WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Multimodal Transformer-Based Phishing Detection with Explainability and Few-Shot Adaptation

Shikha Tiwari[1], Abhay Kumar Sahu[2]

[1]Assistant Professor Amity University Chhattisgarh
[2]Amity University Chhattisgarh

ABSTRACT :

Phishing attacks have become an increasingly persistent and sophisticated threat in the cybersecurity landscape, targeting users across email, messaging platforms, and websites. Traditional detection systems typically rely on static rule-based methods or unimodal machine learning models, often focusing on either the email's textual content or its associated URL. These approaches suffer from limited adaptability to novel phishing techniques and lack transparency in their decision-making processes. In this work, we propose a novel multimodal phishing detection framework based on transformer architectures that effectively integrates heterogeneous information from email content, header metadata, and embedded URLs. The email body and subject are processed using a fine-tuned DistilBERT model, while handcrafted URL and metadata features are encoded through a dedicated multi-layer perceptron (MLP) pipeline.

To address the challenge of limited labelled data for emerging phishing tactics, we incorporate a few-shot learning strategy via transfer learning, allowing our model to generalize to new attack variants with minimal examples. Furthermore, we introduce model interpretability using SHAP (Shapley Additive explanations), which provides human-readable explanations for each prediction—critical for analyst trust and response.

Keywords : Cybersecurity, Intrusion Detection System, Machine Learning, NSL-KDD, Ensemble Learning.

1. INTRODUCTION

Phishing is one of the most prevalent and dangerous forms of cyberattacks, exploiting social engineering to deceive users into revealing sensitive information such as login credentials, credit card numbers, or personal identity data. These attacks often masquerade as legitimate communications from trusted entities and continue to evolve rapidly in both content and delivery mechanisms. The shift towards more targeted and polymorphic phishing campaigns—such as spear phishing and zero-day phishing emails—has rendered traditional rule-based systems and signature-based detection approaches insufficient. Such systems fail to adapt to new, unseen attack variants and provide little to no insight into their decision-making processes.

In this paper, we introduce a multimodal transformer-based architecture that addresses these limitations by:

- Utilizing email text, metadata, and URL features
- Implementing few-shot adaptation via transfer learning
- Applying SHAP-based interpretability.

With the rise of machine learning (ML) and natural language processing (NLP), several automated phishing detection methods have been proposed. While these methods have demonstrated promising performance, they often rely on unimodal input—focusing either on URL analysis or email text alone. This unimodal approach limits the model's ability to capture contextual dependencies between the various components of an email, such as the relationship between its content, header metadata, and the URL. Moreover, most existing solutions do not offer explainability, making it difficult for cybersecurity analysts to trust or validate the predictions.

In this paper, we address these gaps by proposing a multimodal, transformer-based phishing detection system that fuses insights from three data sources: the email body and subject, the metadata (e.g., sender domain, SPF status), and the embedded URLs.

2. RELATED WORKS

Earlier research has applied various methods such as deep learning models like CNN and LSTM, transformer-based models for analyzing URLs, and ensemble techniques for improving accuracy. While these models have shown good performance, they typically rely on single-modal data. There is limited exploration into approaches that combine multiple data types or incorporate model interpretability and low-data learning strategies.

3. PROPOSED METHODOLOGY

3.1. Multimodal Input Design

We construct the input from three views:
- Email body and subject: Passed into a fine-tuned DistilBERT model
- URL features (length, entropy, presence of IP): Extracted numerically
- Metadata (sender domain, SPF status, timestamp): One-hot or encoded

3.2. Model Architecture

We adopt a hybrid architecture:
- TextEncoder: DistilBERT → [CLS] token
- URLMetadataEncoder: MLP with BatchNorm
- FusionLayer: Concatenation → Dense → Softmax

3.3. Few-Shot Learning

We use transfer learning:
- Pretrained DistilBERT on a large email dataset
- Fine-tuned on phishing samples with as few as 20 examples per class

3.4. Explainability

We use SHAP (SHapley Additive Explanations) on the final fusion layer output to highlight which features (e.g., specific words or URL patterns) contributed to a prediction.

4. DATASET AND PREPROCESSING

We use three datasets: the Enron-Spam email dataset, the Nazario Phishing Corpus, and URL data from PhishTank. Preprocessing steps include HTML stripping, tokenization, and feature extraction from URLs. We also compute URL-based metrics like entropy and the use of special characters. To address class imbalance, the SMOTE technique is applied to oversample phishing examples.

4.1. Datasets Used

- Enron-Spam dataset [9]
- Nazario Phishing Email Corps [10]
- PhishTank URLs [11]

4.2. Preprocessing

- HTML cleaning and tokenization for emails
- Feature extraction: entropy, domain age (via WHOIS), number of dots, '@' symbol, etc.
- Balancing dataset using SMOTE

5. EXPERIMENTS

We evaluate our model against traditional machine learning classifiers and deep learning baselines such as CNN+LSTM and URLTran. Metrics used include accuracy, precision, recall, F1-score, and AUC-ROC. Our model was tested under both full-data and few-shot scenarios.

