

Article

Phishing Email Detection Using BERT and RoBERTa

Mariam Ibrahim * and Ruba Elhafiz

Department of Mechatronics Engineering, German Jordanian University, Amman 11180, Jordan;
r.elhafiz@gju.edu.jo

* Correspondence: mariam.wajdi@gju.edu.jo

Abstract

One of the most harmful and deceptive forms of cybercrime is phishing, which targets users with malicious emails and websites. In this paper, we focus on the use of natural language processing (NLP) techniques and transformer models for phishing email detection. The Nazario Phishing Corpus is preprocessed and blended with real emails from the Enron dataset to create a robustly balanced dataset. Urgency, deceptive phrasing, and structural anomalies were some of the neglected features and sociolinguistic traits of the text, which underwent tokenization, lemmatization, and noise filtration. We fine-tuned two transformer models, Bidirectional Encoder Representations from Transformers (BERT) and the Robustly Optimized BERT Pretraining Approach (RoBERTa), for binary classification. The models were evaluated on the standard metrics of accuracy, precision, recall, and F1-score. Given the context of phishing, emphasis was placed on recall to reduce the number of phishing attacks that went unnoticed. The results show that RoBERTa has more general performance and fewer false negatives than BERT and is therefore a better candidate for deployment on security-critical tasks.

Keywords: natural language processing; bidirectional encoder representations from transformers; robustly optimized BERT pretraining approach; peace; justice and strong institutions

1. Introduction

Phishing continues to be one of the most prevalent and harmful forms of cyberattack, preying on human behavior via emails and other online content. Recent research indicates that a large portion of data breaches stem from phishing, and attackers are always evolving their methods to avoid detection [1]. Unlike traditional forms of malware that exploit a system's technical weaknesses, phishing is much harder to defend against since it relies on social engineering.

Phishing techniques are always evolving, and conventional detection methods like blacklisting and rule-based filtering are too static to keep up [2]. These methods fall short of defending against disguised URLs, zero-day attacks, or even adversarial text manipulation. To tackle these issues, there is an increasing focus on machine learning and, more recently, deep learning to uncover the intricate linguistic and contextual clues of phishing emails that, until now, have gone unnoticed [3].

Natural Language Processing (NLP) contributes significantly to this effort by providing modern techniques for pattern analysis, meaning extraction, sense making, and anomaly detection [4]. BERT-based and RoBERTa [5] transformer models have performed

Academic Editor: Roman Makarov

Received: 25 November 2025

Revised: 2 February 2026

Accepted: 6 February 2026

Published: 7 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

exceptionally well on text classification due to contextual embeddings and attention mechanisms. Unlike earlier approaches based on static word embeddings, these models generate contextually relevant embeddings that are useful for identifying complex phishing linguistic signals.

Nevertheless, phishing detection remains problematic. A good number of studies are conducted on datasets that are small, lack linguistic heterogeneity, and do not perform well on obfuscated or adversarial inputs [6]. Moreover, evaluation is often focused on the accuracy metric, which is problematic for security-sensitive applications that require a high recall for the metric [7]. Using the Nazario Phishing Corpus, the transformer-based NLP models BERT and RoBERTa are evaluated in detecting phishing emails. In this case, the architecture of the framework is designed to process the data and conduct fine tuning, after which evaluation on the relevant metrics will be carried out to test the accuracy and robustness of the model.

This paper emphasizes the following key contributions:

1. Creation of a phishing detection pipeline utilizing NLP and transformer models, BERT and RoBERTa, tailored to the Nazario Phishing Corpus.
2. Comprehensive testing and evaluation of the models (with deliberate character, whitespace, and URL obfuscations in emails), focusing on minimizing false negatives with a detailed assessment of the accuracy, precision, recall, and F1-score.
3. Analysis of BERT and RoBERTa models examining the efficiency and performance and adaptability trade-offs to differing phishing content.
4. Analysis of existing gaps and possibilities towards further investigation with the transformer-based NLP model for phishing detection.

2. Literature Review

In recent years, transformer-based NLP models have become increasingly prominent in phishing detection research. For example, the experimental results in [8] demonstrate that the improvement of the AI detection model effectively covers different templates of phishing emails of the same type in black and gray production, as well as variant detection that avoids detection. This approach was proposed in response to the poor detection performance of social work email phishing attacks in complex dimensions. The BERT training model is used for detection. However, compared to AI detection model methods, the detection impact based on the combination of big models and conventional rules is superior.

The Improved Phishing and Spam Detection Model (IPSDM) was introduced by [9] to detect phishing and spam emails, particularly by fine-tuning the BERT family of models. The author shows how IPSDM, the improved version, can more accurately categorize emails in both balanced and unbalanced datasets. The use of LLMs to strengthen information system security is made possible in large part by this study. But the model must be improved by adding more variety and enhancing training performance

PhishingGNN is presented in [10] to detect subtle phishing trends that are missed by conventional algorithms. This hybrid model combines Graph Attention Networks (GATs) to represent email metadata and content as graph structures with DistilBERT for context-aware text analysis. Despite the model's good generalization across benchmarks, the historical breadth of the CEAS_08 dataset could not accurately reflect modern phishing techniques, and there was no proof of effectiveness against new threats like multilingual material, targeted spear-phishing, or LLM-generated assaults.

