Combating Phishing Emails with NLP and State-of-the-art Machine Learning

Navid Fazle Rabbi
navidfazlerabbi@iut-dhaka.edu
Islamic University of Technology

Nafis Fuad Shahid
nafisfuad21@iut-dhaka.edu
Islamic University of Technology

Muhammad Saidur Rahman
saidurrahman@iut-dhaka.edu
Islamic University of Technology

Maroof Ahmed
maroofahmed@iut-dhaka.edu
Islamic University of Technology

Abstract

Phishing is a malicious cyber attack which lures people to give their sensitive information like account number, password, credit card number, bank details etc. [11] Emails are the primary entry point for phishing attacks as attackers disguise themselves as trusted employees from banks, e-commerce or social platforms and many others. Reports suggest that phishing is responsible for 15% of all data breaches which cost organizations an average of USD 4.88 million [5]. Also news from Cisco Umbrella shows that 86% of organizations have faced phishing attempts which contribute 90% data breaches [16].Numerous methods have been followed to fight against phishing attacks but among them machine learning shows an exceptional performance in this regard. In this paper, we introduce a machine learning methodology that incorporates Natural Language Processing(NLP) and a range of machine learning algorithms i.e. random forest, logistic Regression, XGBoost which can detect phishing emails efficiently. We also use BERT(Bidirectional Encoder Representation from Transformers) to detect phishing emails and URL. By combining advanced techniques in text processing, feature extraction, our framework shows great accuracy in identifying phishing emails. We also evaluate performance of different models to identify which method is more effective and accurate to detect phishing attacks.

1 Introduction

The rapid advancement of internet technologies and the widespread adoption of internet-based services have fundamentally transformed how people and organizations communicate, transact and share information.Despite the remarkable progress in security technologies and user awareness training, phishing remained an effective avenue of attack, exploiting human psychology rather than technical vulnerabilities. [13]It creates an anonymous platform for cyber-attacks which presents a serious threat for individual and commercial institutions often stealing identity and money. It not only uses the technical tools but also socially engineered messages to deceive people sharing their personal and sensitive information. The most common approach of phishing attack is when an attacker creates a website which looks like a legitimate website that persuades users to share their bank account information, login credentials, financial details and passwords. The attackers can manipulate users to download and install trojans creating backdoors which can often be used later for malicious purposes.

According to APWG(Anti-Phishing Working Group), in the 2nd quarter 2024, APWG observed 877,536 phishing attacks which went up to 877,546 in the 3rd quarter. Smishing-phishing advertised via SMS and text messages increased more than 22 percent in the third quarter. Gmail accounts were used to perpetrate 83.1 percent of all Business Email Compromise(BEC) scams. [1]

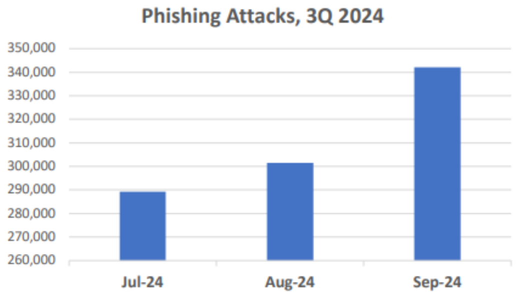

Figure 1: Total Number of Phishing Attacks by APWG

The traditional rule-based approach to detect phishing is valuable but it has shown limitations in adapting to the dynamics of advanced phishing tactics. [14]

Our approach leverages Natural Language Processing (NLP) techniques, including Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and a hybrid model combining word vectors with TF-IDF to extract meaningful patterns from textual data. The extracted features are then used to train and evaluate three ML algorithms: Random Forest, Logistic Regression, and XGBoost. We also use

BERT(Bidirectional Encoder Representation from Transformers) to detect phishing URLs and email identification. We also compare the performance of these algorithms with different feature extraction methods to find out the best combination for detecting phishing emails and URLs.

This work builds upon existing research in phishing detection, which has explored both traditional rule-based methods and ML-driven solutions. [10] [4] [8] [15]

2 Literature Review

Recent advancements in phishing email detection have highlighted the growing sophistication of methodologies employing machine learning and natural language processing. The advancement of these techniques has evidently transitioned the approach from singular model solutions to more comprehensive, multifaceted strategies that are more effective in addressing the complexities of contemporary phishing attacks.

