

PhishLens-XAI: Explainable Multilingual Phishing Detection System

Purva Khedekar¹, Aarti Gawai²

¹Student, Department of Information Technology and Computer Science, D.G. Ruparel College

²Assistant Professor, Department of Information Technology and Computer Science, D.G. Ruparel College, Mumbai, Maharashtra, India

¹khedekarpurva17@gmail.com

²aartigawai1611@gmail.com

Abstract - Phishing is a major cybersecurity threat that can occur through emails, text messages, websites, and URLs in different languages. Many existing phishing detection systems mainly focus on English and provide limited explanations for their predictions. This research proposes PhishLens-XAI: Explainable Multilingual Phishing Detection System, which aims to detect phishing content across multiple languages while explaining the reasons behind each prediction. The proposed system analyzes text and URL features to identify common phishing indicators, including suspicious words, links, domains, and linguistic patterns. Machine-learning algorithms and multilingual transformer models, particularly XLM-RoBERTa, are used to classify inputs as phishing or legitimate. To improve transparency, SHAP and LIME are integrated to identify important features influencing each prediction. The system is evaluated using accuracy, precision, recall, and F1-score, with results compared across selected languages to measure multilingual performance. By combining multilingual phishing detection with explainable artificial intelligence, PhishLens-XAI aims to provide a reliable, transparent, and user-friendly solution that helps users and cybersecurity analysts understand detection results and respond effectively to potential phishing threats.

Key Words: phishing detection, multilingual detection, explainable artificial intelligence, SHAP, LIME

I. INTRODUCTION

The expansion of Internet services, online banking, e-commerce, social networking, and has increased the level of vulnerability of people and companies to cyber-attacks. One of the cyber-attacks that have become common is the use of social engineering attacks such as phishing where the attacker uses fake websites, URLs, emails, and other content to imitate real Internet services and mislead the user into providing their personal information including login information and other sensitive information. Blacklisting methods used to detect phishing attacks commonly face problems when dealing with new and temporary phishing sources. As detection is done based on malicious indicators identified before. Therefore, ML-based phishing attack detection became an important area of research that can help to detect new and unknown phishing attacks using information obtained from URL, webpage, domain, and content analysis [1], [2].

In terms of phishing detection, the methods based on machine learning have proven their capability in automatically detecting patterns of genuine and fake sites.

However, high classification accuracy is not always sufficient for achieving the transparency and user trust in the model. Many ML-based classifiers can be described as black box models since their decisions are presented without a clear list of criteria that is responsible for that result. This problem is especially relevant in the area of cybersecurity where analysts and end-users need an explanation of why a particular website or link was marked as phishing to perform appropriate actions further. In such a situation, Explainable Artificial Intelligence (XAI) is needed since it allows obtaining interpretable information about model decisions. The methods that can help in analyzing the features used in a decision making process and understanding the model behavior include SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). The latest works demonstrate that SHAP and LIME can be successfully applied to feature analysis and explainable phishing detection [3], [4].

Multilingual processing along with explainability is one area of research that can be considered a promising research direction in phishing detection. Multilingual processing requires flexibility to handle variations while keeping intact the features essential for reliable detection of phishing. On the other hand, interpretable systems are expected to provide understandable explanations for their decisions instead of simply making binary decisions on whether a message is genuine or not.

The following paper presents an approach to creating a PhishLens-XAI: An Explainable Multilingual Phishing Detection System. This system combines several components, including multilingual processing, phishing-related feature extraction, classification using machine learning, and explainable artificial intelligence. The system developed by this approach will be capable of analyzing features of URLs and text, extracting verbal data when necessary, classifying the input either as phishing or legitimate, and generating explanations for the features responsible for making the prediction.

The main advantage of the PhishLens-XAI framework can be credited to the use of the multilingual phishing detection process along with the explainable machine learning approach all together into one framework. In particular, the main goal of the framework is to provide not just a phishing prediction itself but also an explainable reason behind the prediction. With the help of combining linguistic and other features related to phishing links/web pages and using the explainable AI, the system tries to make the process more transparent, interpretable, and useful for people. The framework is assessed via several classification metrics (accuracy, precision, recall, F1 score, and others).