Other work has investigated alternative transformer families. For instance, [11] introduced Phishsense-1B, a LLaMA-based detection framework fine-tuned with GuardReasoner. While it outperformed BERT and RoBERTa, the system struggled with generalization, as its accuracy declined on unseen, real-world phishing examples. Similarly, researchers have extended detection to Uniform Resource Locator (URL) and web content.

For instance, [12] presented BERT-PhishFinder, a transformer-based model for URL categorization, but noted limited coverage for multilingual and obfuscated URLs.

Further developments include teacher–student frameworks. For instance, [13] proposed BERTPHIURL, where RoBERTa acted as the teacher for a distilled version (Distil-RoBERTa). This approach maintained efficiency without major loss in accuracy, though the authors argued that knowledge distillation and dynamic model updating require more research to keep pace with evolving phishing tactics.

Several studies have also targeted mobile or resource-constrained environments. Reference [14] introduced Phish-Jam, a mobile ensemble that combined hand-crafted features with transformer embeddings. Although it achieved near real-time detection, the approach faced energy and hardware limitations on low-end devices. Along similar directions, [15] designed PhishTransformer, a hybrid architecture using convolutional neural networks (CNNs) and transformers to detect malicious embedded URLs. While performance was very high, the complexity of training raised barriers for deployment.

Explainability has also become an area of focus. Reference [16] presented EXPLI-CATE, which integrates XAI techniques such as SHAP and LIME into transformer-based phishing detection. The authors observed, however, that interpretability varied across different types of input, and user studies are needed to validate explanation quality. A related contribution, [17], introduced the Language Pack Tuned Transformer Language (LPTTL). By exploiting intentional misspellings (cacography), LPTTL was able to identify ransomware-related phishing emails with high accuracy. Still, handling mixed-language text and irregular spellings remains a challenge, requiring multilingual datasets and standardized orthography.

Other approaches combine linguistic and structural signals. Reference [18] proposed a multimodal framework that integrates BERT, RoBERTa, A Lite BERT (ALBERT), and the Character-Aware Neural Language Model (CANINE) with tabular classifiers to analyze HTML content. While being effective, this hybrid method increased both model complexity and inference time. Reference [19] developed a transformer-based system incorporating Vector Similarity Search (VSS) to quickly compare suspicious inputs with known phishing patterns. Although promising, its reliance on up-to-date embedding databases limits its adaptability to fast-changing phishing material.

In this paper, we sought to overcome several limitations that were identified in previous studies, such as developing a balanced and realistic dataset, which tends to increase evaluation reliability. In addition, we used tokenization, lemmatization, and noise filters, as well as disregarded phishing-specific features to improve resilience. We also performed a head-to-head evaluation of BERT and RoBERTa using the identical preprocessing, training, and evaluation procedures. This separates the effects of pretraining tactics and model design on phishing detection.

3. Preliminaries

3.1. Natural Language Processing (NLP)

Making it feasible for computers to understand, generate, and interpret human language is the aim of NLP, a branch of linguistics and artificial intelligence (AI). NLP may be conceptualized mathematically as a probabilistic modeling task: given a corpus $\mathcal{C} = \{w_1, w_2, \dots, w_n\}$, where each w denotes a token, the goal is to learn parameters θ that minimize a loss like cross-entropy or enhance the likelihood of observed data. Stochastic gradient descent with backpropagation using neural architectures is commonly used for this optimization [20].

Tokens are first separated from the text to create a vocabulary V . After mapping each token $w \in V$ to a continuous vector space, embeddings $v_w \in \mathbb{R}^d$ are obtained, where d denotes the dimensionality of the embedding vector. This mapping is formalized by

Word2Vec models like Continuous Bag-of-Words (CBOW) and Skip-Gram, which optimize via negative sampling and predict target words from context or vice versa. By reducing the distance between words that appear in comparable contexts, these embeddings convey semantic similarity [21].

Prior to deep learning, NLP depended on n-gram models, which estimated the likelihood of a sequence using a fixed history of size $n - 1$. For instance, a trigram model can approximate:

$$P(w_i | w_{i-1}, w_{i-2}) \approx \frac{\text{count}(w_{i-2}, w_{i-1}, w_i)}{\text{count}(w_{i-2}, w_{i-1})} \quad (1)$$

Such models assume the Markov property (the future is determined exclusively by the last $n - 1$ words). While useful for a tiny vocabulary, they are sparse and lack generalizability [22].

Hidden Markov Models (HMMs) are commonly employed in sequence labeling applications such as part-of-speech identification. The Viterbi method is an efficient dynamic programming technique for calculating the most likely sequence of hidden states given observed tokens.

$$\delta_t(s) = \max_{r \in S} (\delta_{t-1}(r) \cdot a_{r,s} \cdot b_{s,o_t}) \quad (2)$$

where $a_{r,s}$ are transition probabilities and b_{s,o_t} are emission probabilities [23].

Neural networks, particularly recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have played an important role in modeling longer-term language connections. Transformer architectures changed NLP by depending solely on self-attention rather than repetition. The primary operation is scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

A series of queries is packed into a matrix Q and the attention function is calculated on each query at the same time. Additionally, the keys and values are crammed into matrices K and V , where the scaling factor is $\sqrt{d_k}$.

Transformers enable concurrent training and capture global dependencies, making them the basis for modern pre-trained models like BERT and RoBERTa [24].