5.1. Baselines

We compare with:
- Logistic Regression
- Random Forest
- CNN+LSTM [5]
- URLTran [7]

5.2. Metrics

- Accuracy
- Precision, Recall, F1-score
- AUC-ROC
- Explanation confidence (SHAP contribution ranking)

5.3. Experimental Setup

- 80/20 train-test split
- PyTorch and HuggingFace Transformers
- 4-core CPU + NVIDIA T4 GPU (Google Colab Pro)
- 5-shot and 10-shot learning evaluations

5.4. Results summary

Model	Accuracy	F1-score	AUC	Explainable
CNN+LSTM	**92.3%**	**91.9%**	**0.91**	✗
URLTran	**94.2%**	**93.8%**	**0.94**	✗
Proposed Model	**96.1%**	**95.7%**	**0.96**	✓
Proposed (Few-Shot, 5 examples/class)	**92.8%**	**91.5%**	**0.93**	✓

Model	F1-score
CNN+LSTM	**91.9%**
URLTran	**93.8%**
Proposed Model	**95.7%**
Proposed (Few-Shot, 5 examples/class)	**91.5%**

Model	AUC
CNN+LSTM	**0.91**
URLTran 0.94	**93.8%**
Proposed Model	**0.96**
Proposed (Few-Shot, 5 examples/class)	**0.93**

The implementation was done in PyTorch with HuggingFace Transformers. Experiments were run on a system with a NVIDIA T4 GPU. Our model achieved 96.1% accuracy and performed competitively even with only 5 examples per class during few-shot learning.
In cybersecurity environments. Our experimental results showed state-of-the-art performance across multiple benchmarks, demonstrating both improved accuracy and generalization.

6. CONCLUSION

In this study, we presented a novel and comprehensive approach to phishing detection by leveraging a multimodal transformer-based architecture that integrates diverse information sources including email text, header metadata, and embedded URLs. Unlike traditional or unimodal machine learning models, our approach provides a more holistic understanding of phishing attempts by capturing the interdependencies across various features of an email. Through the incorporation of a fine-tuned DistilBERT model for textual analysis and a multi-layer perceptron (MLP) for processing structured metadata and URL features, we were able to fuse these modalities effectively to achieve robust classification results.

A key strength of our framework is its ability to perform few-shot learning, allowing it to adapt quickly to new and unseen phishing techniques with minimal training data. This capability is especially important in real-world scenarios where phishing attacks continuously evolve, and labeled data is often scarce. Moreover, by integrating SHAP-based interpretability, we addressed the crucial need for transparency and trust in machine learning systems deployed in future work, we aim to deploy this model in real-time email systems, explore integration with mobile security platforms, and further enhance resilience against adversarial attacks using robust training strategies and continual learning techniques.

REFRENCES :

1. Shikha Tiwari; Abhinav Pandey; Hardik Khamele; Nitish Kumar Rathore; Abhay Kumar Sahu, "Blockchain-Powered Decentralized Platforms for Secure Healthcare Data Exchange", Published in: 2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG).
2. Shikha Tiwari; Ch. Meher Babu; P Shanker; Shahnaz K V; Vandana Roy; Ramgopal Kashyap, "Cross-Lingual Transfer Learning in RNNs for Enhancing Linguistic Diversity in Natural Language Processing", Published in: 2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET).
3. Shikha Tiwari; Ramgopal Kashyap; Vandana Roy, "Integrating Deep Learning to Decode Meningeal Interleukin-17 T Cell Mechanisms in Salt-Sensitive Hypertension-Induced Cognitive Impairment", Published in: 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0.
4. Shikha Tiwari; Rohan Mandhani; Anup Lakra; Aniket Patel, "AI and Blockchain Synergy for Advanced Health Data Processing in IoT", Published in: 2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG).
5. A. Abdelhamid et al., "Phishing Detection Based on Machine Learning Algorithms," IEEE, 2014.
6. L. Bahnsen et al., "DeepPhish: Simulating Malicious AI," Security and Privacy, 2018.
7. M. Marchal and J. François, "URLTran: Improving Phishing URL Detection Using Transformers," arXiv, 2021.
8. S. Gangwar and K. Ghosh, "Phishing Detection using Base Classifier and Ensemble Technique," IJRITCC, 2023.
9. J. Klimt and Y. Yang, "The Enron Corpus," Machine Learning: ECML, 2004.
10. Jose Nazario, "Phishing Email Corpus."
11. PhishTank, "Open Phishing Threat Intelligence," https://www.phishtank.com/