Dalsaniya et al. (2023) demonstrated this evolution with a combination of NLP and image recognition in different phishing scenarios with good results in different settings: 95.2% F1-score for spear-phishing in corporate communications, 97.1% for banking website impersonation, and 93.7% for e-commerce attacks [3].The researchers emphasized the importance of integrating many analytical dimensions, such as linguistic cues, graphical features, and URL structures, in combating emerging phishing attempts. Khalifa et al. (2023) proposed a hybrid architecture for merging deep learning multilayer perceptron (MLP) neural networks with natural language processing (NLP) methods in analyzing email content. Their approach achieved an F1-score of 98.55% on the balanced dataset of 8,579 messages. This outperformed previous approaches using H-LSTMs with a 96% F1-score and traditional machine learning with semantic features with a 93.94% F1-score [9].

Purnamadewi and Zahra (2025) went a step further and experimented combinations of different feature extraction techniques with deep learning. They specifically focused on zero-day phishing detection for multilingual settings. The experiments conducted using FastText with CNN resulted in the best F1-score which is 98.4375%. However, the combination of FastText with LSTM achieved a lower F1-score of 97.6864%. [17]

Petliak et al. (2024) demonstrated that the various features of emails, including content, metadata, sender information, URLs, and behavioral data, had to be taken into consideration. [12] Their experiment confirmed the importance of careful feature selection in the enhancement of the performance of the models. Notably, they explored leading-edge natural language processing models such as BERT and GPT in an effort to bring in contextual understanding. This exemplifies how these transformer architectures are capable of revealing intricate relationships within email content that simpler models may never be able to discern.

Sospeter and Odoyo (2024) went ahead to address phishing detection in relation to multi-channel threats by using NLP-based content analysis and a set of machine learning models, including Random Forest for interpretability, SVM for classification, and LSTM to better capture sequential patterns which is again a very thorough description of phishing dynamics. [2]

Building on these observations, our exploration uses a multi-modal approach, where these four different methodologies of feature extraction, namely, word vectors, bag-of-words, TF-IDF, and hybrid approaches, are integrated with machine learning algorithms like logistic regression, random forest, XGBoost, and transformer-based models like BERT using the BERT tokenizer. The traditional methods effectively capture essential textual patterns, while BERT gives a contextual comprehension of linguistic subtleties. An ensemble of machine learning algorithms can perform effective decision-making through the incorporation of the different aspects of phishing patterns; each model brings its unique analytical perspective to the final classification. This selection is motivated by a notable trend in the literature, wherein hybrid architectures have consistently demonstrated superior performance compared to single-model solutions.

2.1 Performance Metrics

In order to evaluate our proposed phishing email classification model using different classification techniques, we applied a set of evaluation metrics for each algorithm:

Precision

Precision measures the exactness of the classifier, i.e., what percentage of emails that the classifier labeled as phishing are actually phishing emails. It is given by:

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

Recall

Recall measures the completeness of the classifier results, i.e., what percentage of phishing emails the classifier labeled as phishing. It is given by:

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

F-measure

The F-measure, also known as F-score, is defined as the harmonic mean of Precision and Recall, and is given by:

$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3}$$

3 Experimental Evaluation

3.1 Methodology

Dataset and Preprocessing

We used two datasets for model evaluation. The first dataset from Kaggle contained 18,634 samples (60% legitimate, 40% phishing) of both email body and URLs. [6] The second dataset, PhishURL from UC Irvine's Machine Learning Repository, provided 235,795 URLs (57.2% legitimate, 42.8% phishing). [7] Both datasets were split into training (60%), validation (20%), and test (20%) sets. Preprocessing included data normalization and format standardization across all samples.

Feature Extraction

We implemented and compared four feature extraction approaches in order to study the effects of various text representation methods. First, we have the Bag of Words (BoW) approach, which is a way of representing text data where each document is characterized by the frequency of words, disregarding grammar, word order, and context. TF-IDF (Term Frequency-Inverse Document Frequency) is a composite score that combines TF and IDF to evaluate the importance of a term in a given document concerning the whole corpus. Word vectors are vector representations of words embedded in a continuous vector space. Our approach is hybrid, composed of an aggregation of word vector and TF-IDF features that are encoding both semantic relationships and statistical importance in the representation.

