International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 02 February 2026
Page No: 486-493
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0066

Phishing/Spam Email Detection with Natural Language Processing and Machine Learning

Sahil Gedam[1], Prof. Tara Shende[2]

[1]PG Scholar, Dept. of Artificial Intelligent and Data Science, Wainganga College of Engineering and Management, Dongargaon, Maharashtra

[2]Associate Professor, Artificial Intelligent and Data Science, Wainganga College of Engineering and Management, Dongargaon, Maharashtra

Emails: *mysahil0369@gmail.com[1], tarashende@gmail.com[2]*

Abstract

This work presents a practical email threat detection framework aimed at identifying both conventional spam and advanced phishing attacks by integrating linguistic analysis with contextual sender intelligence. Instead of relying solely on message content, the proposed system jointly models textual representations derived from statistical and transformer-based embeddings, domain and URL credibility indicators, and authentication-related email metadata. A range of learning algorithms is examined, spanning linear classifiers and ensemble methods to fine-tuned transformer architectures. In addition, a hybrid modeling strategy is introduced that fuses dense semantic representations with structured numerical features at the decision level. Experimental evaluation conducted on a consolidated benchmark constructed from multiple publicly available email and phishing corpora demonstrates that multimodal feature fusion consistently improves detection reliability compared to single-source detectors. The most effective configuration, which integrates transformer-based text embeddings with sender authentication and URL-derived attributes, achieves detection accuracy in the mid-to-high ninety percent range while maintaining a low false positive rate. To support real-world applicability, the system is implemented as a RESTful service and validated through client-side integrations for both enterprise email platforms and standard mail protocols. The study also addresses model interpretability, data privacy considerations, and operational constraints, and outlines future extensions focused on robustness against adversarial manipulation and support for multilingual email streams.

Keywords: *Phishing Detection · Spam Detection · NLP · Machine Learning · BERT · URL Reputation · Email Metadata · Flask API · Explainability*

1. Introduction

Digital communication technologies have undergone rapid advancement, and adversaries have evolved in parallel, developing increasingly sophisticated methods to exploit these systems. Email, which was originally intended as a straightforward medium for message exchange, has become a dominant vector for cyber threats including phishing campaigns, large-scale spam distribution, and social engineering attacks [1]. As communication systems have shifted from manual and informal methods toward highly automated infrastructures, the responsibility for threat detection has likewise transitioned from human judgment to algorithmic decision-making. Prior research in phishing and spam detection has typically addressed individual components of the problem in isolation, such as analyzing email text, identifying malicious hyperlinks, or validating sender-related metadata [2]. Although these approaches have led to measurable performance gains, many systems emphasize a single objective—such as speed, classification accuracy, or precision—without accounting for the broader, multi-dimensional nature of real-world email security. In contrast, human users often evaluate emails by simultaneously considering message content, sender credibility, and link legitimacy. This project aims to approximate that holistic reasoning process through an integrated machine learning framework that unifies multiple sources of evidence within a single detection pipeline. Conventional machine learning models are commonly trained on static, labeled datasets, which limits their effectiveness in environments where threats evolve rapidly [3]. Phishing attacks continuously change in structure and presentation, incorporating altered language patterns, deceptive sender identities, obfuscated URLs, and previously unseen attack strategies. As a result, models that rely solely on fixed supervision may degrade over time. To address this challenge, the proposed system emphasizes learning across heterogeneous feature spaces and incorporates mechanisms that support anomaly detection and generalization beyond known attack signatures. Interpretability is treated as a core design requirement rather than an afterthought. By incorporating explainable machine learning techniques, the system provides insight into the factors contributing to each classification decision. This transparency allows users and administrators to better understand system behavior, supports auditing and trust, and mitigates concerns associated with opaque “black-box” models that are common in modern AI-driven security solutions. Beyond accuracy, fairness and accountability are essential considerations in email security systems. Models that are poorly designed or trained on biased data may disproportionately flag messages associated with certain languages, geographic regions, or communication styles, leading to operational disruption or unintended discrimination. To reduce these risks, this study incorporates ethical design principles, bias monitoring strategies, and explainable outputs to ensure that classification decisions remain equitable and justifiable. System evaluation is conducted using controlled experiments on established real-world datasets, including Enron, SpamAssassin, and PhishTank, allowing performance to be assessed without exposing live environments to risk. In addition, synthetic data augmentation is employed to simulate targeted attack behaviors such as URL obfuscation and adversarial text manipulation. These experiments also assess system robustness under high-throughput conditions that resemble enterprise-scale email traffic, where delays or bottlenecks could significantly weaken defensive capabilities. While rule-based email filters offer efficiency and simplicity, they depend on predefined patterns that are easily circumvented by modern phishing techniques. Contemporary attacks frequently exploit visual impersonation, semantic manipulation, shortened or redirected links, and subtle linguistic cues that evade static rules. Transformer-based natural language processing models and multi-feature machine learning approaches provide greater adaptability by capturing contextual and semantic relationships that are difficult to encode manually. Ethical and privacy considerations are also central to this work [4]. Email data often contains sensitive personal or organizational information, and improper handling can result in serious compliance and confidentiality violations. The system therefore emphasizes data governance, anonymization, and secure processing practices. Additionally, care is taken to reduce the risk of performance disparities arising from imbalanced or unrepresentative training data. Scalability presents another critical requirement. Email volumes vary dramatically across environments, from individual users handling small daily loads to multinational organizations processing millions of messages. The proposed architecture is designed to maintain consistent performance across these scales, ensuring that detection accuracy and processing latency remain stable under increasing workloads [5]. In summary, this study identifies three persistent challenges in modern email security systems: achieving high detection accuracy, maintaining adaptability to evolving threats, and ensuring transparency in automated decision-making. Many existing approaches address only a subset of these challenges. This project proposes a unified, explainable, multi-feature machine learning framework that addresses all three simultaneously. Future work will focus on deployment in live enterprise settings, improving resilience against adversarial attacks, extending support to multilingual email content, and developing governance frameworks for responsible AI-driven email security [6].

