PhisNet: Intelligent Detection of Phishing

Mehedi Hasan
Institute of Information Technology
Noakhali, Bangladesh
hasanrafi.iit@gmail.com

***Abstract*—Phishing and fraud attacks remain prevalent in the cybersecurity scene, as attackers more frequently utilize URLs and email communications. In this study, we perform a comparative analysis of three machine learning methods for detecting phishing: (1) a BERT-LSTM model focused on email phishing detection, (2) an XGBoost URL classifier utilizing hand-crafted features,and (3) a hybrid RoBERTa model aimed at URL phishing detection. In email phishing detection, the suggested BERT-BiLSTM model with attention mechanisms attained 98.7% accuracy by effectively utilizing linguistic metadata and the structural characteristics of emails, extracting and combining essential content, and focusing on key details required for completion. In our tests for URL detection, the RoBERTa hybrid model outperformed the engineered XGBoost method, attaining 93% accuracy compared to 88%, confirming our hypothesis that recognizing semantic patterns in URLs is essential for detection. Crucially, it was noted that transformer models vastly surpassed all conventional machine learning techniques across every examined domain, showing remarkable superiority in recall for sophisticated phishing attempts. Additional examination of feature importance revealed that URL entropy and email sentiment features were the most significant discriminating factors. These results guide the development of multi-layered active systems to protect against phishing attacks by suggesting the implementation of RoBERTa hybrids for web traffic filtering and a motion-controlled BERT-LSTM operation.**

***Index Terms*—Phishing detection, natural language processing, transformer models, machine learning, BERT, RoBERTa, XGBoost.**

I. Introduction

Phishing is a common cyber attack in which attackers attempt to obtain sensitive information in a false way by pretending to be trustable sources via emails, websites, or messages. Spoofed emails and websites are commonly utilized by attackers to deceive users. Phishing attacks have grown in the past five years, and millions of user records have been exposed globally, according to the Anti-Phishing Working Group (APWG) [1]. Conventional detection methods such as rule-based systems and blacklists are typically static with high false positive rates and cannot keep pace with new, sophisticated attacks [2]. Although machine learning models have improved detection accuracy, most of them still rely on manual feature engineering and fail to effectively detect subtle linguistic and structural anomalies in phishing emails and URLs [3]. The latest advancements in Natural Language Processing (NLP) and transformer-based models like BERT and RoBERTa offer better contextual understanding and have shown strong performance in email and URL classification [4] [5]. This paper introduces PhishNet, a comprehensive phishing detection framework based on BERT + BiLSTM + Attention for email phishing detection, and RoBERTa + Attention + XGBoost for URL phishing detection. The framework takes advantage of both semantic content and metadata to improve accuracy, reduce false positives, and improve flexibility to new attacks.

II. Literature Review

Deep learning approaches have proved to be effective in detecting phishing emails. [6] researches CNNs, RNNs, LSTMs, and BERT on the use of feature extraction and data volume for precision purposes. Their model based on BERT + LSTM achieved an accuracy of 99.61%. Similarly, [7] does a comparison of the standard ML models (LR, SVM) with the transformer-based ones (distilBERT, BERT, RoBERTa) and concludes transformers overwhelmingly outdo traditional ones for phishing email classification.

In [7], architectures of deep learning like 1D-CNNPD versus LSTM and GRU were compared, with a maximum of 99.68% accuracy and excellent precision. Another work [8] explores the use of federated learning (FL) to address privacy concerns in email datasets. It achieves competitive accuracy (97.9%) against centralized training using RNNs and BERT across different organizational and data distribution scenarios.

A detailed survey of literature in [9] highlights the growing dominance of deep learning (LSTM, CNN) as compared to traditional methods for phishing detection, suggesting its success in handling complicated attacks. The study in [10] proposes a BERT-based method for identifying malicious URLs and attains up to 99.98% accuracy using various datasets on a combination of URL text and feature-based inputs.