3.2. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a groundbreaking pre-trained language model based on the Transformer encoder architecture. Unlike unidirectional models, BERT reads context from both the left and right concurrently in all levels, allowing for extensive contextual comprehension. The architectural Details and Mathematical Foundations include the following:

Input Representation: BERT turns input text into embeddings by adding Token embeddings, Position embeddings, and Segment (token-type) embeddings. The encoder stack receives this vector representation X_0 as input [4].

Transformer Encoder Stack: BERT is an encoder-only stack of Transformer blocks. Formally, the model consists of L identical layers (e.g., 12 for BERT_BASE, 24 for BERT_LARGE), each with [25] multi-head self-attention, which is computed per head as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

The feed-forward sublayer is a fully linked two-layer network with ReLU activation. Each sublayer has residual connections and layer normalization.

Pre-training Objectives: BERT is trained with two unsupervised tasks that allow for contextual comprehension: Masked Language Modeling (MLM), which involves randomly masking (usually 15%) tokens and training a model to predict the original words using

bidirectional context, and the Next Sentence Prediction (NSP) model, which predicts whether sentence pairs (A, B), follow each other consecutively. This facilitates the capturing of sentence-level coherence, which is useful for tasks such as quality assurance and natural language inference.

Once pre-trained, BERT may be fine-tuned for downstream tasks by layering a task-specific head (e.g., classification layer) on top of the encoder output and training together [4].

3.3. Robustly Optimized BERT Pretraining Approach (RoBERTa)

RoBERTa is a revised reimplementation of BERT that improves the pretraining technique by making many key changes: training longer with larger batches, deleting the Next Sentence Prediction (NSP) target, utilizing dynamic masking, and training on substantially larger datasets. These adjustments result in significant performance increases across popular NLP benchmarks, including GLUE, RACE, and SquAD [5]. The algorithm and mathematical foundations include:

Input Encoding and Transformer Architecture: RoBERTa uses the encoder-only transformer architecture created by BERT, which features stacked transformer layers with multi-head self-attention, feed-forward sublayers, and residual connections. Input token sequences are similarly embedded using token and positional embeddings [5].

Optimized Pretraining Strategy: RoBERTa improves on BERT's training recipe by [5]: larger batch sizes and longer training sessions that lead to improved optimization and representation quality, removing NSP, as earlier studies indicated it may not be necessary for downstream performance, dynamic masking allows mask patterns to alter each epoch or batch, improving generalization, and using full-sentence inputs instead of fixed-length fragments using Next Sentence Prediction (NSP) mechanisms. These strategic modifications refine the optimization process, increasing capabilities for subsequent jobs.

Training Objective: RoBERTa continues to employ MLM, which predicts masked tokens using a cross-entropy loss but does not include the NSP component, simplifying the loss function to focus just on language modeling [5].

Empirical Performance: Controlling for data, masking approach, and hyperparameters, RoBERTa regularly outperforms the original BERT performance. It improves accuracy across GLUE tasks (e.g., MNLI, QNLI, RTE, STS-B), SquAD, and RACE, demonstrating that superior training dynamics, rather than architecture, are the primary driver of increased performance [5].

3.4. Datasets

Nazario Phishing Corpus Dataset: Jose Nazario, a cybersecurity researcher, established the Nazario Phishing Corpus, which was one of the first publicly available collections of phishing emails. The dataset was published to give academics real-world phishing samples to explore language trends, social engineering methods, and automated detection systems [26]. The corpus comprises around 9500–11,500 phishing emails, depending on the version utilized for study. Emails were acquired during the early 2000s, when phishing became a common criminal method. Emails are saved in plain text format (RFC 822/MIME), which includes headers (e.g., sender, subject, routing information) and body content (HTML or plain text). This early phishing dataset is one of the most legitimate and diversified, as it was obtained via user-reported campaigns and spamtrap with the following content characteristics: frequent use of spoofed domains (e.g., PayPal, eBay, banking institutions), urgency cues in subject lines ("urgent," "verify immediately"), embedded URLs redirecting to fraudulent websites, and requests for credentials or financial information.

Enron Corpus Dataset: To supplement the Nazario Phishing Corpus, which only contains phishing emails, we used the Enron Corpus as a source of real (ham) emails. The Enron dataset was revealed during the Enron Corporation inquiry and includes about 500,000 real-world emails from senior management. This corpus has now become a frequently used standard in text mining and email classification research due to its size and realistic portrayal of non-malicious corporate communication [27]. By selecting valid emails from the Enron corpus, we created a balanced collection of phishing and legitimate communications, allowing for binary classification and adequate detection performance measurement.

From the Enron Corpus, legitimate (ham) emails were selected by randomly sampling messages from the cleaned dataset after removing duplicates, empty emails, and system-generated messages. Sampling was performed using a fixed random seed to ensure reproducibility. No additional content-based filtering was applied, preserving the natural diversity of legitimate corporate communication.

Although the Nazario and Enron datasets are considered legacy corpora, they continue to serve as reproducible benchmarks and enable longitudinal comparisons in recent phishing-detection literature. Importantly, phishing emails consistently exhibit persistent patterns such as urgency cues, impersonation, requests for user credentials, and compliance solicitation. To verify that these datasets remain relevant, we examined phishing terms and structures and observed strong alignment with contemporary phishing reports. Consequently, this study focuses on analyzing model behavior, robustness, and recall-oriented performance within established benchmark datasets rather than claiming dataset novelty.