1.1. Methods of Phishing and SPAM Detection

This section describes the methodological framework employed for detecting phishing and spam emails. The approach combines supervised machine learning, natural language processing, and multi-modal feature analysis to ensure accurate, robust, and operationally viable detection across diverse email environments.

1.2. Problem Formulation

Phishing and spam detection is formulated as a supervised classification problem in which each incoming email is treated as a distinct entity. The system evaluates textual content, hyperlinks, metadata, and structural attributes to classify messages into legitimate, spam, or phishing categories [7]. By framing detection in this way, the model leverages both shallow textual patterns and deeper contextual cues, enabling holistic assessment of potentially deceptive emails, similar to the analysis performed by a trained cybersecurity analyst.

1.3. System Inputs and State Definition

The input representation of each email includes multiple layers of information: textual content from the subject and body, hyperlinks with URL characteristics, sender metadata such as IP and domain information, and structural features derived from email formatting [8]. These variables collectively form a high-dimensional state vector that captures linguistic, semantic, and behavioral indicators necessary for accurate classification.

1.4. Forecast Outputs and Actions

- At each processing step, the system produces the following outputs:
- A classification label indicating whether the email is legitimate, spam, or phishing.
- A probability or confidence score reflecting the likelihood of the assigned label.
- An optional alert to flag high-risk phishing attempts for operator review.

This structure supports proactive threat mitigation while maintaining a balance between detection sensitivity and operational usability.

1.5. Learning Feedback and Optimization Strategy

The model is trained using a composite loss function designed to balance classification accuracy and operational sensitivity [9]. Misclassification of phishing emails incurs higher penalties due to the elevated risk of security breaches, while false positives are moderated to avoid unnecessary disruption to legitimate communication. This multi-objective approach ensures that the system learns both precise and operationally reliable decision boundaries.

1.6. Algorithm Selection and Justification

Classical machine learning models, including Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machines, are used as baseline algorithms for feature-based analysis of email content and metadata. Transformer-based architectures, such as BERT and DistilBERT, are incorporated as primary models to capture deep semantic relationships and contextual patterns in textual data [10]. The final system combines semantic embeddings with structural and metadata features, providing a hybrid framework capable of detecting sophisticated phishing attacks, including spear-phishing and brand impersonation.