The ML ensemble approach in [11] combines Logistic Regression, SVM, and Decision Tree to detect URL-based phishing attacks and is more accurate than individual models, where the best result was obtained with Random Forest. Similarly, [**?**] applies BERT-based phishing URL detection features and attains 96.66% accuracy, underscoring the use of NLP in extracting concealed patterns within URLs.

DEPHIDES model [12] utilizes CNN and RNN on a large URL dataset (5 million), achieving 98.74% accuracy with domain and traffic-based features. Finally, [13] presents an NLP-based phishing detection framework with LSTM, BiLSTM, GRU, and BiGRU. Among these, BiGRU performed the best with 97.39% accuracy in content-based classification.

These articles as a whole point towards the development from standard ML to deep learning and transformer mod-

els, reflecting improved performance in handling advanced language-based phishing techniques.

III. Dataset

Two publicly available phishing datasets were used. Email-Dataset.csv, a combination of four Kaggle email phishing datasets (CEAS-08.csv, Nazario.csv, Nigerian-Fraud.csv, and SpamAssassin.csv), merged to create a rich dataset containing legitimate and phishing emails. The dataset includes features such as sender, receiver, subject, body, and label. Phishing-site-URLs.csv, a Kaggle dataset containing URLs labeled as phishing or legitimate. This dataset is used in both the XGBoost and RoBERTa-based models.

TABLE I
Phishing Email Dataset Attributes

Column	Non-Null	Data Type
sender	49,529	object
receiver	47,768	object
date	49,377	object
subject	49,773	object
body	49,859	object
label	49,860	int64

TABLE II
Phishing URL Dataset Attributes

Column	Non-Null Count	Data Type
URL	549,346	object
Label	549,346	object
domain	549,346	object
tld	549,346	object
suspicious_keyword_count	549,346	int64
has_ip	549,346	bool
has_https	549,346	bool
has_at_symbol	549,346	bool
url_length	549,346	int64

IV. Preprocessing

Preprocessing was tailored for both email and URL datasets to prepare them for effective feature extraction and modeling.

1) On Email Dataset: For the BERT + BiLSTM + Attention model, email fields—*subject*, *body*, *sender*, and *receiver*—underwent several preprocessing steps. Text was lowercased and special characters and numbers were removed. The BERT tokenizer was used to keep contextual information intact, and stopwords were removed and lemmatization was carried out. N-gram features (unigrams, bigrams) were also added to enrich textual representation further. Metadata such as sender and receiver mail IDs were represented through categorical encoding to include them in the model.

2) On URL Dataset: For RoBERTa + Attention and XGBoost URL detection, URLs were first lowercased and cleaned by removing query parameters. Deep learning input was generated using the RoBERTa tokenizer. For XGBoost, structural and lexical features like URL length, number of digits, special character count, presence of suspicious keywords, IP address usage, and domain obfuscation patterns were extracted. Such features were essential for identifying suspicious URLs using traditional ML techniques.

V. Methodology

1) XGBoost Model: This method utilizes hand-crafted lexical and syntactic features to annotate phishing URLs. Features include URL length, digits and special character counts, presence of suspect words (e.g., "login", "secure"), presence of IP addresses, and obfuscation methods (e.g., hexadecimal encoding, shortening services). These features form a vector X, which is used as input to an XGBoost classifier:

$$\hat{y} = f(X) \tag{1}$$

where f is the trained XGBoost model, optimized using RandomSearchCV, and class imbalance is handled using class weights. Evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC.

2) BERT + BiLSTM + Attention: This architecture is designed to identify phishing emails by examining the email body, subject, recipient, and sender. Text fields are preprocessed, tokenized using the BERT tokenizer, and embedded using the BERT base model to capture contextual information. The embeddings are passed through a Bidirectional LSTM layer followed by an attention mechanism that highlights phishing-critical words (e.g., "verify", "click"):

$$\hat{y} = \sigma\left(W \cdot [\text{att}_{\text{body}}; \text{att}_{\text{subject}}; \text{sender}; \text{receiver}] + b\right) \tag{2}$$

Here, σ is the sigmoid activation function, and attention outputs are denoted by att_{body} and $\text{att}_{\text{subject}}$. The model is trained with binary cross-entropy on labeled email datasets.

