Proactive Brand-Targeting Phishing Website Detection using a Hybrid Feature-based Approach with Machine Learning

Nadezda Demidova[1], Philip Lawson[1] and Jake Sloan[1]

[1] *EBAY (UK) LIMITED, 1 More London Place, London, United Kingdom, SE1 2AF*

Abstract

Phishing and online scam sites are on the rise, and the sophistication of these attacks continues to develop. Phishing websites exploit the target brand's identity, using its logo, website design, and reputation to trick customers into divulging sensitive information like login credentials and financial details. This, in turn, can cause financial losses, identity theft, and harm to the brand's reputation, ultimately eroding customer trust. Notably, the number of reported phishing attacks has grown more than five-fold in the last three years. Meanwhile, the number of brands attacked each month has remained relatively consistent. This forces businesses into a highly reactive, defensive mode, unable to get ahead of the problem, while exposing their customers and brand to abuse and financial loss. Moreover, the longer it takes for a business to identify and respond to an attack, the greater the potential damage to their reputation. To mitigate the impact of phishing attacks, businesses need to embrace proactive measures, moving away from purely responsive strategies and to addressing these threats as close to the source of the attack as possible.

Detecting threats that are targeting customers outside of a brand's platform and infrastructure can be challenging. The methods used for distributing phishing attacks are constantly evolving, with cybercriminals targeting new victims and the latest generation of internet users. In addition to classic email attacks, cybercriminals are now also using social networks and instant messaging platforms to reach potential victims, making it difficult for brands to identify and respond to these threats.

While many techniques for combating phishing attempt to address the issue broadly, our approach is focused specifically on brand protection and the abuse of brand assets no matter how a phishing website was distributed to potential victims. We use a combination of features based on URL structure and wording, DOM structure, HTML, and text content, that provide agility and adaptability, allowing us to more precisely detect a wider variety of brand-related phishing websites. These features enable Machine Learning algorithms to capture semantics and create a comprehensive high accuracy model capable of detecting phishing websites across multiple languages. Our approach delivers the proactive detection of classical phishing websites and scam-pages targeting a brand across a range of different scenarios and methods and can be easily adapted to suit the needs of any brand seeking to protect itself and its customers from phishing threats.

Keywords

Phishing, Machine Learning, Cybersecurity, Phishing detection

1. Introduction

According to the Anti-Phishing Working Group (APWG), the number of reported phishing attacks has grown more than five-fold in the last three years [1]. Meanwhile, the number of brands attacked each month has remained relatively consistent.

Phishing attacks have become increasingly sophisticated, posing a significant threat to businesses and their customers. In the typical lifecycle of a phishing URL, cybercriminals first establish their infrastructure, leading to the creation of deceptive phishing pages. Subsequently, phishing campaigns are initiated, attracting traffic to these malicious URLs. During this process, third-party vendors might detect the phishing activities and notify the targeted brands, enabling them to act and add the relevant information to their phishing collection for further investigation.

The time lag between the initiation of a phishing campaign and its detection poses a critical challenge for businesses. Customers remain exposed to phishing infrastructure outside the brand's platform, leading to potential financial losses, identity theft, and damage to the brand's reputation. To address this issue, we

APWG.EU Technical Summit and Researchers Sync-Up 2023, Dublin, Ireland, June 21 & 22, 2023

nadi.demidova@gmail.com (N. Demidova); plawson03@qub.ac.uk (P. Lawson); jsloan@red-button.com (J. Sloan)

0009-0002-9775-2729 (N. Demidova); 0009-0003-3107-5523 (P. Lawson); 0009-0009-5356-7573 (J. Sloan)

sought to develop a proactive approach that identifies phishing URLs and infrastructure earlier in the customer compromise cycle, effectively reducing the exposure time for our customers.

Our approach uses the concept of "Shift Left," emphasizing early identification of phishing assets. To achieve this, we created a custom Anti-Phishing Ecosystem tailored to the unique challenges faced by our brand. A custom solution allows us to leverage our in-depth brand knowledge, the understanding of our business workings, customers, and communication channels in the best possible way.