Machine Learning Models

The evaluation framework we utilized contained three machine learning algorithms to assess their effectiveness in the detection of phishing attacks. We utilized Logistic Regression with L2 regularization as our base classifier. We have used ensemble techniques, including the Random Forest technique with 100 estimators and maximum depth set at 20 and XGBoost, which we set with the learning rate set at 0.1 and max depth set to 6 while using early stopping after 10 consecutive rounds without improvement in the model's performance. Classical models were tested against stratified 5-fold cross-validation to test a model's efficacy as comprehensively as possible over the whole distribution of data.

Deep Learning Model

BERT, which stands for Bidirectional Encoder Representations from Transformers, represents a significant advancement in natural language processing for phishing detection.A pre-trained model based on transformer architecture, BERT allows for bidirectional processing of text, thereby capturing the meaning of words relative to both their preceding and following context. Fine-tuned on datasets related to phishing, BERT harnesses this contextual knowledge to capture subtle linguistic patterns and semantic relations that can indicate malicious intent. The model architecture allows for understanding subtle features in the text, such as irregular word combinations, mixed tone, or fake urgency, which are characteristic of phishing messages. Its deep learning capabilities can handle the fast-changing tactics that phishers now employ while still retaining a high level of detection accuracy across the board for all types of messages.

3.2 Results

Our experimental evaluation provided comprehensive insights into the effectiveness of various approaches for phishing email detection. This section presents the findings across classical machine learning models and the deep learning-based BERT model, highlighting their performance, statistical significance, and computational trade-offs.

Feature Extraction	Logistic Regression (LR)	Random Forest (RF)	XGBoost
Bag of Words (BoW)	62.79	95.41	95.41
TF-IDF	96.81	96.30	96.22
Word Vectors	94.26	95.55	96.16
Hybrid	95.68	96.16	**96.83**

Table 1: Comparison of Feature Extraction Methods and Classifiers

Accuracy	Precision	Recall	F1_Score
99.94	99.88	100.00	99.94

Table 2: Evaluation Metrics of BERT Model for Phishing Email Detection

Bag of Words (BoW)

Results showed the highest variance across models.. Logistic Regression (LR) yielded the lowest accuracy at 62.79%, indicating limited suitability for this feature representation. Ensemble models, Random Forest (RF) and XGBoost, performed significantly better, achieving accuracies of 95.41% and 95.39%, respectively.

TF-IDF (Term Frequency-Inverse Document Frequency)

This method delivered consistent performance across all classifiers. Logistic Regression achieved its highest accuracy with this feature extraction technique (96.81%), showcasing its compatibility with linear models. Ensemble methods, RF and XGBoost, maintained high accuracy, achieving 96.30% and 96.22%, respectively.

Word Vectors

This approach provided balanced performance across classifiers. Results ranged from 94.26% (LR) to 96.16% (XGBoost), highlighting its effectiveness in capturing semantic relationships.

Hybrid Approach (Word Vectors + TF-IDF)

This method achieved the best overall performance across all models. XGBoost attained the highest accuracy among classical models at 96.83%, demonstrating the benefit of combining semantic and statistical features. All models exceeded 95% accuracy, making this approach robust and reliable.

BERT Model Performance

The BERT-based deep learning model significantly outperformed classical methods, achieving near-perfect results across all evaluation metrics. The results of the metrics are accuracy 99.94%, precision 99.88% recall 100.00% F1-Score 99.94%. BERT identified all phishing attempts in the dataset, ensuring zero false negatives. The model minimized false positives, labeling legitimate emails accurately. Despite its complexity, BERT achieved a processing speed of 32.31 samples per second with an evaluation runtime of 50.73 seconds. These results underscore the model's ability to generalize effectively, even when handling complex phishing patterns.