The final consolidated dataset contains 19,000 emails, comprising 9500 phishing samples from the Nazario corpus and 9500 legitimate samples from the Enron corpus. Using a fixed random seed, the dataset is randomly split into training (80%) and testing (20%) subsets. All reported performance metrics and confusion matrices are computed exclusively on the held-out test set. For fair and consistent comparison, all models use the identical train/test split.

3.5. Evaluation Measures

In phishing detection and other text classification tasks, conventional classification measures such as accuracy, precision, recall, and F1-score are often used to evaluate machine learning model performance. These measurements originate from the confusion matrix, which describes classification outcomes in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [28].

Accuracy is defined as the proportion of correctly categorized samples (both phishing and legitimate) compared to the total number of samples [28].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision (also known as Positive Predictive Value) is the percentage of properly recognized phishing emails among all emails anticipated as phishing [29].

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

Recall (also known as Sensitivity or True Positive Rate) is a measure of the model's ability to properly identify genuine phishing emails [30].

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

F1-Score is the harmonic mean of accuracy and recall, resulting in a balanced assessment of the two [29].

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (8)$$

4. Methodology

Dataset Preparation: The experimental technique starts with collecting phishing and legal email corpora. The phishing subset was taken from the Nazario Phishing Corpus, a publicly available collection of around 9500–11,500 real-world phishing emails. Because Nazario only contains phishing samples, real emails were sourced from the Enron Corpus, a huge dataset of business email exchanges that is frequently utilized in text mining and classification studies. To preserve class balance, an equal number of authentic emails from Enron were collected and compared to the phishing emails from Nazario. To guarantee consistency and trustworthiness in the training data, all records were preprocessed, which included duplicate removal, encoding uniformity, and the removal of superfluous formatting.

Preprocessing and Tokenization: NLP models require text to be converted to numerical input. This was accomplished with the WordPiece tokenizer for BERT and Byte-Pair Encoding (BPE) for RoBERTa. Both tokenizers break down email content into sub-word units, allowing the models to handle out-of-vocabulary phrases and phishing-specific lexical changes like obfuscated URLs or misspellings. Each email was trimmed or padded to a maximum length of 256 tokens, with attention masks used to discriminate between content and padding.

Model Selection: We used two cutting-edge transformer-based architectures: BERT is a deep bidirectional language representation model that was trained on large-scale corpora using masked language modeling and next-sentence prediction tasks. Specifically, we employ the BERT-Base (uncased) and RoBERTa-Base pretrained variants, each consisting of 12 transformer layers, 768 hidden units, and 12 attention heads. Its self-attention mechanism gathers contextual dependencies from both left and right directions at the same time, making it very good at detecting subtle indications in phishing content, such as frantic tones or deceitful purpose. RoBERTa enhances BERT by training on bigger datasets, omitting the next-sentence prediction job, and using dynamic masking. Thus, improving resilience against noisy or hostile data, such as phishing efforts that intentionally change phrasing and structure.

Both models were optimized for binary classification (phishing vs. legitimate emails), with a classification head made up of a fully linked layer and softmax activation. Table 1 shows the used hyperparameters for fine-tuning the models.

Table 1. Training hyperparameters used for fine-tuning BERT and RoBERTa models.

Parameter	Value
Pretrained models	BERT-Base (uncased), RoBERTa-Base
Optimizer	AdamW
Learning rate	2.00×10^{-5}
Batch size	16
Epochs	3
Max sequence length	256
Weight decay	0.01
Warmup steps	10% of total steps
Gradient clipping	1

Loss function	Cross-entropy
---------------	---------------

Hyperparameters were chosen according to the standard practices in the literature when fine-tuning transformer models for text classification tasks. A learning rate of 2×10^{-5} is often accepted in the literature as a good compromise between convergence speed and stability of tuning. It turned out that training for 3 epochs was good enough to reach convergence on the balanced dataset. More training resulted in minimal improvements and a greater risk of overfitting. The maximum sequence length and batch size were set to optimize the representation capacity and the constraints of the GPU memory.

Training Procedure: The dataset was separated into training and testing sets based on an 80/20 split. Each model was trained for three epochs with a batch size of 16, using the AdamW optimizer and weight decay regularization to avoid overfitting. A linear learning rate scheduler with warmup was used to stabilize training. Gradient clipping was used to prevent bursting gradients, and all trials were carried out with fixed random seeds to ensure repeatability.

Phishing email detection framework: From raw data to model assessment, the phishing email detection approach shown in Figure 1 creates a structured pipeline. To ensure that phishing and authentic emails are appropriately gathered, cleaned, balanced, and labeled, the process starts with robust data capture and preprocessing. To prepare textual inputs for transformer models, the dataset is first divided into training and testing divisions. Tokenization and attention mask creation are then used. For binary classification, BERT and RoBERTa are optimized, and training is stabilized by the use of gradient clipping and optimization schedules. Confusion matrices and performance measures like accuracy, precision, recall, and F1-score are calculated during evaluation to give a thorough picture of the model's efficacy. Lastly, an interactive inference module enables real-time detection on unseen emails, while comparison analysis and visualization demonstrate the superior model.

