Lightweight Phishing Detection Using Natural Language Processing on Email Metadata

Celinus KIYEA
Department of Cyber Security
Faculty of Computing
Mewar International University
Masaka - Nigeria
ckiyea@gmail.com

Inemesit Isaiah ALUNYO
Department of Cyber Security
Faculty of Computing
Mewar International University
Masaka – Nigeria
inemetukudo@yahoo.com

Abraham DANLAMI
Department of Computer Science,
Faculty of Computing and Information system
Federal University Wukari Taraba State, Nigeria
Abdanabraham@gmail.com

***Abstract*—Internet has influenced and shaped the way people communicate, it has also made security become the fundamental challenging aspect. Its users can be affected by so many security aspects and one of the most common is phishing. Phishing attacks remain among the most prevalent and damaging cybersecurity threats as the number of attacks doubled from 2020 to 2021 recording about 260,000 attacks. Traditional detection methods often depend on full content analysis, introducing privacy risks which is against ethics and computational burdens. This paper proposes a lightweight, privacy-respecting phishing detection system that leverages Natural Language Processing (NLP) techniques applied solely to email metadata. The researchers further perform an ablation study to evaluate the importance of different metadata components and compare the lightweight system against heavy, content-based models. Experiments on public phishing datasets demonstrate that researchers' approach achieves high detection accuracy (up to 93.2%) while significantly reducing computational overhead.**

***Keywords*: Metadata, Phishing Attack, Natural Language Processing, Lightweight**

I. INTRODUCTION

With the advent of Internet that has influenced and shaped the way people communicate, it has also made security become the fundamental challenging aspect. Internet users can be affected by so many security aspects and one of the most common of them and a threat is phishing. This is a type of attack that aims to steal or misuse different kinds of users' sensitive personal information, such as account details, identity information, passwords, and credit card details. Phishers mimic organisational original websites and make it indistinguishable to the eye so as to gather sensitive information about the users. Phishing remains a serious challenge for cybersecurity, especially in corporate and cloud environments. Anti-Phishing Working Group (APWG) is an organization that collects, analyzes, and exchanges a list of credential URLs. It publishes a quarterly report on phishing activities across the globe. The number of phishing attacks has doubled from 2020 to 2021. More than 260,000 reported phishing attacks in July 2021 [1]. Webmail continues to be among the top methods of phishing attempts. There have been an increase in phishing attacks on named brands, from 400 per day in July to 700 in September 2021 despite the fact that there are state laws for penalizing criminals for phishing attacks. Anti-phishing Act of 2005 imposes fines and imprisonment for up to five years or both for a person involved with phishing attacks [2]. California leads the way in having a strong state law on phishing attacks. Traditional approaches often involve parsing full email bodies, analyzing embedded links, or checking attachments operations that are both resource-intensive and potentially intrusive. Therefore, there is a strong need to develop a system that efficiently detects phishing websites

In this work, we propose a lightweight phishing detection system based on email metadata analysis using simple NLP techniques. Our system does not access or process the email body, providing faster, more scalable, and more privacy-compliant phishing detection.

Contributions

a. Design of an NLP-based feature extraction method focused only on metadata.

b. Development and evaluation of lightweight models suitable for real-time phishing detection.

c. Conducting an ablation study to analyze the contribution of each metadata type.

d. Comparative analysis against conventional full-content-based phishing detection methods.

II. RELATED WORK

Phishing is a type of cyberattack that uses fraudulent emails, text messages, phone calls or websites to trick people into sharing sensitive data, downloading malware or otherwise exposing themselves to cybercrime. These attacks are a form

of social engineering activity which takes advantage of human error, fake stories and pressure tactics to manipulate victims into unintentionally harming themselves or their organizations. Most of the hackers in phishing scams do pretend to be someone the victim trusts, like a colleague, boss, authority figure or representative of a well-known brand. The hacker sends a message directing the victim to pay an invoice, open an attachment, click a link or take some other action [3].

The emergence of emails as a convenient means of communication in modern days has created a lot of problems as phishing is concerned.
Different methods and technologies have been used by researchers to detect and counteract the phishers, but they keep increasing tremendously. Some researchers have used machine learning models, while others have focused on manual add-ins and natural language processing methods on email text. Traditional methods, such as the blacklist approach proposed by [4], in this method, the blacklist was set to contain the sender blacklist and the link blacklist. In the detection process, a check is carried out on the sender's address and the link address in the message sent to determine whether it is on the blacklist or not. All these are to determine the authenticity of the email whether it is not a phishing email or not. A report is manually sent to the email user.

