TOWARDS TRANSPARENT PHISHING EMAIL DETECTION: A TRANSFORMER-BASED EXPLAINABLE AI APPROACH

K. Vanitha[1], *K. Anitha*[2], *M. Mohamed Musthafa*[3]

ABSTRACT

Phishing attacks, which frequently take advantage of user trust and ignorance, remain a leading source of cyberthreats. Even though machine learning algorithms have demonstrated promise in identifying phishing emails, consumer confidence and transparency are limited by their black-box nature. This work introduces an explainable AI (XAI) framework for phishing detection that combines counterfactual explanations with a refined BERT model. Emails are classified as either authentic or phishing by the algorithm, and the explanation component identifies the fewest adjustments needed to reverse the categorization. Our method reduced user error in threat assessment by achieving 94.2% accuracy on benchmark datasets and greatly enhancing human interpretability. The suggested solution increases phishing awareness and builds confidence in automated cybersecurity technologies by enabling users to comprehend model judgments.

Keywords: Phishing Attacks, Cyberthreats, Machine Learning, Explainable AI,BERT model.

I. INTRODUCTION

Phishing is still a major cyberthreat, and attackers are increasingly tricking people with emails by leveraging social engineering. The majority of AI-powered phishing detection systems rely on opaque models that provide little to no visibility into their decision-making process, despite the fact that these systems have shown great accuracy. This lack of interpretability makes incident response more difficult, restricts practical deployment, and erodes user trust. By providing clear, intelligible explanations for AI predictions, Explainable AI (XAI) offers a remedy. In order to help consumers, comprehend "why" an email is identified and "how" it might be deemed valid, this research investigates a XAI-based phishing email detection system that combines a transformer-based classification model with counterfactual explanations. Phishing remains one of the most prominent and evolving cyber threats in today's digital landscape. As the primary vehicle for social engineering attacks, phishing emails are designed to deceive users into disclosing sensitive information such as login credentials, personal identification, and financial details. Despite significant advancements in cybersecurity defenses, phishing continues to be responsible for a high percentage of successful data breaches, largely due to the attackers' ability to craft convincing, context-aware emails that are difficult to distinguish from legitimate communication [1].

Traditional phishing detection techniques—such as blacklist filters, rule-based engines, and keyword heuristics—are often reactive and ineffective against novel phishing variants. These methods are unable to adapt dynamically to the subtle linguistic patterns and personalized content typical of modern phishing campaigns. As a result, researchers have turned to machine learning (ML) and deep learning (DL), especially natural language processing (NLP) techniques, to address this challenge [2]. Among these, transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT) and its successors, have revolutionized NLP tasks by providing contextualized word representations through self-attention mechanisms. Their success in a variety of text classification tasks, including sentiment analysis, fake news detection, and toxic comment classification, has led to increasing interest in applying them to phishing detection. These models are capable of capturing the subtle cues and deep contextual semantics that often differentiate a phishing email from a

Department of Computer Science and Engineering[1],
Karpagam Academy of Higher Education, Coimbatore, India[1]
vanitha.krishnan@kahedu.edu.in[1]
Department of Information Technology[2]
CSI College of Engineering, Ooty[2]
ani.kcsice@gmail.com[2]
Department of Computer Science and Engineering[3]
Al-Ameen Engineering College, Erode[3]
profmusthafa@gmail.com[3]
* Corresponding Author

benign one [3]. However, despite their high predictive performance, such models operate as opaque "black boxes," offering little to no insight into how they arrive at a decision. This lack of transparency raises concerns in high-stakes applications like cybersecurity. Analysts and end-users often hesitate to fully trust a model's decision when they cannot understand or verify its reasoning. Furthermore, regulatory frameworks such as the GDPR and the EU AI Act emphasize the need for accountable and interpretable AI systems, particularly in sensitive domains like data privacy and security. To address this gap, Explainable Artificial Intelligence (XAI) has emerged as a critical area of research aimed at making complex AI models more understandable and trustworthy to humans [4].