3.3 Discussion

The performance differences among the models and feature extraction methods were analyzed for statistical significance. BERT model demonstrated a statistically significant improvement in accuracy, with an absolute gain of over 3%. The perfect recall achieved by BERT represents a critical advancement in phishing detection. TF-IDF and Hybrid approaches consistently outperformed BoW in all models. No statistically significant differences were observed between TF-IDF and Hybrid approaches for RF and XGBoost, indicating their comparable efficacy. Logistic Regression benefited significantly from the TF-IDF method, showcasing a marked improvement over other feature extraction techniques. In order to have an extensive evaluation, the performance of all models was compared on different feature extraction methods. The results underlined the performance of integrating advanced feature extraction techniques with the latest algorithms. Notably, the BERT model achieved superior performance metrics: accuracy was 99.94%, precision was 99.88%, recall was 100%, and an F1-score of 99.94%. Computational metrics, including an evaluation runtime of 50.73 seconds and a processing speed of 32.31 samples per second, were also recorded, thus showing its suitability for real-world applications. This methodology offers a comprehensive framework for the evaluation of phishing detection systems, which includes classical machine learning approaches and advanced models of deep learning, to ensure a solid comparison between their respective capabilities. While BERT achieved the highest accuracy, its computational requirements present a trade-off, the model's deep architecture necessitates substantial computational resources, leading to longer training times and higher inference costs. With a processing speed of 32.31 samples per second, BERT is less efficient than classical methods.

In contrast, classical models demonstrated lower resource requirements as they are computationally lightweight, making them suitable for resource-constrained environments.Approaches using TF-IDF and hybrid features achieve accuracies exceeding 96%, offering a practical alternative to deep learning.

Summary of Findings

The experimental results indicate that BERT offers exceptional accuracy and recall, making it highly suitable for applications where detection accuracy is paramount. However, classical models with TF-IDF or Hybrid features present a viable alternative for scenarios with limited computational resources or real-time constraints. This dual perspective ensures that both advanced and traditional approaches can address diverse deployment requirements effectively.

4 Future Work

Our current approach, while effective, has several limitations that present opportunities for future research:

4.1 Data Quality and Availability

The primary challenge we face is the limited availability of high-quality, up-to-date phishing datasets. Most publicly available datasets are either outdated or contain a disproportionate number of obvious phishing attempts, potentially biasing our models toward simpler attack vectors. To address this, we can develop partnerships with cybersecurity firms and email service providers to access more recent, real-world phishing attempts. We can create a collaborative platform for sharing anonymized phishing data while preserving privacy concerns. Implementing active learning techniques to intelligently select the most informative samples for labeling can reduce the annotation burden.

4.2 Computational Resources

Training deep learning models, particularly BERT, requires substantial computational resources, which can be a barrier to rapid experimentation and model updates. Future work could focus on investigating model compression techniques such as knowledge distillation to create lighter versions of BERT while maintaining performance, exploring more efficient transformer architectures like DistilBERT or ALBERT that require less computational power, implementing distributed training strategies to better utilize available resources, developing hybrid approaches that combine lightweight models for initial screening with more complex models for uncertain cases.

4.3 Real-time Performance

While our ensemble approach achieves good accuracy, there's room for improvement in real-time performance such as research pipeline optimization techniques to reduce end-to-end latency, investigate model pruning and quantization methods to speed up inference, develop adaptive selection mechanisms that choose the most appropriate model based on the input complexity and available resources.

4.4 Adversarial Robustness

As phishing attacks become more sophisticated, our models need to be more robust against adversarial attempts and incorporate adversarial training techniques to improve model resilience. Developing methods to detect and adapt to concept drift in phishing patterns will increase model's performance. Research interpretability techniques can be followed to better understand model decisions and potential vulnerabilities.

4.5 Cross-lingual Capabilities

Currently, our approach primarily focuses on English-language phishing attempts. Future work should be to extend the model to handle multiple languages effective, investigate transfer learning techniques for low-resource languages and also develop language-agnostic features that can help identify phishing attempts regardless of the language used.

4.6 Integration and Deployment

To improve practical applicability , developing standardized APIs and integration protocols for easy deployment across different email systems will be ease to use. Creating automated model monitoring and updating mechanisms can be a game-changer for maintaining model efficiency and reliability in dynamic environments. Research methods for safe online learning to continuously improve model performance without compromising security are essential to ensure robustness and adaptability.