Visualization and Comparison: To improve interpretability, confusion matrices were produced for each model to show the distribution of true positives, true negatives, false positives, and false negatives. A side-by-side performance comparison was performed between BERT and RoBERTa, allowing for an examination of their comparative strengths in phishing detection.

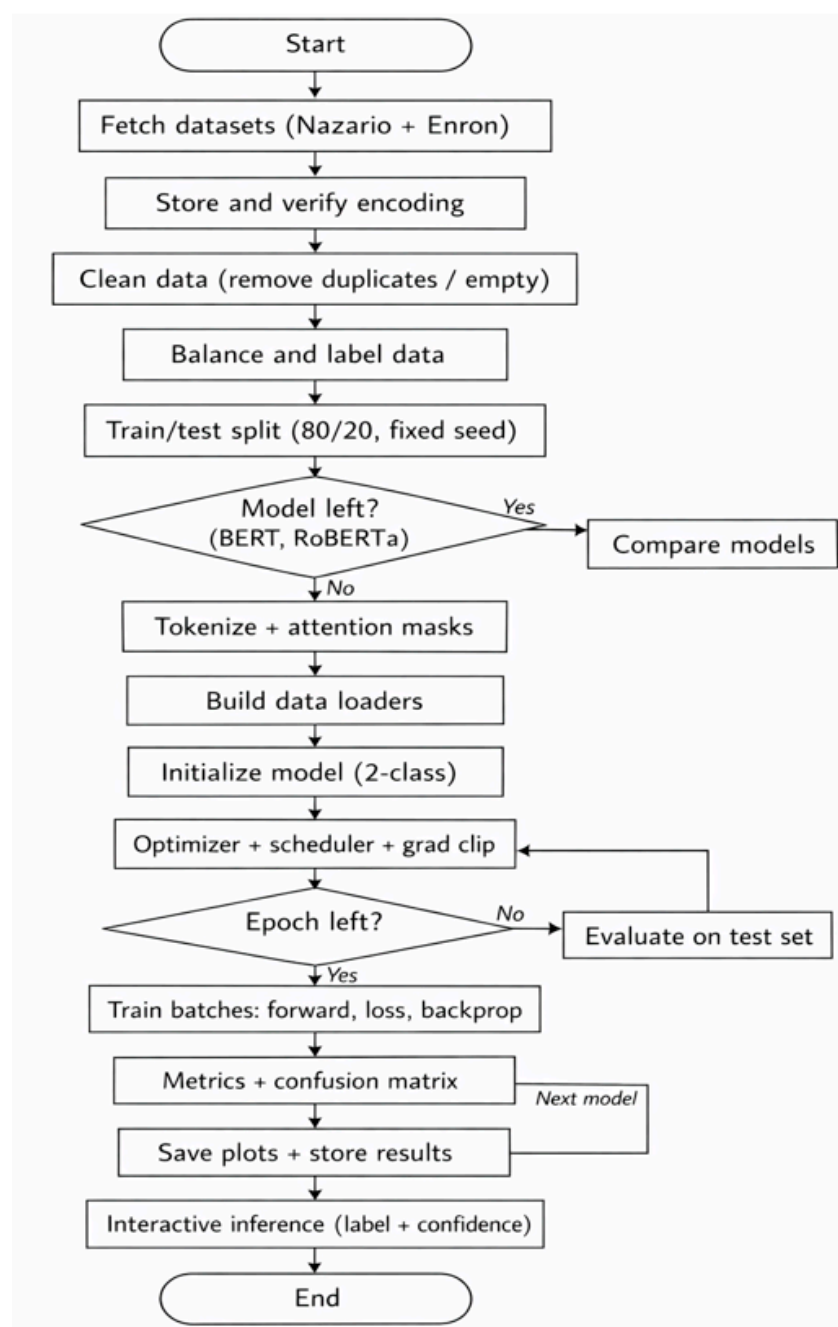


Figure 1. Flowchart of the proposed phishing email detection framework.

5. Results and Discussion

The confusion matrices in Figures 2 and 3 pertain only to the predictions made on the 20% test subset. The matrices show the class counts according to the specific distribution of the test split, not the fully balanced dataset.

To ensure fairness, both BERT and RoBERTa were evaluated using the identical test set, and confusion matrices were regenerated using this fixed split.

Figures 2 and 3 show the confusion matrices of BERT and RoBERTa, respectively. It can be seen that BERT correctly identified the majority of phishing and valid emails; however, it produced more false negatives (phishing emails misclassified as legitimate). This vulnerability is essential in phishing detection since undiscovered phishing emails can offer serious security risks. RoBERTa revealed a more robust balance, successfully lowering both false positives and false negatives. This enhancement is especially significant in

real-world security applications, as eliminating false negatives is critical to averting intrusions.