In a similar case, [5], were able to divide the detection solutions into three main parts, they identified the blacklist, heuristics and the data mining. Going by this they highlighted that the blacklist was the simplest and lightest but for the fact that they are updated frequent lists of the previous detected phishing URLs and IP addresses. This further explains that they cannot be able to prevent zero-day attacks. Besides they are time consuming as the updates are done manually. Heuristics solutions on the other hand embody the real characteristics that are eminent in the phishing attacks. Carrying out a set of identified general heuristics tests, will make it possible to handle and detect the zero-day phishing attacks because it also strikes a balance between efficiency and effectiveness. A research carried out by [6] argues that though machine learning algorithms offer high detection accuracy, they require enough computational resources and constant updates to remain effective against the continuously changing and emerging phishing techniques. Heuristic-based approaches have proven to be good and fast in detection especially as they strike a balance between effectiveness and efficiency. They also require a lower resource demands but may struggle with new or sophisticated attacks.

The option of using NLP in Phishing attacks detection lies on the fact that, since most phishing attacks depend fundamentally on manipulative language, therefore, Natural Language Processing (NLP) is considered one of the most appropriate and effective techniques for detecting them. Also, the traditional techniques such as signature-based or blacklist methods have been seen inefficient in the presence of novel or obfuscated attacks again enables NLP systems to analyse the semantic, syntactic, and contextual properties of email content, thereby identifying deceptive patterns that are not tied to fixed signatures[7]. NLP use is relevant and appropriate because it aligns with the state of the art and industry standard. This is because the most modern platforms continuously rely on the NPL and machine learning strategies to span detection across a large number of emails, to counter adaptive phishing attacks some high level of intelligence to threats[8].

In the same way, [5] and [9] still went further to explore the next solution by using data mining techniques that takes the advantage of various different machine learning algorithms such as K-nearest Neighbour, support vector Machines, K-means and the likes to classify and cluster the newly arriving emails. In order to apply this detection method, the training and testing stages must be carried out. Therefore, a list of parameters to be used for the emails analysis is needed alongside a well labeled dataset for such a purpose. In the same manner [10] also, detected phishing emails by employing the basic machine learning methods. These include decision trees, logistic regression, random forests, and SVM combined with supervised classification methods. Furthermore, as events and new development continued unfolding themselves, artificial intelligence evolved and was used in conjunction with machine learning algorithms to detect phishing emails. [11], applied a combination of Natural Language Processing and Machine Learning for the detection of phishing email. Various features like semantic, syntax and contextual features have been used previously by [9] and [12] respectively in this area. In the same manner, [13] in an attempt to address this problem, proposed a model that leverages neural networks which include CNN, RNN, LSTM, and transformer-based models together with the Natural Language Processing (NLP) to identify malicious patterns and improve the accuracy of phishing detection.

III. THE MODEL ARCHITECTURE

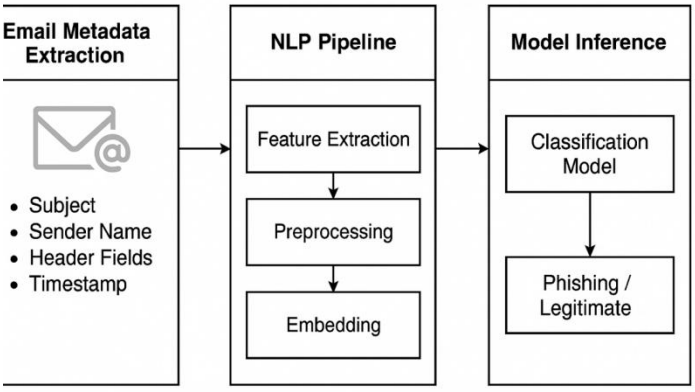

Fig. 1: Architecture Lightweight Phishing Detection

Architecture of the lightweight phishing detection system using NLP on email metadata is made up of three main sections which are the Email Metadata Extraction, NLP Pipeline and the Model Inference. The system extracts emails key metadata fields such as subject, sender name, header fields and timestamp, processes them through a streamlined NLP pipeline for feature extraction, pr-processing, and embedding, and finally classifies the email as phishing or legitimate using a lightweight machine learning model.

A. Methodology

Data Collection: In order to carry out this research effectively three databases were used to obtain datasets. The Dataset Sources included MeAJOR Corpus email datasets, Phishing validation emails dataset and Open-MalSec v0.1, each of which contains 5000, 3000, and 2000 legitimate emails and phishing email logs respectively making a total of 10,000 datasets altogether.

i. **Filtering:** this was done by extracting only email headers and metadata fields. In order to reduce overheard and ensure privacy as declared by ethical policy as state in General Data Protection Regulation (GDPR), the full message bodies were also filtered out.