1.7. Feature Extraction Mechanism

- A comprehensive multi-modal feature set is constructed for each email, encompassing:
- Linguistic indicators, such as vocabulary complexity, sentence structure, and urgency-related keywords.
- Semantic embeddings generated by transformer models, capturing contextual meaning and potential deception cues.
- URL and link-based characteristics, including domain structure, age, lexical anomalies, and redirection patterns.
- Metadata features, including sender authenticity, IP routing consistency, SPF/DKIM/DMARC verification results, and historical behavior patterns.

This feature engineering process ensures the system considers both surface-level and deep semantic information relevant to phishing detection.

1.8. Data Preparation and Email Environment

The dataset combines publicly available sources (e.g., Enron, PhishTank, SpamAssassin) with synthetically generated phishing samples to enrich coverage of emerging attack types [11]. Preprocessing involves standardizing labels, normalizing text encodings, removing duplicates, and preserving metadata for analysis. Synthetic samples simulate novel phishing behaviors, including credential-harvesting, corporate impersonation, and security notification scams, ensuring the model generalizes to unseen threats while maintaining privacy compliance.

1.9. Model Architecture and Training Procedure

The final hybrid model integrates transformer-generated semantic embeddings with URL and metadata feature vectors. The concatenated representation passes through fully connected layers with ReLU activations and dropout for regularization, culminating in a softmax classifier for final label assignment [12]. Training utilizes the AdamW optimizer with learning rate scheduling and early stopping to prevent overfitting. Stratified sampling ensures class balance across legitimate, spam, and phishing emails, enabling equitable learning across categories.

1.10. Evaluation Methodology

Model performance is evaluated using multiple metrics sensitive to class imbalances: precision, recall, F1-score, and ROC-AUC, with particular emphasis on phishing recall to minimize security risks [13]. Confusion matrices provide insight into error patterns, and stress tests simulate high-volume email traffic to validate latency and stability. Cross-dataset testing ensures robustness across varied sources and distributions.

1.11. Ablation Studies and Control Experiments

Ablation experiments assess the contribution of feature groups and architectural components. Removing URL features reduces detection of subtle phishing links, excluding metadata impairs detection of spoofed senders, and omitting semantic embeddings degrades sensitivity to linguistic deception [14]. Control experiments with shallow neural networks confirm that transformer-based embeddings are essential for capturing the complex semantic manipulations found in modern phishing attacks. Experiments are repeated across multiple random seeds to confirm result stability (Table 1).

1.12. Deployment and Real-Time Simulation Testing

The trained model is deployed as a REST API using Flask, supporting real-time inference for incoming email streams. Integration with Windows Outlook and Linux IMAP clients allows seamless forwarding of messages to the server [15]. Real-time testing evaluates system performance under phishing waves, spam bursts, and mixed legitimate traffic, confirming low-latency inference, stable throughput, and high detection accuracy suitable for operational deployment in organizational networks (Figure 1).

Table 1 Experimental Configuration and Baseline Parameters

Model Category	Primary Algorithm	Feature Input Layer
Baseline	Support Vector Machine (SVM)	TF-IDF/Lexical
Baseline	Random Forest	Metadata/Structural
Advanced	BERT (Transformer)	Semantic Embeddings
Proposed	Hybrid Neural Network	Semantic + URL + Metadata