II. Review of Literature

Phishing detection has changed from traditional methods such as blacklisting and rules based methods into the machine learning (ML), natural language processing (NLP), transformer, and explainable artificial intelligence (XAI). The recent research on this topic does not concentrate only on phishing detection, rather, it is also focused on finding the reasoning behind a specific prediction made by the system.

Calzarossa et al. have created an explainable machine learning method for phishing detection and found the most relevant features of websites for phishing versus genuine websites discrimination [1]. In the other paper, Calzarossa et al. have used bag-of-words features to detect phishing based on webpage text and give understandable information about the prediction [2]. Greco et al. identified the problem of XAI in phishing detection and focused on explanations for normal users [3]. Yakandawala et al. used LIME for explanations of individual phishing emails predictions [4].

Detection of the phishing emails in multiple languages is also an issue to consider. Staples et al. compared machine learning models and transformers in English, French, and Russian phishing emails [5]. The study proved that XLM-RoBERTa performed efficiently in cross-language phishing detection. Apeke et al. compared BERT, RoBERTa, mBERT, and XLM-RoBERTa for cross-language phishing detection as well [6]. It resulted in XLM-RoBERTa being more consistent in performance when tested on different languages. Thus, the use of multilingual models will help detect phishing in emails that are sent in different languages.

Some of the most recent papers have processed the combination of multilingual detection and explainability. Shah et al. applied NLP, XGBoost, SHAP, and SMOTE for multilingual phishing analysis [7]. Lai et al. applied ML models with SHAP and LIME for phishing URL detection [8]. Chai et al. suggested the multimodal technique for phishing detection that considers several types of phishing information and explains its predictions [9]. Pandya et al. used NLP, OCR, behavioral analysis, ML models, SHAP, and LIME for phishing website and email detection [10].

From the above-mentioned studies, it becomes proof that multilingual and explainable phishing detection techniques have largely been considered separately.

Multilingual researchers are more concerned with the detection accuracy and cross-language aspects of phishing, whereas explainability-oriented studies concentrate on the analysis of the decision making of models. Hence, there is a need for a system that combines both the capabilities.

The proposed “PhishLens-XAI” is a combination of “multilingual phishing detection and explainable machine learning techniques”. This system not only detects which website is phishing and which one is legitimate, it will also explain the reasons behind the phishing websites in multiple languages in which users are comfortable.

Ref.	Main Approach	Multilingual	XAI	Main Focus	Limitation
[1]	Explainable ML	No	✓	Phishing website features	Limited multilingual evaluation
[2]	Bag-of-Words + ML	No	✓	Text-based phishing detection	Language-specific
[3]	XAI-based phishing detection	No	✓	User-understandable explanations	Mainly user awareness
[4]	ML + LIME	No	✓	Phishing email detection	Limited multilingual analysis
[5]	ML + Transformers	✓	Limited	Cross-language phishing detection	Explanation is not the main focus
[6]	BERT, mBERT, XLM-R	✓	Limited	Cross-lingual and adversarial detection	Limited XAI analysis
[7]	NLP + XGBoost + SHAP	✓	✓	Multilingual phishing analysis	Requires broader evaluation
[8]	ML + SHAP/LIME	No	✓	Phishing URL detection	Mainly URL-based
[9]	Multimodal attention	—	✓	URL/text/visual information	More complex architecture
[10]	NLP + OCR + ML + SHAP/LIME	—	✓	Website and email detection	Detailed multilingual evaluation limited
Phish Lens-XAI	Multilingual ML + XAI	✓	✓	Multilingual detection + explanation	To be evaluated experimentally

Table -1: Comparison of Existing Phishing Detection Studies

III. Problem Identification

Phishing attacks can happen through the emails, websites, URLs, and messages written in different languages. Existing phishing detection systems can achieve good results in identifying the phishing content. However, many existing systems are mainly focused on English-language data or focus on only one type of phishing information, such as emails, websites, or URLs [5], [6]. Another major challenge is the limited availability of phishing datasets in different languages. Because of this limited data, it is difficult to develop a detection system that can provide similar and reliable performance across multiple languages [5].