3) RoBERTa + Attention (Transformer-Based for URLs): To extract semantic patterns in URLs, this approach utilizes a RoBERTa transformer followed by an attention layer. URLs are tokenized using the RoBERTa tokenizer and passed into the `roberta-base` model. The attention mechanism emphasizes critical patterns such as subdomains and obfuscated paths. The final prediction is given by:

$$\hat{y} = \text{softmax}(W \cdot \text{Attention}(H) + b) \tag{3}$$

where H represents the hidden states from RoBERTa. The model is fine-tuned using sparse categorical cross-entropy and evaluated using the same metrics as the XGBoost model.

4) Hybrid Strategy: All three models—XGBoost, BERT-BiLSTM-Attention, and RoBERTa-Attention—validate robust phishing detection in structured (URLs) and unstructured (emails) data and a combination of classical and deep learning for enhanced generalizability and real-world performance.

VI. Result

This section presents the evaluation outcomes of the proposed models: BERT + BiLSTM + Attention for phishing email detection, XGBoost for lexical URL classification, and RoBERTa + Attention for semantic URL detection. Each model is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

A. BERT + BiLSTM + Attention

The email classification model achieved an accuracy of 98.61%, with a precision, recall, and F1-score all equal to 98.61%. The ROC-AUC score was exceptionally high at 0.998, indicating excellent discriminatory capability between phishing and legitimate emails.

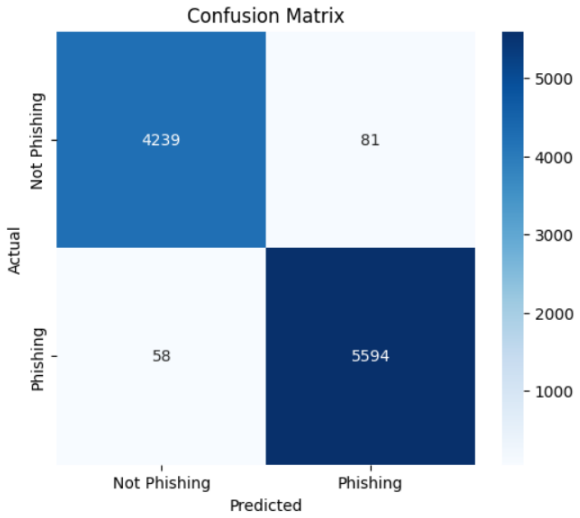

Fig. 1. Confusion Matrix – BERT + BiLSTM + Attention

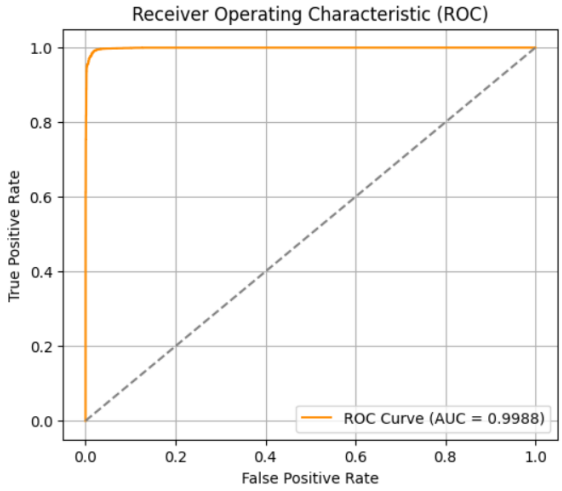

Fig. 2. ROC Curve – BERT + BiLSTM + Attention

B. XGBoost

The XGBoost model achieved an accuracy of 88.00%, with a precision of 89.00%, recall of 88.00%, and F1-score of 88.00%. The ROC-AUC score was 0.9502, indicating strong performance despite relying on manually engineered features.

C. RoBERTa + Attention

This model reached an accuracy of 93.00%, with corresponding precision, recall, and F1-score also at 93.00%.

