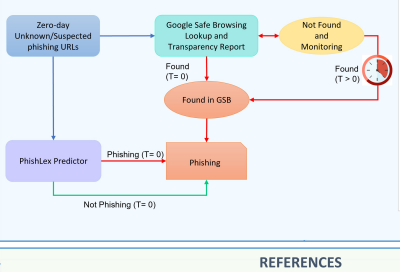
Classifiers	Accuracy	AUC	FNR	Recall	Specificity	Precision	F1score
AdaBoost Classifier	0.9432	0.9862	0.0544	0.9456	0.9395	0.9598	0.9527
Decision Tree Classifier	0.9495	0.9473	0.0412	0.9588	0.9352	0.9577	0.9583
Gaussian NB	0.6757	0.9164	0.5167	0.4833	0.9699	0.9609	0.6431
Gradient Boosting Classifier	0.9496	0.9894	0.0465	0.9535	0.9435	0.9627	0.9581
Hist Gradient Boosting Classifier	0.9644	0.9943	0.0333	0.9667	0.9609	0.9743	0.9705
J48	0.8381	0.9091	0.0637	0.9363	0.6878	0.8210	0.8749
KNeighbors Classifier	0.9397	0.9705	0.0591	0.9409	0.9377	0.9585	0.9496
LGBM Classifier	0.9649	0.9944	0.0327	0.9673	0.9613	0.9745	0.9709
MLP Classifier	0.9572	0.9918	0.0389	0.9611	0.9511	0.9678	0.9644
Random Forest Classifier	0.9700	0.9953	0.0274	0.9726	0.9661	0.9777	0.9751
XGB Classifier	0.9506	0.9889	0.0495	0.9505	0.9507	0.9672	0.9588
Logistic Regression	0.8927	0.9560	0.0871	0.9129	0.8619	0.9100	0.9114

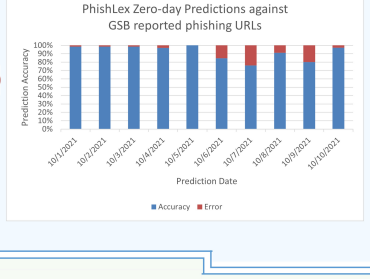
Study		Worst Case	Best Case	Average Case
PhishLex	Decision Tree Classifier	0.948053	0.949497	0.948972
	MLP Classifier	0.954705	0.957155	0.955794
	Random Forest Classifier	**0.968928**	**0.970022**	**0.969501**
Gupta et al., 2021 [16]	Decision Tree Classifier	0.901313	0.902101	0.901755
	MLP Classifier	0.886608	0.894398	0.890954
	Random Forest Classifier	**0.923851**	**0.925777**	**0.924696**
Sahingoz et al., 2019 [32]	Decision Tree Classifier	0.924085	0.926185	0.925511
	MLP Classifier	0.940248	0.944449	0.93621
	Random Forest Classifier	**0.951568**	**0.952327**	**0.951239**
Weedon et al., 2017 [44]	Decision Tree Classifier	0.894692	0.909983	0.909178
	MLP Classifier	0.901535	0.913254	0.908814
	Random Forest Classifier	**0.916762**	**0.936525**	**0.928841**

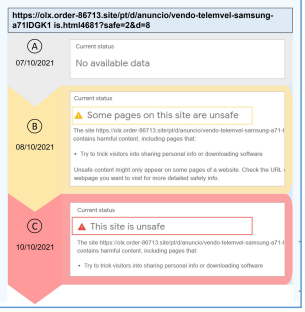


PhishLex - ZERO-DAY PHISHING ATTACK DETECTION

For our experiment we chose a 10-day observation period where we collected 100 unknown and online URLs from PhishTank.com that are reported on the same day. Then we evaluated them through our PhishLex predictor and reported the results. At the same time, we ran them through Google's Safe Browsing (GSB) and reported the outcome. We continued a rolling check on all Not Found URLs to observe when they get alerted in GSB. As of the 11th day, GSB status showed "Phishing" for 95% of the phishing URLs we detected in zero-day.







REFERENCES

[1] Google Security Blog. https://security.googleblog.com/2019 /08/understanding-why-phishing-attacks-are.html, Accessed 2021
[2] Adam Oest, Penghui Zhang, Brad Wardman, Eric Nunes, Jakub Burgis, Ali Zand, Kurt Thomas, Adam Doupé, and Gail-Joon Ahn. Sunrise to sunset: Analyzing the endto-end life cycle and effectiveness of phishing attacks at scale. [3] Brij B. Gupta, Krishna Yadav, Imran Razzak, Konstantinos Psannis, Arcangelo Castiglione, and Xiaojun Chang. A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. Computer Communications, 175:47–57, July 2021.
[4] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, and Banu Diri. Machine learning based phishing detection from URLs. Expert Systems with Applications, 117:345–357, March 2019.
[5] Martyn Weedon, Dimitris Tsaptsinos, and James Denholm-Price. Random forest explorations for URL classification. In 2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), pages 1–4. IEEE, June 2017