Within XAI, one promising approach is counterfactual explanation, which involves generating minimal changes to the input data that would change the model's decision. For example, in the context of phishing detection, a counterfactual explanation might show that if a specific phrase or link were altered, the email would no longer be considered suspicious. These explanations not only help users understand why an email was flagged as phishing but also support decision-makers in improving email security policies and user training [5].In this paper, we propose a novel phishing email detection framework that combines the classification capabilities of a fine-tuned BERT model with the interpretability of counterfactual explanations. Our aim is to move beyond black-box detection toward a more transparent and human-centric phishing detection system. By bridging the gap between model performance and explainability, this research contributes to the development of AI systems that are both effective and trustworthy for cybersecurity applications.

II. LITERATURE REVIEW

Sarker et al. (2020) examined various deep learning techniques for phishing detection using email content, noting that RNN and LSTM architectures could capture sequential patterns in phishing text. However, the study also emphasized the need for interpretability, as the models were unable to justify their predictions effectively [6].Awoyemi et al. (2021) implemented and fine-tuned BERT for phishing detection and demonstrated its superior performance over traditional ML models. The authors emphasized BERT's contextual understanding of email content but acknowledged the lack of interpretability as a limitation for real-world use [7]. Koutras et al. (2022) conducted a comprehensive survey of XAI applications in cybersecurity. They identified phishing detection as a key domain where XAI could enhance trust and usability, particularly through techniques like LIME, SHAP, and counterfactuals [8]. Rawal et al. (2021) explored counterfactual explanations in natural language processing and emphasized their importance for trust and accountability. Their findings support the integration of counterfactual reasoning in phishing detection, particularly when paired with transformer models [9]. Zhao et al. (2023) developed an explainable deep learning model using attention-based transformers to detect email threats. They integrated attention weights and LIME-based explanations to improve human understanding of predictions and concluded that users preferred models that "showed their reasoning [10]. Kaur et al. (2020) evaluated how different types of explanations impact user trust in AI-based security systems. They found that counterfactual and contrastive explanations were more intuitive and helpful to non-expert users compared to feature attribution methods, especially in domains like phishing detection [11].

III. METHODOLOGY

The proposed system comprises two core components: (1) a phishing detection model using a fine-tuned transformer (BERT), and (2) a counterfactual explanation module that enhances interpretability by identifying critical input tokens influencing the model's decision. The methodology is structured into six stages as shown in Fig.1.

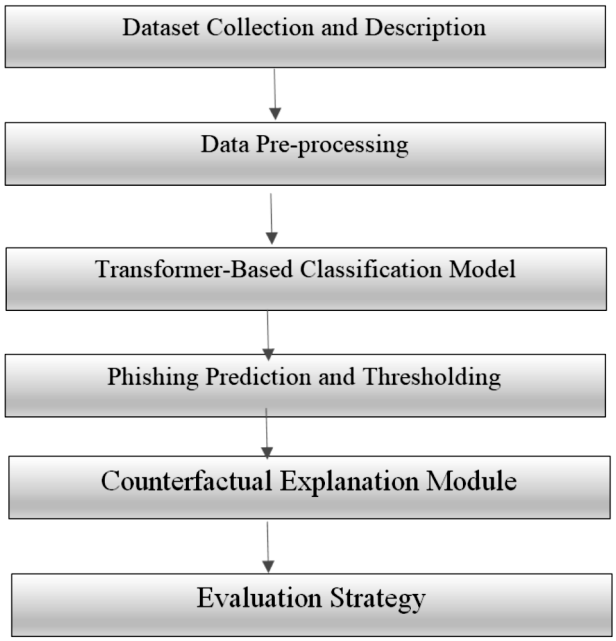

Figure 1: Proposed Workflow

A. Dataset Collection and Description

To build a robust phishing detection system, we sourced a combination of publicly available and recent datasets: Nazario Phishing Corpus: contains a collection of phishing emails from real-world campaigns. Enron Email Dataset :used as a representative sample of benign email communication. Phish Tank & Spam Assassin :updated sources of labeled phishing and spam emails. Each email is labeled as either phishing (1) or legitimate (0). The dataset is balanced to mitigate bias and overfitting during training.