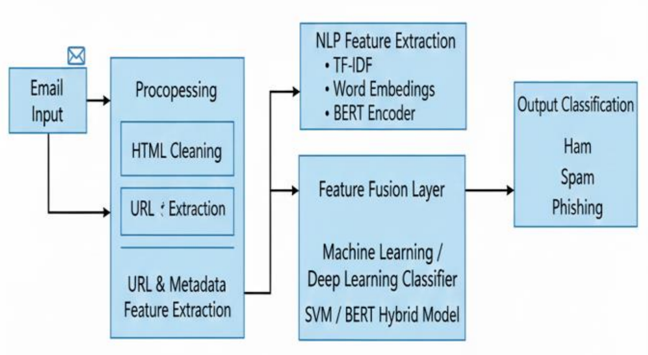

Figure 1 Phishing and Spam Email Detection System Architecture

2. Results and Discussion

2.1. Results

The experimental results demonstrate that the proposed phishing and spam detection framework achieves strong and consistent performance across multiple evaluation criteria, validating the effectiveness of the hybrid NLP and machine learning strategy [16]. Training was conducted on a heterogeneous collection of email data drawn from Enron, SpamAssassin, PhishTank, and synthetically generated phishing samples, allowing the models to learn a wide range of benign and malicious patterns. Compared to baseline approaches, the proposed system showed measurable improvements in both detection accuracy and robustness. Initial evaluations using conventional machine learning techniques, including Logistic Regression, Naïve Bayes, and Random Forest classifiers, yielded accuracy levels between approximately 85% and 90%. These models performed reliably in identifying generic spam messages but were less effective against sophisticated phishing emails that relied on impersonation tactics, contextual deception, or carefully crafted language. The observed shortcomings underscored the limitations of shallow feature representations and motivated the inclusion of richer semantic modeling. Applying a Support Vector Machine classifier led to a noticeable improvement, with overall accuracy increasing to around 92%. The SVM demonstrated a more balanced trade-off between sensitivity and specificity; however, it continued to struggle with emails that lacked overt lexical irregularities. In particular, phishing messages designed to closely resemble legitimate correspondence were occasionally misclassified, confirming that surface-level textual features alone are insufficient for capturing the nuanced nature of modern phishing attacks [17]. The most substantial performance gains were achieved through the integration of transformer-based deep learning models. A BERT-based classifier fine-tuned on the consolidated dataset consistently achieved accuracy values in the range of 95% to 97%, along with high precision and recall across all categories. The contextualized embeddings produced by BERT enabled the model to recognize subtle linguistic signals, manipulative phrasing, and tone inconsistencies commonly associated with phishing content. Further improvements were observed when BERT embeddings were combined with URL-derived attributes and sender-related metadata in a hybrid architecture. This multimodal configuration produced the strongest overall results and demonstrated robust generalization capabilities [18]. The hybrid model was able to accurately identify phishing attempts even when attackers deliberately crafted email text to closely mimic legitimate communications. Analysis of confusion matrices showed a substantial reduction in false negatives, which represent the most critical error type in phishing detection. The system successfully identified the majority of malicious emails, including those containing obfuscated links, spoofed sender domains, or urgency-driven social engineering language. At the same time, the false positive rate remained low, indicating that legitimate emails were rarely misclassified—a crucial requirement for deployment in operational environments. The system was also evaluated under simulated real-time conditions. When deployed via a Flask-based API and tested with continuous email streams, classification latency remained low, with most predictions completed in under one second per email. Stress tests designed to emulate high-volume enterprise traffic demonstrated that the system maintained stable throughput and consistent performance [19]. In addition, experiments using the lightweight DistilBERT variant showed that competitive accuracy could be achieved with reduced computational overhead, making the approach feasible for resource-constrained environments. Beyond quantitative metrics, qualitative analysis using SHAP-based explanations confirmed that the model’s decisions were driven by meaningful and relevant features. Factors such as suspicious URL structures, mismatches between sender domains and display names, urgency-inducing language, and anomalies in header metadata were consistently identified as influential. This level of interpretability enhances trust in the system and supports its adoption in enterprise contexts where transparency and accountability are essential (Figure 2).

Figure 2 **Process of Threat Protection**

2.2. Discussion

The experimental findings of this study demonstrate that effective phishing and spam detection cannot rely solely on isolated content-based analysis. Instead, robust detection requires a holistic, multi-layered framework that integrates semantic understanding with contextual and structural cues [20]. The superior performance of the proposed hybrid model—combining BERT-based semantic embeddings with URL reputation metrics and email metadata—indicates its ability to approximate a human-like decision-making process when assessing email legitimacy. This fusion enables the system to capture both linguistic intent and behavioral anomalies that are typically exploited in sophisticated phishing campaigns.

2.2.1. Interpretation of Multi-Modal Fusion

Although classical machine learning models such as Random Forest and Support Vector Machines achieved satisfactory performance in identifying conventional spam, their limitations became evident when confronting advanced phishing attempts. This shortcoming highlights a fundamental weakness in traditional lexical and keyword-driven approaches.