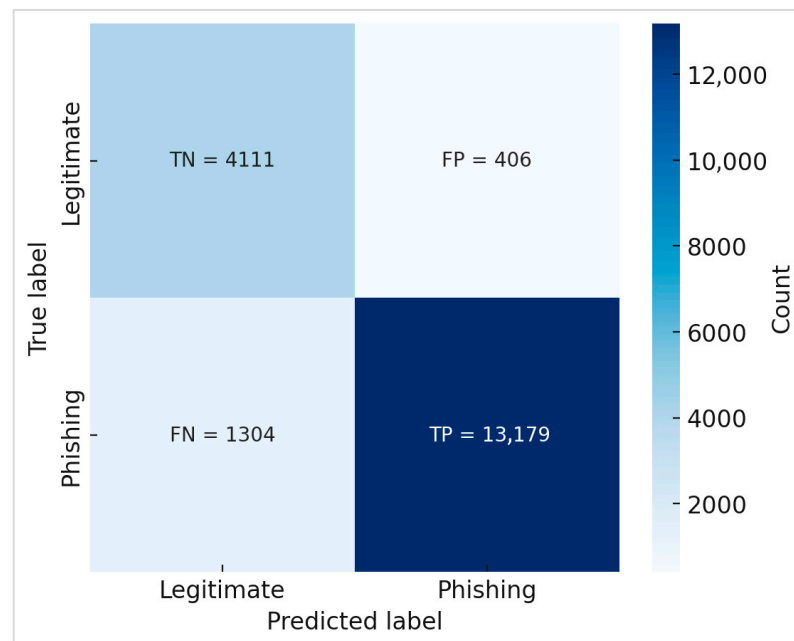


Figure 2. Confusion matrix of BERT model.

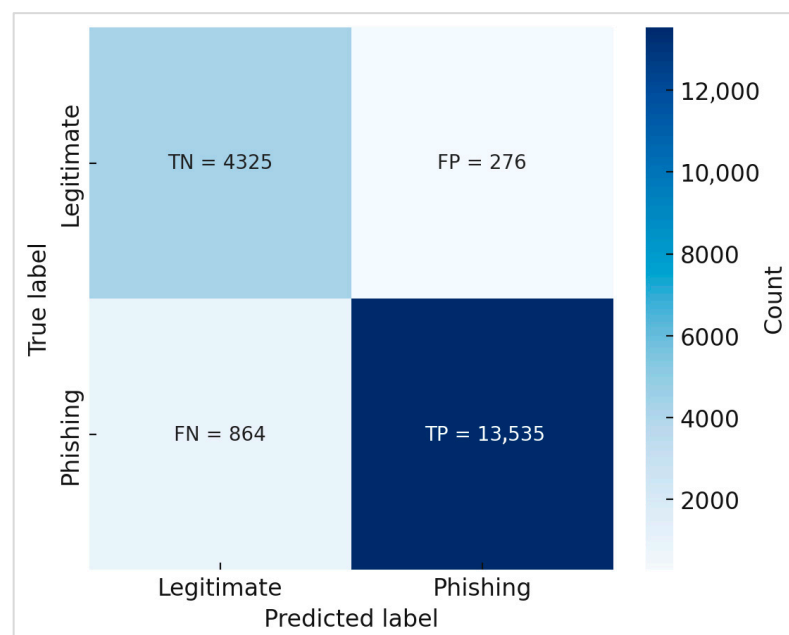


Figure 3. Confusion matrix of RoBERTa model.

The resulting performance, as shown in Figure 4, is as follows: for BERT, the accuracy = 0.91, precision = 0.97, recall = 0.91, and F1 = 0.94. Meanwhile, for RoBERTa, accuracy = 0.94, precision = 0.98, recall = 0.94, and F1 = 0.96.

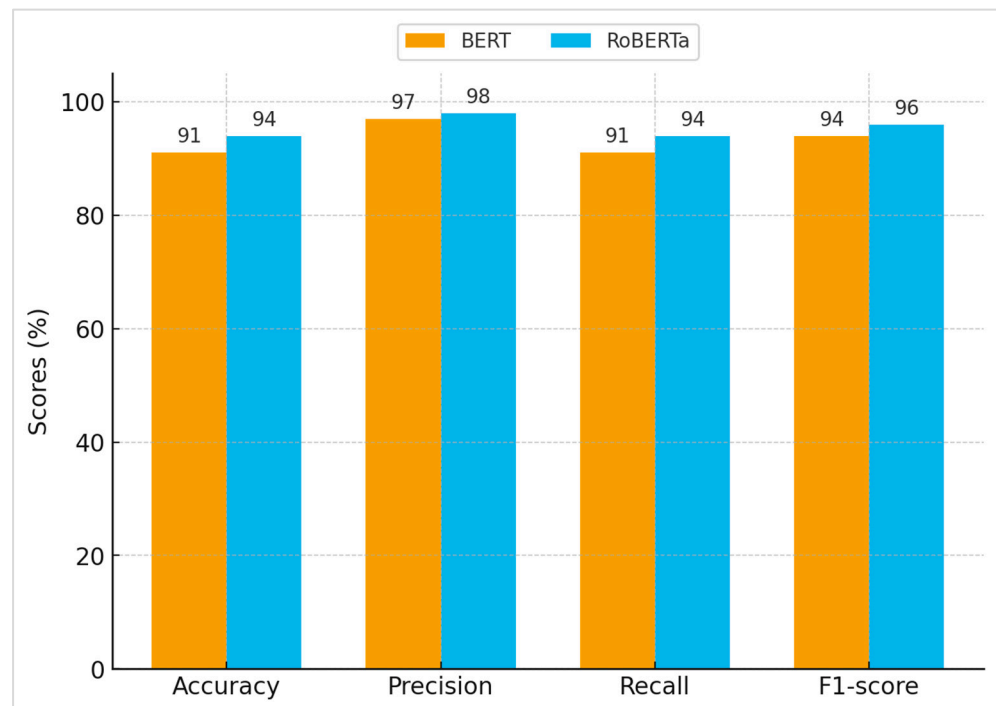


Figure 4. Performance metrics of BERT and RoBERTa models.