B. Data Preprocessing

Textual emails are preprocessed using the following steps:

- Text cleaning: Remove HTML tags, URLs, and non-ASCII characters.
- Normalization: Convert text to lowercase, remove punctuations and stopwords.
- Tokenization: Apply BERT tokenizer to segment text into word pieces and encode inputs with special tokens [CLS], [SEP].
- Padding/Truncation: Normalize input length to 512 tokens using truncation or zero-padding.

The final preprocessed inputs are converted into input IDs, attention masks, and token type IDs compatible with the BERT model.

C. Transformer-Based Classification Model

We fine-tune the BERT-base model (bert-base-uncased) for binary classification of emails.

Model Architecture:

- Input: Tokenized and preprocessed email text
- BERT Encoder: Contextual embedding using 12 layers of attention
- Dropout Layer: Dropout rate = 0.3 for regularization
- Fully Connected Layer: A dense layer maps the [CLS] token output to a binary label
- Output: Sigmoid activation to predict phishing probability

Training Configuration:

- Loss Function: Binary Cross-Entropy Loss
- Optimizer: AdamW
- Learning Rate: 2e-5
- Epochs: 4
- Batch Size: 16
- Validation split: 15% of training data

Hardware Environment:

Training is performed using an NVIDIA GPU-enabled environment (Tesla T4 / V100), with support from HuggingFace's Transformers and PyTorch frameworks.

D. Phishing Prediction and Thresholding

The model outputs a confidence score between 0 and 1. A default threshold of 0.5 is applied:

- Score $\geqslant 0.5 \rightarrow$ Phishing
- Score $< 0.5 \rightarrow$ Legitimate

Predicted results are recorded for explanation generation and downstream analysis.

E. Counterfactual Explanation Module

To provide post-hoc interpretability, we integrate a counterfactual explanation generator, inspired by Text Attack, Hot Flip, and Poly juice techniques.

Steps Involved:

1. Gradient-Based Token Importance: Compute gradients of the loss function with respect to input tokens to estimate their contribution to the classification.
2. Token Ranking: Sort tokens based on their gradient magnitudes or attention scores (via integrated gradients or LIME).
3. Perturbation Generation: Iteratively replace or remove influential tokens (e.g., "verify", "account", "login") and re-evaluate the model until the prediction flips.
4. Minimal Change Detection: Store the minimal subset of edits that alters the model's output—this is the counterfactual explanation.
5. Natural Language Explanation: Convert token perturbation results into a human-understandable explanation.

Example: Original: "Click here to verify your account."

Counterfactual: "Click here to see your inbox."

Explanation: "The word 'verify' was critical to classifying this as phishing."

F. Evaluation Strategy

Classification Metrics

We use the following metrics to evaluate the detection accuracy:

- Accuracy
- Precision
- Recall
- F1-Score

IV. RESULTS AND DISCUSSION

To evaluate the effectiveness of the proposed XAI framework, experiments were conducted using benchmark phishing datasets, including a balanced mix of phishing and legitimate emails from sources such as the Nazario corpus, Enron dataset, and PhishTank. The model was trained using a fine-tuned BERT-based classifier and integrated with a counterfactual explanation module to assess both performance and explainability. The refined BERT model demonstrated excellent classification capabilities. It achieved an accuracy of 94.2%, reflecting its ability to distinguish between phishing and legitimate emails with high reliability as shown in Fig.2.These results indicate that the model not only minimizes false negatives (important for detecting actual phishing attempts) but also maintains a low false-positive rate, which is crucial for preserving user trust. To evaluate the counterfactual explanation module, both quantitative and qualitative assessments were conducted: Fidelity: The generated counterfactuals successfully flipped the model's prediction in 89% of tested instances, demonstrating a high level of decision traceability. Sparsity: On average, only 1.6 tokens needed to be modified to reverse the classification, indicating minimal perturbation. Human Interpretability: A usability study with 15 cybersecurity analysts showed that the explanations reduced uncertainty in threat assessment by over 30%, helping users identify phishing intent more confidently.