Machine Learning (ML) plays a central role in our custom solution. By harnessing ML capabilities, we gain a competitive advantage in staying ahead of evolving threats and detecting new zero-day attacks. ML allows us to be agile and adaptive, enabling swift responses to emerging phishing patterns. Additionally, we continue to leverage trusted external data sources and collaborate with valuable insights from partners to strengthen our approach further.

In this paper, we present our hybrid feature-based approach with machine learning for proactive brand-targeting phishing website detection. Our custom solution focuses on brand protection and the incorporation of internal signals and data sources, providing a comprehensive and highly accurate model capable of detecting a wide variety of brand-related phishing websites across multiple languages and distribution methods.

By implementing our proactive approach, businesses can fortify their defenses, protect their customers from phishing attacks, and safeguard their brand reputation. As cybercriminals continually refine their strategies, the need for early identification and agile detection becomes paramount in the fight against phishing threats. Our research aims to contribute to the evolving field of cybersecurity, empowering businesses to take a proactive stance against brand-targeting phishing attacks.

The paper has the following structure: Section 2 provides an overview of related work in the field, laying the foundation for our contributions. Moving forward, Section 3 outlines the key elements of our methodology, presenting our approach, system overview, and data collection methodology. In Section 4, we delve into the critical process of feature engineering, detailing how we transform raw data into insights. Section 5 introduces the models that power our detection system. To assess model performance, Section 6 elaborates on the evaluation metrics we have chosen. Section 7 is dedicated to presenting our results and review. Finally, in Section 8, we conclude with remarks that summarize our findings and pave the way for future research endeavors.

2. Literature review

The domain of phishing detection has been a focal point in cybersecurity research, driven by the increasing sophistication of cybercriminal activities. Researchers have proposed various machine learning-based solutions to tackle this pervasive threat, each with unique approaches and features. In this section, we review a selection of pertinent studies that contribute to the advancement of phishing detection methodologies.

One study by Das Guptta, S., Shahriar, K.T., Alqahtani, H. et al. [2] advances hybrid feature-based phishing website detection. The authors leverage URL and hyperlink features for real-time accuracy, minimizing reliance on third-party systems. This addresses the challenge of new websites and zero-hour attacks.

Q. A. Al-Haija and A. A. Badawi [3] propose an efficient phishing website detection system focusing on URL patterns. Machine learning techniques, including neural networks and decision trees, classify authentic and phishing sites effectively.

Arun Kulkarni, Leonard L. Brown III [4] delve into machine learning classifiers such as decision trees, Naive Bayesian classifier, SVM, and neural network to distinguish real from fake websites4. Real-world URL datasets exhibit their prowess.

Additionally, A. Ghimire, A. Kumar Jha, S. Thapa, S. Mishra and A. Mani Jha [5] champion a machine learning-driven approach detecting phishing URLs. Balanced datasets and varied algorithms reveal high precision, recall, and F-score potential.

S. Zaman, S. M. Uddin Deep, Z. Kawsar, M. Ashaduzzaman and A. I. Pritom [6] demonstrate the effectiveness of Naive Bayes, J48, and HNB classifiers in phishing detection. Innovative feature selection enhances accuracy.

Lastly, P. Yang, G. Zhao and P. Zeng [7] propose multidimensional feature-based phishing detection with deep learning. Character sequence features facilitate quick deep learning-based classification, complemented by URL statistics, webpage code, and text features.

In summary, the reviewed studies collectively contribute to the ongoing efforts in phishing detection using machine learning-based approaches. The variety of methodologies and feature sets underscores the need for adaptable and comprehensive solutions to counter the dynamic nature of phishing attacks.

This paper brings novelty by emphasizing brand-specific abuse, combining structural and textual features, and promoting the collection of compatible clean training samples for effective phishing detection.