Another important problem is the “lack of clear explanations”. A machine-learning model can classify an email or website as phishing, but it may not be explained why the website/content was classified as phishing.

For example, a user may receive a prediction that an email is spam, but the user may not know which words, URL features, or other characteristics are responsible for generating the model to make that decision. Explainable artificial intelligence (XAI) methods can help to solve this problem by showing the important features that influence the model's prediction [1], [3], [4]. Calzarossa et al. developed an explainable machine-learning method that provides a phishing prediction and identifies the features that are related to that prediction [1]. This shows that providing an explanation along with the prediction can make a phishing detection system easier to understand.

Therefore, the main problem addressed by this research is the “lack of an integrated phishing detection system that can work with multiple languages and also explain the reasons behind its predictions”. Existing studies have identified multilingual phishing detection and explainable phishing detection, but these two capabilities are not always combined into a one system [1], [5], [6]. The proposed “PhishLens-XAI” aims to solve this problem by combining multilingual phishing detection with an XAI. The system is targeted to identify whether the given content is phishing or genuine and also provide information about the important features that influenced the prediction.

Problem Statement:

“Existing phishing detection systems have limitations in handling phishing content written in multiple languages and in explaining their prediction results. Therefore, there is a need for a system that can detect phishing across multiple languages and provide simple and understandable reasons for its decisions. The proposed PhishLens-XAI addresses this problem by combining multilingual phishing detection with explainable artificial intelligence.”

IV. Objectives of the Study

The specific objectives of the study are:

1. To collect and create a dataset that contains the phishing and genuine emails, URLs, or text in the selected languages for training and testing the proposed system.
2. To preprocess the collected data by removing unnecessary information, handling duplicate or missing data, and preparing the text and URLs for analysis.
3. To identify important phishing features from the text and URL data, such as suspicious words, unusual URLs, urgent requests, and other patterns related to the phishing.
4. To develop a multilingual phishing detection model that can classify the given input as either phishing or genuine across the selected languages.
5. To apply an Explainable AI (XAI) technique to identify the important features that are responsible for the model's prediction.
6. To provide simple and understandable explanations so that users and security analysts can understand why an input was classified as a phishing or genuine.
7. To measure the performance of the proposed PhishLens-XAI system using standard measures such as accuracy, precision, recall, and F1-score.
8. To compare the detection performance across the selected languages and determine whether the proposed system provides reliable results for different languages.

V. Hypothesis of the Study

The hypothesis of this study is developed to decide whether the proposed PhishLens-XAI system can effectively detect the phishing content in user preferable languages and provide useful explanations for its predicted results.

Null Hypothesis (H_0): There is no meaningful difference in the phishing detection performance of the proposed PhishLens-XAI system across the selected languages.

Alternative Hypothesis (H_1): There is a meaningful difference in the phishing detection performance of the proposed PhishLens-XAI system across the selected languages.

VI. Research Design-Methodology

A. Research Design:

The research will be conducted based on an experimental and measurable research design method for designing and testing the proposed "PhishLens-XAI: Explainable Multilingual Phishing Detection system". The main objective of this system is to check whether the provided text or URL is "phishing or not".

This system combines multilingual text processing, URL feature extraction, machine learning models, and Explainable Artificial Intelligence (XAI). The steps involved in this research process include the collection of data, data preprocessing, language identification, feature extraction, model training, phishing classification, generating the explanation, and performance measuring.

The performance of the system will be measured by following performance metrics and results will be compared for the selected languages.

B. Data Collection and Preparation:

Within each language, the dataset includes "phishing and legitimate emails, text, and URLs". There are two classifications given to every sample: "phishing (1)" and "legitimate (0)".

Data is tested before it is entered into a model. Duplicate, empty records are identified as unnecessary and removed during the testing process. The text is filtered and prepared for further processing, while URLs are normalized for feature extraction.

This dataset is subdivided into "training, validation and testing" sets. The representation of phishing samples and genuine samples are maintained in each set through the stratified sampling.

C. Multilingual Text Processing:

This system is proposed to handle phishing messages in different languages. First, the language in which each sample text is written. After normalizing the text, it is time to classify it.