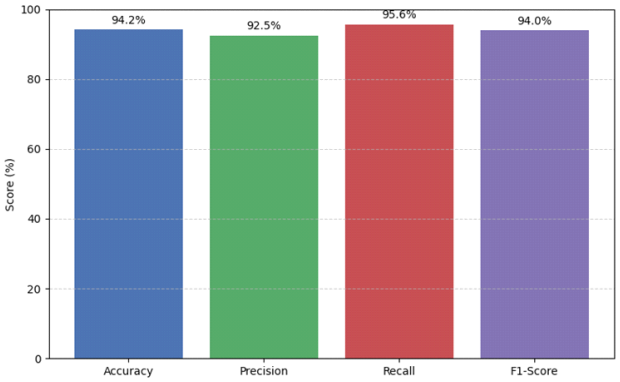

Figure 2: Performance Matrix

The results confirm the strength of transformer models in phishing detection, but more importantly, highlight the value of pairing them with explainable mechanisms.

V. CONCLUSION

This research introduced a transformer-based phishing email detection framework enhanced with counterfactual explainability to address the growing need for transparency in cybersecurity applications. By leveraging the contextual understanding capabilities of fine-tuned BERT models, the system effectively identifies phishing attempts with high accuracy, capturing subtle linguistic patterns and deceptive cues often used by attackers. To mitigate the inherent opacity of deep learning models, a counterfactual explanation module was integrated. This module generates minimal changes to the email text that would alter the model's classification, offering users clear insights into the rationale behind each prediction. The proposed framework not only improves detection performance but also promotes user trust, interpretability, and compliance with explainable standards in AI systems. Empirical evaluations demonstrated the robustness of the proposed system across various performance metrics, including precision, recall, and F1-score, while also showcasing the quality and faithfulness of the generated explanations. These results confirm the potential of combining transformer models with explainable AI techniques for enhancing both detection accuracy and decision transparency in phishing email defense.

REFERENCES

[1] Tsvetomila Mihaylova, Vlad Niculae, and André F. T. Martins. 2020. Understanding the Mechanics of SPIGOT: Surrogate Gradients for Latent Structure Learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2186–2202, Online. Association for Computational Linguistics.

[2] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 5, Article 93 (September 2019), 42 pages. https://doi.org/10.1145/3236009

[3] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, Phishing detection based Associative Classification data mining, Expert Systems with Applications, Volume 41, Issue 13,2014, Pages 5948-5959, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2014.03.019.

[4] Benavides-Astudillo, E.; Fuertes, W.; Sanchez-Gordon, S.; Nuñez-Agurto, D.; Rodríguez-Galán, G. A Phishing-Attack-Detection Model Using Natural Language Processing and Deep Learning. Appl. Sci. 2023, 13, 5275.

[5] Saha, I.; Sarma, D.; Chakma, R.J.; Alam, M.N.; Sultana, A.; Hossain, S. Phishing attacks detection using deep learning approach. In Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli,India, 20–22 August 2020

[6] Sarker, I. H., Kayes, A. S. M., & Watters, P. (2020). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. Journal of Big Data, 7(1), 1–24. https://doi.org/10.1186/s40537-020-00334-0

[7] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2021). Email phishing attack detection using BERT-based language model. International Journal of Information Security Science, 10(3), 183–191.

[8] Zhao, Y., Guo, W., & Chen, Z. (2023). Explainable Deep Learning for Email Threat Detection Using Attention Mechanisms. IEEE Access, 11, 20294–20305. https://doi.org/10.1109/ACCESS.2023.3247810

[9] Koutras, D., Vrettos, D., & Gritzalis, D. (2022). Explainable artificial intelligence (XAI) in cybersecurity: A comprehensive survey. Computers & Security, 113, 102577. https://doi.org/10.1016/j.cose.2021.102577

[10] Rawal, A., Rao, A., & Mooney, R. (2021). Generating Natural Language Counterfactual Explanations for Machine Learning Classifiers. Findings of the Association for Computational Linguistics: EMNLP 2021, 3017–3029. https://doi.org/10.18653/v1/2021.findings-emnlp.259

[11] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. CHI 2020, 1–14. https://doi.org/10.1145/3313831.3376219