3. Methods

3.1. Definitions and notations

Table 1
Definitions and notations

Term	Definition
URL	Address of a given unique resource on the Web
Phishing URL	Address of a phishing content on the Web
Document Object Model (DOM)	It defines the logical structure of documents and the way a document is accessed and manipulated

Page content	Captured web page source, when given phishing URL is requested in browser.
FQDN	Domain name that specifies its exact location in the tree hierarchy of the Domain Name System (DNS). It specifies all domain levels, including the top-level domain and the root zone.
TLD	Top level domain
Subdomains	All domains on the left of second-level domain
Path	The path refers to the exact location of a page, post, file, or other asset. It is often analogous to the underlying file structure of the website. The path resides after the hostname and is separated by "/" (forward slash).
Directories	Folder in a path (directory names separated by "/")
Parameters	*goes after "?" symbol.* Extra parameters provided to the Web server.
Anchor	Represents a sort of "bookmark" inside the web resource.

3.2. Approach

There is a common approach that underlies the Customer Compromised Cycle (Figure 1) and basic off-platform anti-phishing strategy:

1. Cybercriminal infrastructure setup
2. Phishing page creation
3. Phishing campaign launch
4. As campaigns gain momentum, third-party vendors identify and share this information.
5. This prompt notifications, add relevant data to our phishing-collection, and take necessary actions.

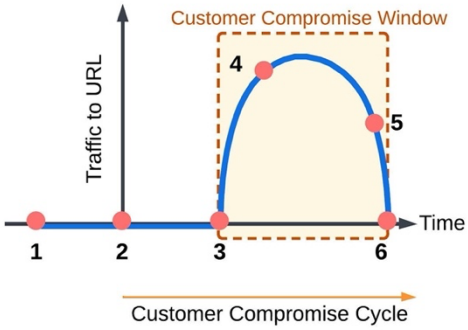

Figure 1: Customer Compromise Cycle

Our approach is designed to minimize our customers' exposure to off-platform phishing infrastructure and focuses on early identification of phishing assets. At its core, our solution integrates a machine learning model that plays a pivotal role in automating and scaling the phishing page detection process. This model utilizes a combination of features meticulously selected to ensure high accuracy in detection. A subset of these features revolves around the use of brand assets, aligning with our concentrated approach tailored for a specific brand. This synergy empowers our system to process large volumes of suspicious URLs from diverse sources, elevating our overall phishing detection efficacy.

The incorporation of machine learning allows us to proactively address evolving threats, including the detection of new zero-day attacks. This adaptability and agility are integral to staying ahead in the rapidly changing landscape of online security.

3.3. System overview

At a high level, our solution follows a streamlined workflow (Figure 2) to detect and mitigate phishing threats:

1. Data Collection: Our system actively collects URLs that exhibit suspicious characteristics from diverse sources. These sources encompass various avenues, including new domains, SSL Certificate stream data, our internal signals, and other repositories of potentially suspicious URLs.
2. Data Retrieval: From the gathered URLs, the system extracts the content of the web pages associated with these URLs.
3. Data Processing: Raw data is subjected to a comprehensive processing phase to derive meaningful data points that are conducive to effective phishing detection.
4. Feature Extraction: The system transforms the processed data points into a structured numeric representation.
5. Model Evaluation: Utilizing the numeric representation, our machine learning model takes over. It evaluates each sample and provides a verdict: whether the URL is indicative of phishing or not.
6. Action and Collection: If the model identifies a URL as phishing, we initiate an appropriate response.

This process has a feedback loop, as the insights gleaned from the collected data continuously contribute to the refinement and evolution of our machine learning model. This iterative approach ensures that our model remains adaptive to emerging trends and effectively addresses new challenges that may arise in the dynamic landscape of phishing threats.