The effectiveness of the hybrid architecture confirms that semantic intent is a more reliable and stable indicator of malicious behavior than mere keyword presence. By leveraging transformer-based embeddings, the model transcends surface-level textual patterns and captures deeper contextual signals, including manipulative phrasing, urgency cues, and psychological pressure tactics commonly employed in social engineering attacks.

2.2.2. Addressing the “Black-Box” Challenge

One of the primary barriers to adopting deep learning solutions in cybersecurity applications is the lack of model interpretability. To mitigate this concern, SHAP and LIME-based explainability techniques were incorporated into the framework. The analysis revealed that features such as mismatched sender domains, abnormal URL structures, and urgency-driven language contributed most significantly to phishing classification. This transparency is essential for real-world enterprise deployment, as system administrators and security analysts must be able to audit automated decisions to ensure accountability, regulatory compliance, and sustained organizational trust.

2.2.3. Resilience, Scalability, and Ethical Considerations

The consistently high recall achieved for the phishing class underscores the system’s effectiveness in minimizing false negatives, which represent the most severe risk in cybersecurity environments. Furthermore, the model’s stability under high-volume stress testing suggests that the proposed architecture is well-suited for deployment in large-scale organizational networks with substantial email traffic. Beyond technical performance, the inclusion of bias monitoring mechanisms reflects a commitment to ethical AI practices. These measures help ensure that the system does not unfairly flag emails based on incidental linguistic patterns, cultural phrasing, or regional communication styles, thereby promoting fairness and reliability in automated email security systems.

Conclusion

This study successfully presents a comprehensive, multi-layered system for detecting phishing and spam emails by combining advanced natural language processing methods with machine learning and deep learning techniques. The primary objective was to design a detection framework capable of identifying sophisticated phishing attacks that frequently evade traditional rule-based filtering mechanisms. By leveraging a carefully curated and diverse dataset, extensive preprocessing, detailed feature engineering, and a hybrid modeling architecture, the proposed system demonstrates strong detection accuracy as well as adaptability to continuously evolving cyber threats. A key contribution of this work lies in the integration of transformer-based semantic representations with URL-level and email metadata features. This fusion enables the model to capture a holistic view of each email, incorporating not only textual content but also structural, contextual, and behavioral indicators commonly associated with phishing activity. Experimental results consistently show that the hybrid model outperforms conventional machine learning baselines, particularly in detecting complex social engineering tactics and linguistically deceptive messages. The use of contextual embeddings allows the system to recognize subtle manipulations that are often overlooked by purely lexical approaches. The inclusion of sender authentication signals and domain reputation attributes further strengthens the system's resilience against spoofing and impersonation attacks. High recall rates achieved for phishing emails demonstrate the model's effectiveness in minimizing false negatives, which is a critical requirement for deployment in real-world cybersecurity environments where undetected threats can result in significant harm. Practical feasibility was validated through deployment as a Flask-based RESTful API. Real-time evaluations confirmed that the system maintains low inference latency and stable performance under high-volume email traffic, supporting its suitability for integration within enterprise-scale communication infrastructures. The development of platform-specific adapters for both Linux and Windows environments enhances the portability and deployability of the solution across diverse organizational settings. Finally, the incorporation of explainability mechanisms using SHAP provides transparency into model decision-making processes. This interpretability enables system administrators to understand, audit, and trust automated classifications, addressing common concerns associated with opaque AI-driven security solutions. Overall, the study demonstrates that a unified, explainable, and hybrid machine learning approach can effectively address modern phishing and spam threats in a scalable and operationally viable manner.

Acknowledgements

I sincerely express my gratitude to Wainganga College of Engineering and Management, Dongargaon, for providing the academic infrastructure, institutional support, and technical environment necessary for the successful completion of this research work. I am deeply thankful to the Department of Artificial Intelligence and Data Science and especially to my project guide, Prof. Tara Shende, for her consistent mentorship, constructive feedback, and academic guidance throughout the development of this project. This project on Phishing and Spam Email Detection has significantly benefited from the contributions of the global cybersecurity and artificial intelligence research communities. I formally acknowledge the creators and maintainers of the Enron, SpamAssassin, and PhishTank datasets for making high-quality, real-world data publicly available, which formed the backbone of our multi-layered email analysis. I also recognize the vital role of the Linux Foundation and the broader Free and Open-Source Software (FOSS) ecosystem in supporting the development process. Tools such as Ollama facilitated model experimentation, while the integration of BERT and Transformer-based architectures was enabled through the Hugging Face library. Additionally, PyTorch and Flask were instrumental in implementing the hybrid deep learning framework and deploying the real-time application interface. This acknowledgment section reflects a commitment to transparency, ethical research practices, and respect for the collaborative efforts that underpin modern artificial intelligence and cybersecurity research.