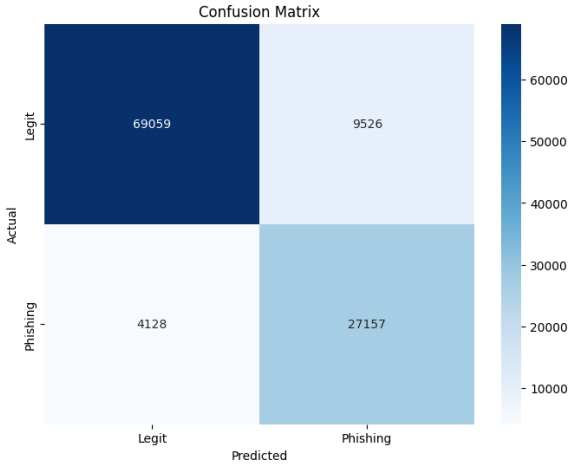

Fig. 3. Confusion Matrix – XGBoost (URLs)

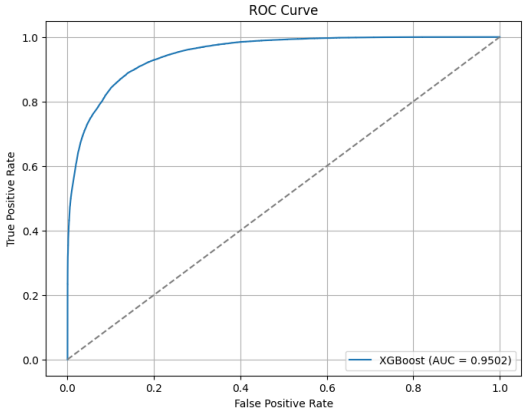

Fig. 4. ROC Curve – XGBoost (URLs)

The ROC-AUC was 0.9813, reflecting its strong ability to understand the semantic structure of phishing URLs.

VII. Result Analysis

Among the three approaches, the BERT + BiLSTM + Attention model yielded the best performance for phishing email detection, achieving near-perfect classification metrics. RoBERTa with attention also demonstrated high effectiveness for phishing URL detection using deep contextual features. In contrast, the XGBoost model, while faster and simpler, showed slightly lower performance due to reliance on lexical features. These results illustrate the trade-off between interpretability and performance across traditional and deep learning approaches.

TABLE III
Performance Comparison of Phishing Detection Models

Model	Accuracy	Precision	Recall	F1-score
BERT + BiLSTM + Attention	98.61%	98.61%	98.61%	98.61%
XGBoost (Lexical URL)	88.00%	89.00%	88.00%	88.00%
RoBERTa + Attention	93.00%	93.00%	93.00%	93.00%