Metadata Features Extracted: The various metadata features that were extracted for the research were as highlighted below:

`From`, `To`, `Reply-To`, `Subject`, `Return-Path`

`Received:` headers (including originating IP, timestamps) **`Authentication-Results`** (SPF, DKIM, DMARC)

Timestamp (e.g., time sent, day of week, hour of day)

Subject line text (used with lightweight NLP)

a) ***Preprocessing and Feature Engineering****:* the data collected need to be transformed into more relevant features so that the accuracy as well the predictive power of machine learning model can improved. The following steps were taken into consideration

Text Normalization (NLP on Subject Line): the following actions were taken and converted to lowercase
Remove punctuation, digits, and stop words
Tokenization using lightweight tokenizer
Extract character/word n-grams such as unigrams or bigrams

Metadata-Based Features:
Domain mismatch: Does "**`From`**" domain match `Return-Path`?
Suspicious keywords in `Subject` (e.g., "urgent", "verify", "reset", "account", "click")
Sender IP reputation (optional via external blacklist)
Unusual sending time (e.g., weekend, late night)
Reply-to mismatch with `From` address
Header anomalies: Missing SPF/DKIM/DMARC

B. Feature Engineering

The feature engineering process was carried out by designing features across four groups so as to make the transformation process efficient and easier as seen below:

(i) **Subject Line NLP Features**: TF-IDF vectors, keyword indicators ("urgent", "action required", "verify"). Subject length, punctuation marks count.
(ii) **Sender Features**: this was made up of Domain mismatch indicators and Editing distance between sender domain and organization domain.
(iii) **Header Features**: involved Number of "Received" lines and Presence/absence of "Reply-To".
(iv) **Timestamp Features**: made up of the Out-of-hours sending detection.

C. Models

a. Lightweight Models: The lightweight model was built by using:

Logistic Regression: a statistical analysis method that predicts binary outcome of 'yes' for phishing emails and 'no' for legitimate emails

Random Forest: was employed for the decision-making process

Single-layer MLP (Multi-Layer Perceptron) is a lightweight neural network algorithm type that learns to classify linearly separable patterns

b. Heavyweight Comparison Models: this phase includes the body of the emails, all the text in the emails were considered, though it seems to have gone against the code of conduct and also wastes a lot of processing time. The process was categorised into two as below:

BERT (Bidirectional Encoder Representations from Transformers) is a machine learning model. It is used for NLP tasks analysis. Therefore, it was employed to fine-tuned on full email bodies for sentiment analysis and summerisation of texts

CNN for email body and headers was used to detect and determine patterns on the emails and their headers.

It is important to note here that all lightweight models used only metadata, while heavy models processed the full email content.

IV. EXPERIMENTS

A. Experimental Setup

In a search for efficiency and effectiveness, 80% of the total data was used for model training while 20% was used for real experiments. To make sure the performance of the model was well tested a 5-fold cross-validation was considered in the process.

Metrics: These are the parameters that were used to measure the validity of the experiment on the models carried out. They include Accuracy, Precision, Recall, F1-Score, Inference Time. This can be seen clearly from table 1 below.

B. Baseline Results (Lightweight Models)

Table 1: Baseline results for the lightweight model

Model	Accuracy	Precision	Recall	F1-Score	Inference Time (ms)
Logistic Regression	91.5%	90.7%	92.3 %	91.5 %	3 ms
Random Forest	93.2%	92.8%	93.7 %	93.2 %	5 ms
Lightweight MLP	92.5%	91.9%	93.1 %	92.5 %	6 ms

It shows from table1 that Logistic Regression though low in performance list in all the other metrics ranging from the accuracy through precision recall F1 score, performed excellently in inference time which is recorded as 3ms. This is immediately followed by Random Forest with an Inference Time of 5ms, then Lightweight MLP that marked 6ms Inference Time. Conclusively, when time is critical, Logistic Regression offers the best. While Random Forest offers the best detection performance when accuracy is the priority and Lightweight MLP yields a balance of moderate high accuracy and low latency hence are required where accuracy and speed are required.

C. Ablation Study: Feature Group Importance

We evaluated model performance after removing one group of features at a time: the purpose was to be have a knowledge of the role and importance of each parameter in the exercise.

Table 2: Ablation Study of important group features

Feature Removed	Accuracy Drop
Subject Line Features	-7.2%
Sender Features	-5.5%
Header Features	-3.8%
Timestamp Features	-2.1%

Findings from table 2 reveal that Subject line with and accuracy drop of -7.2% and sender-based features with a drop of -5.5% contribute the most to phishing detection.