These findings suggest that RoBERTa's optimal pretraining tactics (longer training, dynamic masking, and the elimination of next sentence prediction) improve its ability to handle the misleading and combative character of phishing emails. Its greater recall is especially useful for reducing false negatives, ensuring that fewer phishing emails get through detection. This is critical in security-sensitive applications.

BERT showed high accuracy, indicating that it seldom misidentified valid emails as phishing, making it useful in situations where avoiding false alarms is crucial. In comparison, RoBERTa had a higher F1-score and a greater recall, indicating a better capacity to detect phishing attempts without sacrificing precision. This demonstrates that RoBERTa performs better across the board: it not only minimizes false positives, making it suitable for environments where avoiding unnecessary alerts is important, but also achieves higher recall, making it especially valuable in mission-critical scenarios where missing a phishing attempt poses a major security risk.

To examine resilience to predictive phishing techniques, we employed tests with deliberate character, whitespace, and URL obfuscations in emails. When assessing the two models under obfuscation, we saw that RoBERTa has, and maintains, a higher recall than BERT, suggesting that RoBERTa is more resilient to phishing texts that have been manipulated. This analysis goes beyond conventional accuracy assessments and emphasizes how models respond to more plausible attacks.

We evaluated our models with those in the literature, and our models performed better overall. For example, [7] proposed CATBERT, a context-aware Tiny BERT model for identifying social engineering emails, which obtained an accuracy of around 90% on their dataset. While effective, these results fall short of the performance of our fine-tuned transformer models, where BERT and RoBERTa attained higher accuracy. This demonstrates the benefit of combining bigger, optimized transformer designs with improved pretraining techniques, which were more robust in detecting minor phishing cues and lowering misclassification rates.

Furthermore, in reference [31], the authors employed the same datasets as those used in this study to train their BERT module, and while they had greater accuracy than our

BERT module (93%), they had lower perception (92%) and lower F1 (93%). Furthermore, our RoBERTa module outperformed their BERT module across all measures.

Table 2 presents examples from both the phishing and legitimate email categories to demonstrate the challenges involved in detecting phishing emails. Some phishing emails are written with professionalism and contain no apparent signs of malicious intent, while other legitimate emails contain requests and links that might superficially be indicative of phishing. These instances are proof that contextual and semantic reasoning are needed.

Table 2. Representative examples of phishing and legitimate emails used to illustrate task difficulty.

Category	Example Email Text
Phishing	"Your account access has been temporarily limited. Please review the attached document to restore full service."
Phishing	"Unusual sign-in activity detected. Verify your identity using the secure link below."
Legitimate	"Please review the attached purchase order and confirm if everything is correct."
Legitimate	"IT has completed system maintenance. Details are available on the internal portal."

In order to gain an in-depth understanding of the reason for misclassification, we manually reviewed a sample of emails that had been misclassified by the two models in question. We identified three main error categories. First, the emails that are phishing attempts look like authentic emails from a company. They may use a professional tone and may not contain a rush or threat language, which makes it hard, even for human readers, to detect them. Second, some of the emails that are authentic contain links, attachments, and prospective actions that are phishing-like, thus presenting the potential for false positives. Third, small numbers of misclassifications can also be a result of the presence of unclear and conflicting labels in the datasets because the corpora that are considered to be outdated can have an even borderline, or some samples that have been labeled indiscriminately may contain a mixture of misclassifications. Thus, most of the cases of misclassifications occur due to this being the case rather than from the models being deficient to some extent.

6. Conclusions

This paper provided an NLP-powered phishing email detection system based on fine-tuned transformer models, notably BERT and RoBERTa, which was tested on a balanced dataset that included both the Nazario Phishing Corpus and real emails from the Enron dataset. Through rigorous preprocessing, model tuning, and comparison analysis, we showed that both models exhibit good detection performance, with RoBERTa surpassing BERT across key criteria. BERT demonstrated great accuracy (0.91); its greater false negative rate reveals a weakness in mission-critical applications where undiscovered phishing emails can pose significant hazards. In comparison, RoBERTa performed better overall, with an F1-score of 0.96, an accuracy of 0.94, and a recall of 0.94. These findings support RoBERTa's optimal pretraining method, which improves robustness against adversarial or disguised phishing attempts and increases resilience in real-world circumstances. Future work should prioritize scalability, linguistic support, and explainability to guarantee broad application across varied situations. Future studies will also extend the comparative analysis to additional transformer architectures, including lightweight models such as DistilBERT and alternative pretraining strategies such as ELECTRA and related variants, to examine accuracy–efficiency trade-offs. In addition, simpler deep learning baselines such as CNN and hybrid CNN–LSTM architectures will be explored to

provide broader benchmarking against non-transformer neural models. Integrating real-time detection pipelines, adversarial training, and hybrid multimodal characteristics (such as URLs, metadata, and HTML content) may further improve detection accuracy and robustness.