It takes into account phishing patterns such as a:
Urgent messages;
Requesting passwords or personal information;
Account verification requests;
Financial requests; and
Suspicious or misleading instructions.

XLM-RoBERTa is used for categorizing text into different languages. The previous study reviewed the efficiency of XLM-RoBERTa for detecting cross-language and multilingual phishing.

D. Feature Extraction:

This system is proposed to take advantage of the features that are obtained by analyzing the Text and URLs.

Some of the attributes listed below:

1) Text Features:

This document contains suspicious language, urgent slang utilized for validating credentials, financial terminology, impersonation practices, and other forms of phishing. Phrases can be written again or sent to email for confirmation.

2) URL Features:

The majority of URLs possess various features such as length, dots and subdomains (such as domain names), special characters, IP addresses, suspicious words in URL descriptions or patterns, HTTPS usage, and unique URL occurrence.

Identifying attributes linked to the phishing is achieved through this classification model's use of these traits. The XAI component benefits from their knowledge.

E. Model Development:

Machine-learning and multilingual models are used in the proposed system to identify the phishing attempts.

XLM-RoBERTa is used for multilingual text analysis.

XGBoost or Random Forest is used to identify the source of information for classification using URLs. Baseline models are used for comparison and it can also be provided by other methods such as SVM or Logistic Regression.

Combining the text and URL predictions yields for the final prediction using following equation:

$$P_{final} = \alpha P_{text} + (1 - \alpha) P_{URL}$$

A weight assigned to a text model is (P_{α}), while $p(R)$ is the prediction probability from the URL model.

Depending on the result stated that the emails/text or URLs are Phishing Or Legitimate. According to the selected threshold.

F. Explainable Artificial Intelligence:

Explainable Artificial Intelligence (XAI) is a key aspect of the proposed PhishLens-XAI solution. "SHAP" and "LIME" are applied for explaining the predictions which are made by the model.

SHAP is used to reveal the features that have the biggest impact on the prediction. SHAP also enables the assessment of feature importance throughout the whole dataset.

LIME provides the explanation of the prediction using the identification of the most significant words/word combinations, phrases, or URL attributes.

For example, the explanation may look as follows:

"Prediction": Phishing

"Confidence": 95%

"Important features": Urgency, credentials request, and suspicious URL.

The explanation helps users and security analysts in understanding the reasoning behind the prediction. It directly solves the issue that is described in the problem statement of the research.

G. Proposed System Architecture:

The main components of the proposed PhishLens-XAI system are shown below in the Fig.1

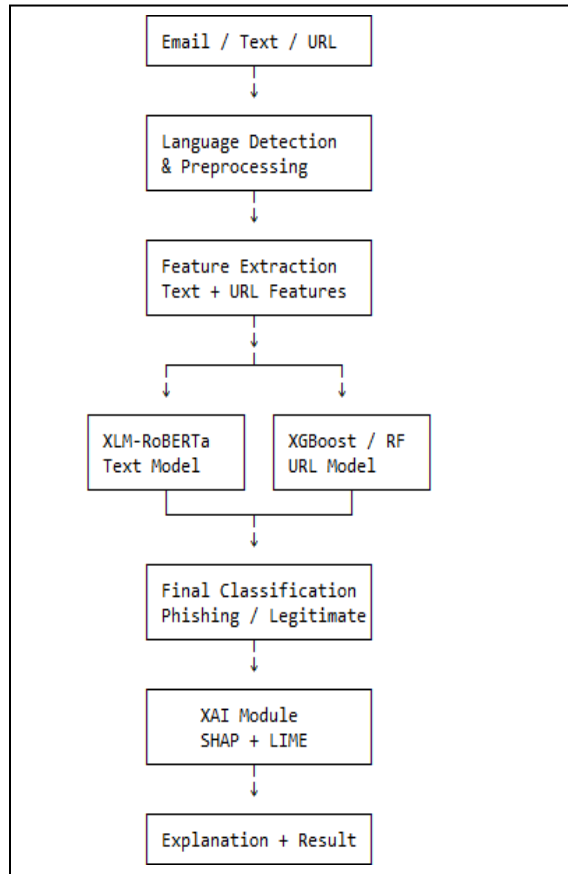


Fig.1 Architecture of the proposed PhishLens-XAI system

H. Research Methodology Flow:

The complete research process is shown in Fig. 2.

