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 Existing publicly available phishing datasets are outdated, limited in size and often depend on third-party services.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/

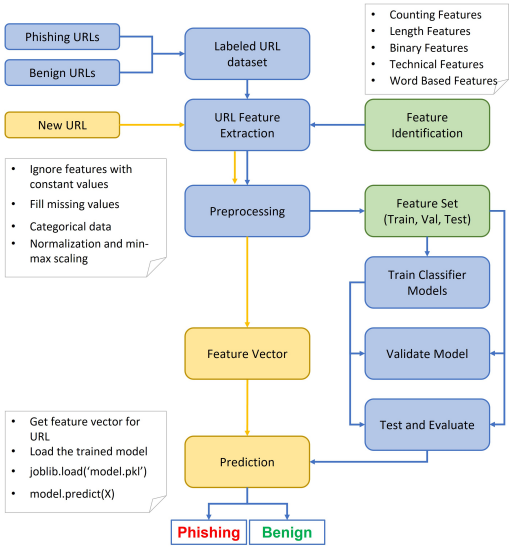

Fig. 1. Proposed Approach for PhishLex zero-day phishing detector

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.