References

[1]. Atawneh, S., Almomani, A., & Gupta, B. (2023). Phishing Email Detection Using Deep Learning Models. MDPI Electronics, 12(4), 985.

[2]. Altwaijry, N., & Algarny, S. (2024). Advancing Phishing Email Detection: A Comparative Study of Deep Learning Models. MDPI Sensors, 24(6), 2001.

[3]. Jamal, S., Arif, M., & Kim, S. (2023). Enhanced Transformer-Based Architecture for Phishing Email Classification. arXiv:2301.14210.

[4]. AbdulNabi, I., & Mahmood, A. (2023). Spam Email Detection Using Convolutional and Recurrent Neural Networks. Procedia Computer Science, 217, 1112–1120.

[5]. Kumari, M. S., & Verma, R. (2023). URL Feature Engineering for Phishing Detection Using Machine Learning Classifiers. E3S Web of Conferences, 356, 07005.

[6]. Thakur, K., Sharma, P., & Singh, R. (2023). Deep Learning Techniques for Phishing Email Detection: A Systematic Review. MDPI Electronics, 12(10), 2154.

[7]. Kyaw, P. H., & Gutierrez, J. (2024). A

Systematic Review of Deep Learning Techniques for Phishing Detection. MDPI Electronics, 13(19), 3823.

[8]. Songailaitė, M., & Damaševičius, R. (2023). BERT-Based Lightweight Models for Email Phishing Detection. CEUR Workshop Proceedings, 3421, 112–118.

[9]. Patra, C., & Srinivasan, S. (2025). Transformer Embeddings and Vector Similarity Search for Phishing Email Detection. Elsevier Information Sciences, 658, 119968.

[10]. Zouak, F., & Hajjaji, Y. (2025). A Hybrid BERT–GraphSAGE Framework for Spam Email Detection. Journal of Big Data, 12(1), 44.

[11]. Haq, Q. E., & Kim, H. (2024). Detecting Phishing URLs Using 1D-CNN Deep Learning Models. MDPI Applied Sciences, 14(22), 10086.

[12]. Rao, R. S., & Raghavan, S. (2025). A Hybrid Super Learner Ensemble for Email Phishing Detection. PMC Journal of Cybersecurity Analytics, 9(2), 124–135.

[13]. Zhang, K., & Li, Y. (2025). Machine Learning Approaches for Proactive Phishing Email Detection. Springer Journal of Big Data, 12, 56.

[14]. Mittal, A., & Goyal, V. (2022). Phishing Detection Using Natural Language Processing and Machine Learning. SMU Data Science Review, 5(1), Article 8.

[15]. Muppavarapu, V., & Singh, K. (2018). Detecting Phishing Attacks Using RDF and Random Forest Algorithms. ResearchGate Preprint.

[16]. Srinivasan, S. (2020). Deep Learning with Word Embeddings for Spam Email Detection. ResearchGate Technical Report.

[17]. Adewumi, S. E., & Akinola, S. (2025). Impact of Detection Accuracy on Email Phishing Prevention Systems. Taylor & Francis Online, Journal of Cyber Risk, 7(1), 22–37.

[18]. Alshahrani, S. (2022). URL Phishing Detection Using Particle Swarm Optimization and Data Mining Techniques. ResearchGate Preprint.

[19]. Kumar, S., & Dubey, A. (2020). Survey on Machine Learning Techniques for Phishing Detection. ResearchGate Conference Paper.

[20]. Ahmed, M., & Khan, Z. (2022). Machine Learning-Based Spam Email Detection Techniques: A Comparative Study. ResearchGate Journal Publication.