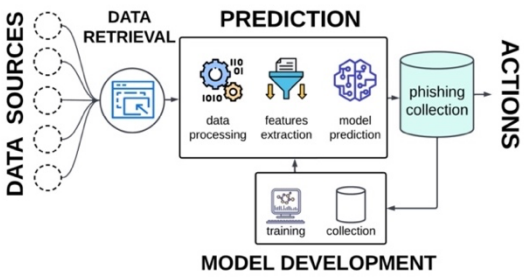

Figure 2: High-level System Architecture

3.4. Data collection

As our model makes its decisions based on features extracted from URLs and page content, this is what we needed to gather for our learning collection. To facilitate the training of our detection model, the acquisition of a substantial number of examples for each distinct targeted group was imperative. Presently, our dataset comprises more than 62,000 samples. Within the "phishing" category, we encompass a broad spectrum of deceptive pages designed to exploit our customers, ranging from traditional phishing schemes that replicate login interfaces to fraudulent support pages employing vishing tactics.

The efficacy of machine learning hinges on the caliber of the data it learns from, coupled with the algorithms' capacity to assimilate it. The quality of data exerts a direct influence on the model's performance; it can exclusively glean insights from the data it is provided. As such, the data must meet the following criteria:

- Relevance
- Non-duplication
- Accurate labeling
- A combination of recent and historical data
- Representative of real-world production scenarios
- Sourced from diverse origins.

Mitigating data selection bias is also paramount. This bias manifests when the collected data inadequately encapsulates the full spectrum of possible information or information combinations that the model may encounter in practical scenarios.

For instance, consider the analogy of fruits and vegetables. In reality, these come in a myriad of colors. However, if data collection predominantly focuses on red fruits and green vegetables, it would introduce data selection bias.

To ensure diversity within each targeted group, strategic sampling is essential. For instance, if top-level domains (TLDs) are employed as features and phishing samples are available for each TLD, solely featuring ".com" samples in the clean dataset could predispose the model to label anything else as phishing. Similarly, the nature of web pages must be adequately represented; relying solely on top ranked pages might not correlate well with our phishing dataset, which predominantly mimics login and registration forms. Therefore, a comprehensive collection is necessary to mirror analogous representations in the clean dataset.

Moreover, our solution leverages visible text prevalent in phishing pages, necessitating the accumulation of authentic instances that deploy similar terminologies without malicious intent. This includes legitimate pages from our customers who feature their businesses on our platform, personal websites, or articles about our brand. To bolster our clean dataset, we employed search engines with intelligent search queries to curate samples embodying diverse feature combinations. This strategic approach ensured that our clean dataset matched the multifaceted nature of the phishing collection we had amassed over time.

As a result, our model is equipped to discern nuanced patterns and characteristics in both phishing and legitimate content, enhancing its predictive accuracy in real-world scenarios.

For the evaluation of our model's performance, the dataset was divided into a training set and a test set in an 80:20 ratio, allowing us to assess its performance on previously unseen data.

4. Feature engineering

Feature engineering is a fundamental step in converting raw webpage data into numeric vectors that can be effectively utilized by machine learning algorithms for phishing detection. Since machine learning algorithms operate on numerical data, we need to find a suitable representation for each sample that provides valuable information to the model, enabling it to distinguish phishing instances effectively. In our solution, we adopt a hybrid feature-based approach, combining URL structure and wording, DOM structure, HTML, and text content to create numeric vectors for each webpage sample.

4.1. URL structure and rank features

The first type of features are URL-based features, such as the number of subdomains used, the count of path folders, and the domain's association with highly ranked domains based on the number of referring subnets.

This structural foundation, as depicted in Figure 3, forms the basis upon which our URL-based features are constructed.

Figure 3: Parts of a URL

As a result, the following features are extracted:

1. Does the root domain name rank within the top 1 million of widely recognized domains, based on the number of referring subnets?
2. Fully Qualified Domain Name (FQDN) rank – indicating whether the domain resides in the first 1000, first 100000, within the top 1 million, or outside this range.
3. Presence of brand-related keywords within the URL.
4. Count of path directories.
5. Count of subdomains.

4.2. Page content structure and links features

The second type of features we employ are based on the webpage's structure. Additionally, we analyse the links used on the page to identify any brand assets or links to the original brand logo. Since cybercriminals

often copy the original page, there is a high chance of finding traces left behind. Furthermore, we assess the DOM structure counts to determine the presence of forms and inputs on the page, contributing to effective phishing detection.