D. Lightweight vs Heavyweight Comparison

Table 3: Comparison result of lightweight vs Heavyweight

Model	Accuracy	Inference Time (ms)	Model Size (MB)
Random Forest (Metadata Only)	93.2%	5 ms	22 MB
BERT Fine-tuned (Full Email)	96.8%	480 ms	420 MB
CNN-based Model (Full Email)	95.2%	230 ms	180 MB

Table 3 reveals that the heavy models achieved a higher accuracy of about 3 to 4% gain. This gain is also accompanied by a higher cost of approximately 80 to 100 times, a slower inference and an increase model sizes.

V. DISCUSSION

Our metadata-only approach yields:

Competitive performance with minimal resource consumption.

Superior privacy by avoiding body/content inspection. This also leads to good ethics and conduct compliance

High scalability for large-volume email systems.

The ablation study confirms that subject line and sender metadata are strong indicators of phishing.

VI. CONCLUSION AND FUTURE WORK

This study shows that lightweight phishing detection based on email metadata and NLP are both feasible and highly effective. Without analyzing email bodies, we can achieve real-time phishing protection suitable for scalable, privacy-preserving deployments. For future studies this work can be expanded by integrating the federated learning to update privacy preserving models. It can also be extended to accommodate adversarial robustness testing against crafted metadata by the attackers.

REFERENCES

[1] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques," *IEEE Access*, vol. 10, pp. 65703–65727, 2022, doi: 10.1109/ACCESS.2022.3183083.

[2] H. Kim and J. H. Huh, "Detecting DNS-poisoning-based phishing attacks from their network performance characteristics," *Electron. Lett.*, vol. 47, no. 11, pp. 656–658, May 2011, doi: 10.1049/el.2011.0399.

[3] M. Kosinski, "Evaluating large language models in theory of mind tasks," *Proc. Natl. Acad. Sci.*, vol. 121, no. 45, Nov. 2024, doi: 10.1073/pnas.2405460121.

[4] D. Li, Q. Chen, and L. Wang, "Phishing Attacks: Detection and Prevention Techniques," Aug. 2024, doi: 10.5281/ZENODO.12789572.

[5] M. khonji, A. Jones, and Y. Iraqi, "An Empirical Evaluation for Feature Selection Methods in Phishing Email Classification," 2013.

[6] R. Jayaprakash *et al.*, "Heuristic machine learning approaches for identifying phishing threats across web and email platforms," *Front. Artif. Intell.*, vol. 7, p. 1414122, Oct. 2024, doi: 10.3389/frai.2024.1414122.

[7] M. A. Elberri, Ü. Tokeşer, J. Rahebi, and J. M. Lopez-Guede, "A cyber defense system against phishing attacks with deep learning game theory and LSTM-CNN with African vulture optimization algorithm (AVOA)," *Int. J. Inf. Secur.*, vol. 23, no. 4, pp. 2583–2606, Aug. 2024, doi: 10.1007/s10207-024-00851-x.

[8] P. H. Kyaw, J. Gutierrez, and A. Ghobakhlou, "A Systematic Review of Deep Learning Techniques for Phishing Email Detection," *Electronics*, vol. 13, no. 19, p. 3823, Sept. 2024, doi: 10.3390/electronics13193823.

[9] K. Anindita, "Detection of Phishing Websites Using Data Mining Techniques," vol. 2, no. 12, 2013.

[10] V. Anu, N. B. Harikrishnan, R. Vinayakumar, and K. Soman, "Phishing Email Detection Using Classical Machine Learning," 2018.

[11] R. Verma, and N. Hossain, "Semantic feature selection for text with application to phishing email detection.," *Int. Conf. Inf. Secur. Cryptol. Cham Springer Int. Publ.*, pp. 455-468)., Nov. 2013.

[12] R. Verma, N. Shashidhar, and N. Hossain, "Detecting phishing emails the natural language way," *Comput. Secur. 2012 17th Eur. Symp. Res. Comput. Secur. Pisa Italy Sept. 10-12 2012 Proc. 17 Pp 824-841 Springer Berl. Heidelb.*, pp. 824–841, 2012.

[13] K. Thakur, M. L. Ali, M. A. Obaidat, and A. Kamruzzaman, "z, R., & Hossain, N. (2013, November). Semantic feature selection for text with application to phishing email detection. In international conference on information security and cryptology (pp. 455-468). Cham: Springer International Publishing.," *Electronics*, vol. 12, no. 21, p. 4545, Nov. 2023, doi: 10.3390/electronics12214545.