Author Contributions: Conceptualization, M.I. and R.E.; methodology, M.I. and R.E.; software, M.I. and R.E.; validation, M.I. and R.E.; formal analysis, M.I. and R.E.; investigation, M.I. and R.E.; resources, M.I. and R.E.; data curation, M.I. and R.E.; writing—original draft preparation, M.I. and R.E.; writing—review and editing, M.I. and R.E.; visualization, M.I. and R.E.; supervision, M.I.; project administration, M.I.; funding acquisition, M.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The processed dataset can be released upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Singh, A. From past to present: The evolution of data breach causes (2005–2025). *LatIA* **2025**, *3*, 333.
2. Liu, R.; Lin, Y.; Yu, S.Y.S.; Teoh, X.; Liang, Z.; Dong, J.S. PiMRef: Detecting and Explaining Ever-evolving Spear Phishing Emails with Knowledge Base Invariants. *arXiv* **2025**, arXiv:2507.15393.
3. Abdelhamid, N.; Ayesh, A.; Thabtah, F. Phishing detection based associative classification data mining. *Expert Syst. Appl.* **2014**, *41*, 5948–5959.
4. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
5. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
6. Salloum, S.; Gaber, T.; Vadera, S.; Shaalan, K. A systematic literature review on phishing email detection using natural language processing techniques. *IEEE Access* **2022**, *10*, 65703–65727.
7. Alhuzali, A.; Alloqmani, A.; Aljabri, M.; Alharbi, F. In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets. *Appl. Sci.* **2025**, *15*, 3396.
8. Li, H.; Yang, J.; Li, Y.; Li, K. Email phishing attack detection based on BERT transformer model. In *Proceedings of the International Conference on Optics, Electronics, and Communication Engineering (OECE 2024)*, Wuhan, China, 26–28 July 2024; SPIE: Cergy-Pontoise, France, Volume 13395, pp. 1040–1045.
9. Jamal, S.; Wimmer, H. An improved transformer-based model for detecting phishing, spam, and ham: A large language model approach. *arXiv* **2023**, arXiv:2311.04913.
10. Safran, M.; Musleh, A. PhishingGNN: Phishing Email Detection Using Graph Attention Networks and Transformer-Based Feature Extraction. *IEEE Access* **2025**, *13*, 131390–131399.
11. Blake, S.E. Phishsense-1B: A Technical Perspective on an AI-Powered Phishing Detection Model. *arXiv* **2025**, arXiv:2503.10944.
12. Aljofey, A.; Bello, S.A.; Lu, J.; Xu, C. BERT-PhishFinder: A Robust Model for Accurate Phishing URL Detection with Optimized DistilBERT. *IEEE Trans. Dependable Secur. Comput.* **2025**, *22*, 4315–4329.
13. Hussan, P.H.; Mangj, S.M. BERTPHIURL: A Teacher-Student Learning Approach Using DistilRoBERTa and RoBERTa for Detecting Phishing Cyber URLs. *J. Future Artif. Intell. Technol.* **2025**, *1*, 417–428.
14. Rao, R.S.; Kondaiah, C.; Pais, A.R.; Lee, B. A hybrid super learner ensemble for phishing detection on mobile devices. *Sci. Rep.* **2025**, *15*, 16839.
15. Asiri, S.; Xiao, Y.; Li, T. Phishtransformer: A novel approach to detect phishing attacks using url collection and transformer. *Electronics* **2023**, *13*, 30.
16. Lim, B.; Huerta, R.; Sotelo, A.; Quintela, A.; Kumar, P. EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability. *arXiv* **2025**, arXiv:2503.20796.
17. Abiramasundari, S.; Ramaswamy, V. Cacography based Ransomware Email Phishing Attack Prevention using Language Pack Tuned Transformer Language Model. *Sci. Rep.* **2025**, *15*, 21526.

18. Çolhak, F.; Ecevit, M.İ.; Uçar, B.E.; Creutzburg, R.; Dağ, H. February. Phishing website detection through multi-model analysis of html content. In *Proceedings of the International Conference on Theoretical and Applied Computing, Kanjirapally, India, 13–15 September 2023*; Springer Nature Singapore: Singapore; pp. 171–184.
19. Patra, C.; Giri, D.; Nandi, S.; Das, A.K.; Alenazi, M.J. Phishing email detection using vector similarity search leveraging transformer-based word embedding. *Comput. Electr. Eng.* **2025**, *124*, 110403.
20. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536.
21. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
22. Brown, P.F. Class-based n-gram models of natural language. *Comput. Linguist.* **1990**, *18*, 18.
23. Churbanov, A.; Winters-Hilt, S. Implementing EM and Viterbi algorithms for Hidden Markov Model in linear memory. *BMC Bioinform.* **2008**, *9*, 224.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
25. Phuong, M.; Hutter, M. Formal algorithms for transformers. *arXiv* **2022**, arXiv:2207.09238.
26. Nazario, J. Phishing Corpus. Original Release of the Nazario Phishing Dataset. Available online: <https://monkey.org/~jose/phishing/> (accessed on 2 December 2025).
27. Klimt, B.; Yang, Y. The Enron Corpus: A New Dataset for Email Classification Research. In *Proceedings of the European Conference on Machine Learning, Pisa, Italy, 20–24 September 2004*; Springer: Berlin/Heidelberg, Germany; pp. 217–226.
28. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
29. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006*; pp. 233–240.
30. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
31. Otieno, D.O.; Namin, A.S.; Jones, K.S. The application of the bert transformer model for phishing email classification. In *Proceedings of the 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torino, Italy, 26–30 June 2023*; IEEE: Piscataway, NJ, USA, pp. 1303–1310.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.