Figure 4: Snippet of page content with highlighted links

Figure 5: Snippet of page content with highlighted input, form, button elements

As a result, the following features are extracted:

1. Number of links in <link>/<script> /<img>/<a> tags (links to brand assets, links brand with keywords, non-brand related).
2. Number of inputs.
3. Number of forms.
4. Number of buttons.
5. Forms methods used (attribute specifies how to send form-data).
6. Use of original brand logo.

4.3. Tag names features

The third type of features revolves around unique ID names of HTML elements, class names, and form names. We extract them from html and map with dictionaries of the most frequent terms from phishing pages. By doing so, we create a linkage between these HTML element identifiers and common phishing patterns, enhancing the model's capability to identify suspicious content.

Figure 6: Snippet of page content with highlighted element attributes

The id attribute specifies a unique identifier for an HTML element. The value of the id attribute is usually unique within the HTML document. The class attribute is often used to point to a class name in a style sheet. It can also be used by JavaScript to access and manipulate elements with the specific class name.
The construction of dictionaries adheres to the following process:

1. Compilation of unique term sets from each distinct document within the phishing segment of the training set.
2. Aggregation of these sets into a comprehensive list of terms.
3. Retention of the most frequently occurring terms through a counting mechanism.

For each sample within the dataset, we employ a count vectorization technique to align the extracted terms with the prepared dictionaries. This alignment is grounded in the frequency of occurrence exhibited by each token within the entire text of the respective sample.

To add further significance to the numeric vectors, we perform TF-IDF (term frequency-inverse document frequency). This statistical measure evaluates the relevance of a word to a document within a collection of documents. It considers both how frequently a word appears in a document and its inverse document frequency across the entire dataset:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}, \quad (1)$$

$$idf(t,D) = log\frac{1+N}{1+|\{d \in D: t \in d\}|}, \quad (2)$$

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D), \quad (3)$$

where ft,d is the raw count of a term in a document, i.e., the number of times that term t occurs in document d. Note the denominator is simply the total number of terms in document d (counting each occurrence of the same term separately). N: total number of documents in the corpus N = |D|.

|{d∈D:t∈d}| : number of documents where the term t appears.

By applying TF-IDF, we emphasize the importance of each term in the context of phishing detection.

Even when focusing solely on the most frequently occurring phishing terms to map textual information, the resultant array of variables remains substantial. In addressing this, and with the dual aim of distilling valuable insights while mitigating overfitting, we employ Principal Component Analysis (PCA). This technique serves to condense the dimensions of our data vectors, effectively retaining the maximal information within more compact representations.

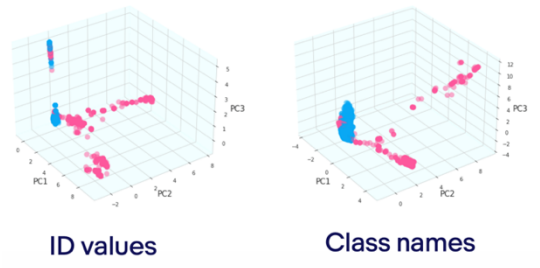

Figure 7: First 3 Principal Components on unseen data (red – Brand-targeting Phishing, blue – "clean" samples)

Figure 7 provides a visualization of the first three principal components of the test data, showcasing the distribution of ID values and class names. These test samples were initially mapped onto dictionaries that were constructed using the pre-built training data. Following this mapping, we applied PCA to achieve the visualization depicted in the figure, vividly illustrating the successful separation and differentiation of samples using these chosen features.

4.4. Visible text features

Our fourth set of features involves visible text obtained from the webpage, which we categorize into four parts: Title, URL (treated as text), Body (entire visible text), and Footer. To make the text more informative, we map it with dictionaries containing the most frequent terms and phrases extracted from known phishing pages. This step empowers the model to recognize key indicators of phishing attempts, such as the presence of "login" or "register" in the Title or "copyrights" in the Footer.