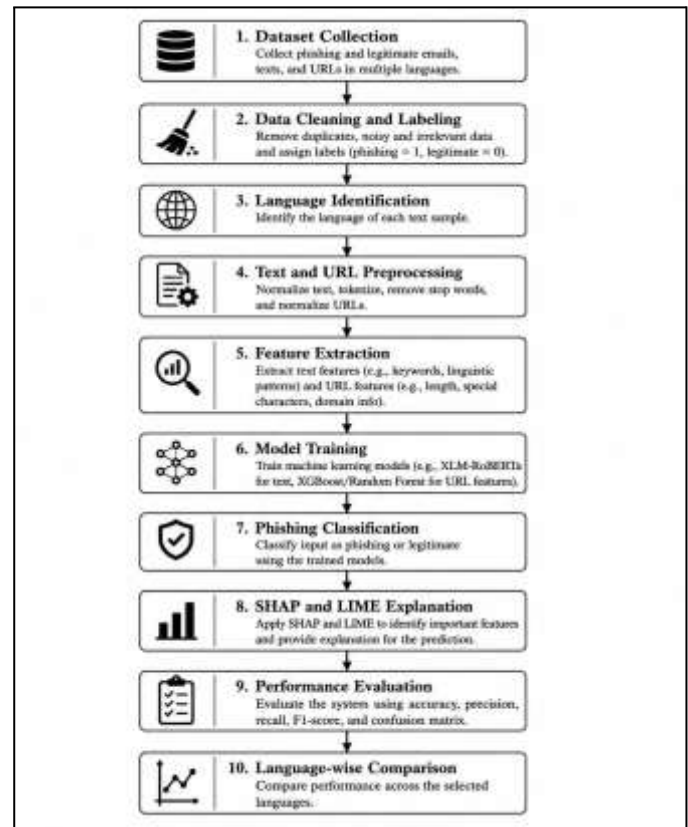


Fig.2 Research Methodology Flowchart

I. Performance Measuring:

The performance of the proposed system is measured using the following methods:

1. Accuracy: Accuracy measures the overall number of the correct predictions.
2. Precision: Precision measures the correctness of phishing predictions.
3. Recall: Recall measures how many actual phishing samples are correctly detected.
4. F1-score: F1-score provides the combination of the Recall and Precision methods.

The performance is calculated for each selected language and compared across the languages. This evaluation is important because the objectives of the study include determining whether the proposed system provides compatible results for different languages.

J. Implementation Environment:

The proposed system can be implemented using "Python", Python libraries such as "Pandas, NumPy, Scikit-learn, XGBoost, Transformers, SHAP, and LIME" can be used for the data processing, model development, classification, and explanation generation.

The dataset division, model settings, and testing procedure should be recorded to support reproducibility of the experiment.

VII. Significance of the Study

The proposed PhishLens-XAI: Explainable Multilingual Phishing Detection System is significant because it aims to improve phishing detection for content written in different languages while also explaining the reasons behind the detection results. Existing phishing detection methods often focus on a limited number of languages and may not provide clear explanations for their predictions.

The study can benefit general users by helping them identify suspicious emails, messages, and URLs and understand the common signs of phishing. The use of XAI methods such as SHAP and LIME can provide simple information about the features that influenced a phishing prediction. This may help users better understand and respond to suspicious content.

The study can also support security analysts and researchers by providing information about the features that contribute to phishing detection. The multilingual approach allows the performance of the system to be examined across different languages, supporting the development of more inclusive phishing detection systems.

Finally, this research contributes to the field of cybersecurity and explainable artificial intelligence by combining multilingual phishing detection with explainable predictions in a single framework. The proposed approach can also provide a foundation for future research involving additional languages, larger datasets, and improved phishing detection techniques.

3. CONCLUSIONS

Phishing attacks are a serious cybersecurity problem because they can be found in emails, messages, and URLs written in different languages. This research proposed PhishLens-XAI: Explainable Multilingual Phishing Detection System to solve two main problems: detecting phishing in different languages and explaining why the system gives a particular prediction.