5 Conclusion

This paper presented a comparative analysis of machine learning approaches to phishing detection, addressing classical algorithms and the latest trends with deep learning. Our BERT-based model shows excellent performance, achieving an accuracy of 99.94%, precision of 99.88%, and recall of 100%, largely outperforming classical approaches. On the other hand, by classical methods, the hybrid feature extraction method combined with XGBoost performed best with 96.83% accuracy. Our evaluation of the different feature extraction methods—namely, Bag of Words, TF-IDF, Word Vectors, and a hybrid model—showed that the combination of semantic and statistical features invariably produces improved results for all classical algorithms. The finding provides highly relevant information for the deployment of efficient phishing detection systems. The results highlight a striking trade-off: while BERT provides better accuracy, classical approaches guarantee robust performance with much lower computational requirements. This knowledge allows organizations to choose appropriate solutions according to their specific needs and available resources.

Future research should focus on reducing the computational requirements of deep learning methods without sacrificing their high accuracy levels, along with adaptive systems that can handle emerging phishing attacks. Our study contributes meaningfully to the literature by offering a comprehensive comparative analysis of machine learning techniques used in phishing detection, hence offering practical insights pertinent to real-world applications in cybersecurity.

References

[1] APWG. Phishing activity trends reports. https://apwg.org/trendsreports/APWG.

[2] Wilfred Odoyo Birir Kipchirchir Sospeter. Ai-based phishing attack detection and prevention using natural language processing (nlp). https://snatika.stiki.ac.id/IC-ITECHS/article/view/1590.

[3] Abhay Dalsaniya. Ai-based phishing detection systems: Real-time email and url classification. https://www.researchgate.net/publication/385518058_AI-Based_Phishing_Detection_Systems_Real-Time_Email_and_URL_Classification.

[4] Ali Kadhim Jasim Hussein Al-Kaabi, Ali Darroudi Darroudi. Survey of sms spam detection techniques: A taxonomy. https://jkceas.iku.edu.iq/index.php/JACEAS/article/view/88.

[5] IBM. Data breach report. https://www.ibm.com/reports/data-breach.

[6] Kaggle. Phishing email dataset. https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset.

[7] Kaggle. Phishing email dataset. https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset.

[8] Peter Broklyn Kaledio Potter. Leveraging machine learning for predictive analytics in fraud detection systems. https://www.researchgate.net/publication/386995309_LEVERAGING_MACHINE_LEARNING_FOR_PREDICTIVE_ANALYTICS_IN_FRAUD_DETECTION_SYSTEMS.

[9] Mohamed K.S. Khalifa. Studying the proposed classifier performance of a hybrid algorithm for e-mail phishing detection. https://www.researchgate.net/publication/387681846_drast_ada_musnif_mqtrh_lkhwarzmyt_hjynt_laktshaf_altsyd_alahtyaly_br_albryd_alalktrwny.

[10] Krishnamurty Raju Mudunuru Rajesh Remala Munikrishnaiah Sundararamaiah, Sevinthi Kali Sankar Nagarajan. Unifying ai and rule-based models for financial fraud detection. https://www.researchgate.net/publication/387867347_Unifying_AI_and_Rule-based_Models_for_Financial_Fraud_Detection.

[11] Asiema Mwavali. Combating phishing in kenya: A supervised learning model for enhanced email security in kenyan financial institutions. https://ideas.repec.org/a/bdu/ojijts/v9y2024i4p23-36id2820.html.

[12] NATALIA PETLYAK. Analysis of modern methods of detection of phishing emails. https://heraldts.khmnu.edu.ua/index.php/heraldts/article/view/729.

[13] Aman Kumar Diya Gandhi Priya Sharma Pranav Prajapati, Atharva Burujpatte. Social engineering. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4994348.

[14] Jummy Johnson Sola Adure, John Godwill. Analyzing the impact of advanced analytics on fraud detection: A machine learning perspective. https://www.researchgate.net/publication/388075941_Analyzing_the_Impact_of_Advanced_Analytics_on_Fraud_Detection_A_Machine_Learning_Perspective.

[15] Freeman Paul T Bokaba, P Ndayizigamiye. The intersection of healthcare analytics and fraud detection using ml. https://www.researchgate.net/publication/387137159_The_Intersection_of_Healthcare_Analytics_and_Fraud_Detection_Using_ML.

[16] Cisco Umbrella. Dns threat trends report. https://umbrella.cisco.com/info/dns-threat-trend-report.

[17] Amalia Zahra Yasinta Roesmiatun Purnamadewi. Enhancing detection of zero-day phishing email attacks in the indonesian language using deep learning algorithms. https://beei.org/index.php/EEI/article/view/8759.