Before converting text into a numeric representation, we perform text pre-processing, that helps to put all text on equal footing. It involves following steps:

1. Translation to English
2. Removing non-ASCII characters
3. Conversion to lowercase
4. Removing punctuation
5. Removing numbers
6. Removing extra spaces

Subsequently, the processed text is translated into a numeric format using pre-established dictionaries containing the most prevalent terms or tokens, derived from the phishing data within the training set. This process involves the application of both count vectorization, which captures token frequency across the entire text of each sample, and the TF-IDF statistical technique.

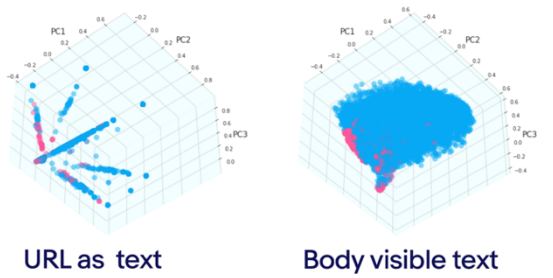

Figure 8: First 3 Principal Components on unseen data (red – Brand-targeting Phishing, blue – "clean" samples)

As we have capacity to translate all text into English, we have incorporated features that indicate the original language group of the text. This capability enables us to detect phishing attempts aimed at our global customer base within a unified model. For instances where such capabilities are unavailable, we recommend the creation of separate models for each language or the exclusion of page visible text as a feature. However, the URL text can still be considered for use.

5. Classifiers

We compared the performance of different classifiers using varying combinations of feature sets to enhance phishing detection accuracy. The selected classifiers are Logistic Regression, Random Forest, and XGBoost, each offering distinct advantages.

Logistic Regression: This classic algorithm suits straightforward tasks with linear relationships between features and outcomes, providing an interpretable baseline for comparison.

Random Forest: By aggregating the outputs of multiple decision trees, Random Forest effectively captures intricate feature interactions and minimizes overfitting.

XGBoost: Known for its predictive power, it constructs an ensemble of weak learners and iteratively improves their performance, accommodating various data types and complex patterns.

6. Evaluation metrics

Balanced accuracy is a better metric to use with imbalanced data. It accounts for both the positive and negative outcome classes and does not mislead with imbalanced data.

$$BA = \frac{1}{2} \cdot \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right), \quad (4)$$

Where:

TP – true positive (the correctly predicted positive class outcome of the model),

TN – true negative (the correctly predicted negative class outcome of the model),

FP – false positive (the incorrectly predicted positive class outcome of the model),

FN – false negative (the incorrectly predicted negative class outcome of the model).

7. Results

In this section, we evaluate the performance of three classifiers across five distinct feature sets. Our evaluation is based on cross-validation, a technique where the training data is split into multiple folds (we use five folds), with each fold serving as validation data in a rotation. During each iteration, the model is trained on four folds and validated on the fifth, and this process is repeated five times to ensure robust results.

The five feature sets are as follows:

Features1: URL-Based Features. This set includes features related to URL structure, domain ranking, and the URL as text.

Features2: Encompassing both URL and page content-based countable features, as well as features derived from page structure and brand assets analysis.

Features3: Tag Names Features. Features extracted from unique HTML element IDs, class names, and form names.

Features4: Visible Text Features. Features derived from visible text across different parts of the webpage, including Title, URL, Body, and Footer.

Combined: A comprehensive set comprising all the features from the previous sets.

Table 2
Comparison of different classifiers and features combination. Mean balanced accuracy and standard deviation.