The proposed system uses multilingual text processing, URL features, machine-learning models, and Explainable AI (XAI). It classifies the input as phishing or legitimate and uses SHAP and LIME to show the important features that affected the prediction. The system is evaluated using accuracy, precision, recall, and F1-score, with performance compared across the selected languages.

The main benefit of PhishLens-XAI is that it does not only provide a phishing prediction. It also provides an explanation of the decision. This can help users and security analysts better understand common phishing signs and make better decisions.

However, the performance of the system may be different for different languages because the amount and quality of available data may vary. Future research can improve the system by using larger datasets, adding more languages, and testing the system against new phishing attacks.

Overall, PhishLens-XAI provides a simple approach for detecting phishing across multiple languages while also explaining the reasons behind its predictions. It can provide a useful foundation for future research in multilingual and explainable phishing detection.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my **research supervisor/guide** for providing valuable guidance, suggestions, and continuous support throughout this research work.

I am thankful to my **institution and department** for providing the necessary facilities, resources, and academic support required to complete this research.

I would also like to thank all the **faculty members** for their valuable suggestions and encouragement during the research work.

I am deeply grateful to my **family and friends** for their constant support, motivation, and encouragement throughout the completion of this research.

Finally, I would like to thank all the researchers, research papers, publicly available resources, and open-source tools that helped me in understanding and developing this research on **PhishLens-XAI: Explainable Multilingual Phishing Detection System**.

I sincerely acknowledge the support and encouragement received from everyone who helped me complete this research work.

REFERENCES

- [1] M. C. Calzarossa, P. Giudici, and R. Zieni, "Explainable machine learning for phishing feature detection," *Quality and Reliability Engineering International*, vol. 40, no. 1, pp. 362–373, 2024, doi: 10.1002/qre.3411.
- [2] M. C. Calzarossa, P. Giudici, and R. Zieni, "Explainable Machine Learning for Bag of Words-Based Phishing Detection," in *Explainable Artificial Intelligence*, 2023.
- [3] F. Greco, G. Desolda, and A. Esposito, "Explaining Phishing Attacks: An XAI Approach to Enhance User Awareness and Trust," in *Proc. ITASEC*, 2023.
- [4] Y. L. D. H. Yakandawala *et al.*, "Explainable AI for Transparent Phishing Email Detection," in *Proc. ICATC*, 2024.
- [5] D. Staples, S. Hakak, and P. Cook, "A Comparison of Machine Learning Algorithms for Multilingual Phishing Detection," in *Proc. PST*, 2023.
- [6] S. Apeke, T. Djiraikinan, and M. Barate, "Phishing detection by Transformers: Adversarial Robustness and Cross-Lingual Generalization," *Indian Journal of Science and Technology*, vol. 19, no. 12, 2026.

[7] Shah *et al.*, “Multilingual Phishing Analysis: Feature-Based Scoring and Level Classification,” in *Proc. ICSFT*, 2026.

[8] H. K. M. Lai *et al.*, “Explainable Machine Learning Framework for Phishing URL Detection,” in *Proc. CSPA*, 2026.

[9] Y. Chai *et al.*, “An Explainable Multi-Modal Hierarchical Attention Model for Developing Phishing Threat Intelligence,” *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 2, pp. 790–803, 2022.

[10] H. Pandya *et al.*, “Detection of Phishing Websites and Emails Using Explainable Machine Learning Models,” *Journal of Information Systems Engineering and Management*, 2025.

[11] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016, doi: 10.1145/2939672.2939778.

[13] A. Conneau, K. Khandelwal, N. Goyal, *et al.*, “Unsupervised Cross-lingual Representation Learning at Scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020, doi: 10.18653/v1/2020.acl-main.747.

[14] M. Khonji, Y. Iraqi, and A. Jones, “Phishing Detection: A Literature Survey,” *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 2091–2121, 2013.

[15] Y. Zhang, J. I. Hong, and L. F. Cranor, “CANTINA: A Content-Based Approach to Detecting Phishing Web Sites,” in *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pp. 639–648, 2007.

[16] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Learning to Detect Malicious URLs,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.