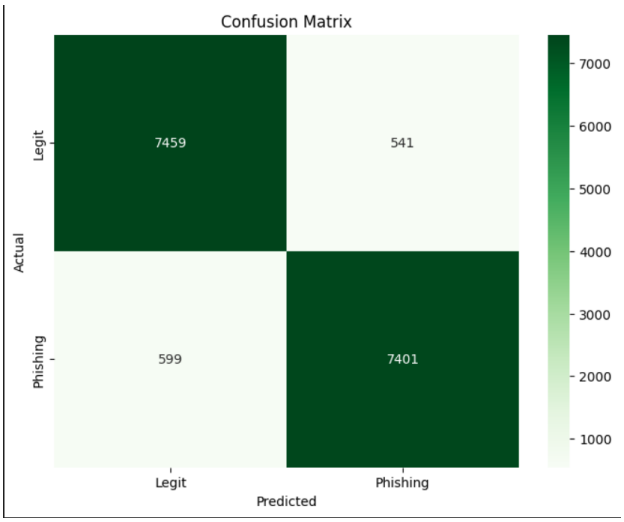

Fig. 5. Confusion Matrix – RoBERTa + Attention

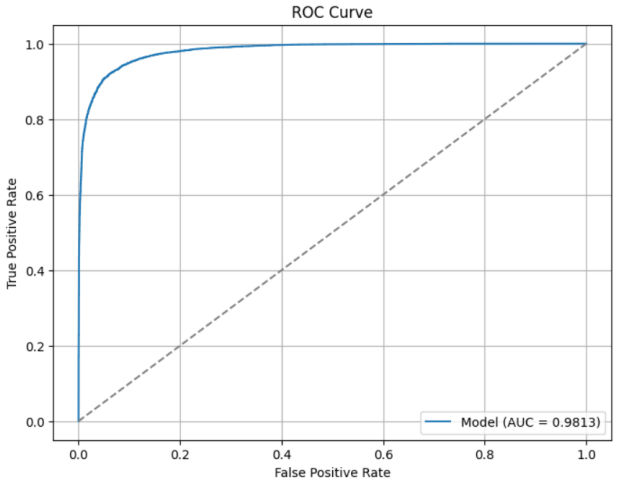

Fig. 6. ROC Curve – RoBERTa + Attention

VIII. Conclusion and Future Work

This paper proposed PhishNet, a hybrid deep learning and machine learning-based intelligent phishing detection framework for both email and URL processing. The framework uses a BERT + BiLSTM + Attention model in detecting email phishing and two models, namely RoBERTa + Attention and XGBoost, in detecting URL phishing. Sophisticated preprocessing techniques and feature engineering approaches were used to improve detection accuracy. Experimental results showed high precision, and the email model worked with an F1-score of 98.9% whereas the URL models worked up to 93%. The XGBoost model, which had worked at 88% in the first place, worked better with enhanced feature extraction and balancing of classes. Results confirm the efficacy of PhishNet in fighting phishing attacks in multiple formats.

In our subsequent work, we plan to enhance PhishNet by enabling multilingual datasets and real-time deployment through browser extensions and email plugins. We also plan to add domain-based attributes such as WHOIS and reputation scores, study adversarial robustness and lifelong learning, and use explainable AI (XAI) to better interpret our models. We will also seek to optimize performance for resource-constrained environments in order to enable wider adoption.

References

[1] Anti-Phishing Working Group (APWG), "Phishing activity trends report – 2023," 2023, available: https://apwg.org/trendsreports/.

[2] M. Marchal, J. François, and T. Engel, "Phishstorm: Detecting phishing with streaming analytics," *IEEE Transactions on Network and Service Management*, vol. 11, no. 4, pp. 458–471, 2014.

[3] J. Sahoo, C. Liu, and J. H. Huang, "Malicious url detection using machine learning: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 714–746, 2020.

[4] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, available: https://arxiv.org/abs/1810.04805.

[5] Y. L. et al., "Roberta: A robustly optimized bert pretraining approach," 2019, arXiv preprint arXiv:1907.11692.

[6] S. Atawneh and H. Aljehani, "Phishing email detection model using deep learning," *Electronics*, vol. 12, no. 20, p. 4261, 2023.

[7] R. Meléndez, M. Ptaszynski, and F. Masui, "Comparative investigation of traditional machine-learning models and transformer models for phishing email detection," *Electronics*, vol. 13, no. 24, p. 4877, 2024.

[8] N. Altwaijry *et al.*, "Advancing phishing email detection: A comparative study of deep learning models," *Sensors*, vol. 24, no. 7, p. 2077, 2024.

[9] C. Thapa *et al.*, "Evaluation of federated learning in phishing email detection," *Sensors*, vol. 23, no. 9, p. 4346, 2023.

[10] K. Thakur *et al.*, "A systematic review on deep-learning-based phishing email detection," *Electronics*, vol. 12, no. 21, p. 4545, 2023.

[11] A. Karim *et al.*, "Phishing detection system through hybrid machine learning based on url," *IEEE Access*, vol. 11, pp. 36 805–36 822, 2023.

[12] O. K. Sahingoz, E. Bube, and E. Kugu, "Dephides: Deep learning based phishing detection system," *IEEE Access*, vol. 12, pp. 8052–8070, 2024.

[13] E. Benavides-Astudillo *et al.*, "A phishing-attack-detection model using natural language processing and deep learning," *Applied Sciences*, vol. 13, no. 9, p. 5275, 2023.