Head 1	LR	RF	XGBoost
Features1	0.7082	0.8306	0.8362
	(0.007)	(0.007)	(0.008)
Features2	0.8804	0.9799	0.9766
	(0.005)	(0.001)	(0.001)
Features3	0.9439	0.9849	0.9851
	(0.006)	(0.002)	(0.002)
Features4	0.9705	0.9803	0.9898
	(0.005)	(0.003)	(0.003)
Combined	0.9857	0.9897	0.9941
	(0.002)	(0.002)	(0.001)

The performance of different classifiers and feature sets was evaluated using cross-validation on the training data. XGBoost demonstrated the highest mean balanced accuracy, particularly when utilizing the combined feature set. When applied to the test data, the XGBoost model with the comprehensive features achieved an impressive accuracy of 99.8342.

To identify the key drivers of accurate predictions, we examined the top 20 most influential features:

1. Classes names. PC #0 (First principal component)
2. Form names. PC #0
3. ID values. PC #0
4. Title. PC #25
5. Classes names. PC #21
6. # non-brand related <a>
7. Classes names. PC #4
8. Title. PC #17
9. Title. PC #6
10. Title. PC #16
11. URL contains brand keyword
12. Body text. PC #13
13. ID values. PC #3
14. Classes names. PC #18
15. Domain rank not in 1m
16. # subdomains
17. ID values. PC #2
18. # brand's assets <img>
19. Title. PC #13
20. Title. PC #0

8. Conclusion and further work

In this study, we presented a comprehensive approach for detecting phishing pages that target our brand's customers. By leveraging a hybrid feature-based approach, encompassing URL structure, HTML elements, text content, and brand-specific signals, we developed a robust detection model. Through rigorous evaluation, we demonstrated the effectiveness of our approach in accurately identifying phishing attempts.

While our approach shows promising results, there are opportunities for further enhancement and exploration. We plan to explore integration with social media phishing detection, develop better strategies to counter cloaking and filtering techniques, optimize takedown processes, and leverage the potential of Large Language Models. These endeavors aim to reinforce our brand's cybersecurity measures and protect our customers from evolving threats.

Acknowledgements

We extend our gratitude to the Anti-Phishing Working Group (APWG) for their valuable insights and resources that contributed to the success of our research. Special thanks to our colleagues at eBay for their continuous support and collaboration throughout this project. We also acknowledge the contributions of Marc Green and the broader cybersecurity community for their discussions and feedback, which enriched our understanding and approach.

References

[1] Anti-Phishing Working Group (APWG). Phishing activity trends report. 3rd Quarter 2022. URL: https://docs.apwg.org/reports/apwg_trends_report_q3_2022.pdfP. S. Abril, R. Plant, The patent holder's dilemma: Buy, sell, or troll?, Communications of the ACM 50 (2007) 36–44. doi:10.1145/1188913.1188915.

[2] Das Guptta, S., Shahriar, K.T., Alqahtani, H. et al. Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques. Ann. Data. Sci. (2022). https://doi.org/10.1007/s40745-022-00379-8

[3] Q. A. Al-Haija and A. A. Badawi, "URL-based Phishing Websites Detection via Machine

Learning," 2021 International Conference on Data Analytics for Business and Industry (ICDABI), Sakheer, Bahrain, 2021, pp. 644-649, doi: 10.1109/ICDABI53623.2021.9655851.

[4] Arun Kulkarni, Leonard L. Brown III, "Phishing Websites Detection using Machine Learning," International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, 2019, pp. 8. URL: https://thesai.org/Downloads/Volume10No7/Paper_2-Phishing_Websites_Detection_using_Machine_Learning.pdf

[5] A. Ghimire, A. Kumar Jha, S. Thapa, S. Mishra and A. Mani Jha, "Machine Learning Approach Based on Hybrid Features for Detection of Phishing URLs," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 954-959, doi: 10.1109/Confluence51648.2021.9377113.

[6] S. Zaman, S. M. Uddin Deep, Z. Kawsar, M. Ashaduzzaman and A. I. Pritom, "Phishing Website Detection Using Effective Classifiers and Feature Selection Techniques," 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, 2019, pp. 1-6, doi: 10.1109/ICIET48527.2019.9290554.

[7] P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," in IEEE Access, vol. 7, pp. 15196-15209, 2019, doi: 10.1109/ACCESS.2019.2892066.