International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 5, Sep-Oct-2026, ISSN (Online): 2395-566X

AI Phishing Email Detector: A Review of Transformer-Based NLP, Sender and Url Analysis, and Defence Against AI-Generated Phishing

Kurhade Suyash Rajaram[1] Prof. R. V. Varhade[2] Prof. M. C. Kolse[3] Prof. S. K. Said[4]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India; Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India; Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India; Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India

Abstract—Phishing emails deceive individuals into revealing credentials and personal data, and attackers can now use large language models to write fluent, personalized phishing messages at scale that slip past rule-based filters. This paper reviews AI techniques for an adaptive phishing email detector that analyses message text, sender information and embedded URLs in real time. It traces detection from structural and feature-based approaches through natural-language analysis of intent and urgency to transformer models such as BERT and RoBERTa fine-tuned for phishing classification. URL analysis with deep character models, sender-authentication checks and behavioural signals are reviewed. Particular attention is given to the new threat of AI-generated phishing, including studies showing that language models can craft convincing spear-phishing emails and that the same models can help detect them. A five-layer architecture comprising email ingestion, multi-view feature extraction, transformer-based and URL classifiers, risk fusion with explanation, and user interaction with feedback is proposed. Research gaps, objectives, a comparison of models and challenges including adversarial adaptation, false positives, privacy and multilingual phishing are discussed. The review concludes that adaptive, explainable, multi-view detection is necessary as phishing becomes AI-assisted.

Index Terms—Phishing email detection, NLP, BERT, RoBERTa, URL analysis, AI-generated phishing, large language models, email security

I. INTRODUCTION

Phishing has always relied on persuasion: an email that creates urgency, imitates a trusted sender and asks the recipient to click a link or share a password. Traditional filters catch many such messages through blocklists, keywords and spelling mistakes. Large language models remove many of these tell-tale signs, producing well-written messages tailored to each target.

Hazell [1] demonstrated that large language models can generate realistic, personalized spear-phishing emails cheaply, and Heiding *et al.* [2] compared AI-generated and human-crafted phishing and also tested language models as detectors of malicious intent. Detection must therefore rely on deeper understanding of content, context and links rather than surface cues.

This paper reviews AI techniques for an adaptive phishing email detector and proposes a real-time architecture.

1. Theoretical Background

a. Signals In A Phishing Email

Fig. 1 organizes detection signals into content (intent, urgency, requests for credentials or payment), sender (authentication results, display-name spoofing, reputation), links (URL structure and destination) and context (unusual timing or first-time sender).

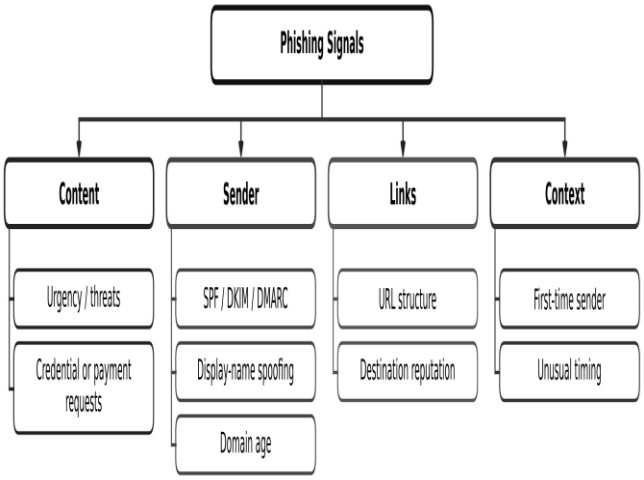


Fig. 1. Signals used by an AI phishing email detector.

2 Transformer Language Models

BERT [3] and RoBERTa [4] are pre-trained on large text corpora and can be fine-tuned with labelled phishing and legitimate emails to capture subtle cues of manipulation even when grammar is perfect.

© 2026 IJSRET

3 Survey Of Existing Approaches

Fette *et al.* [5] combined structural and link features in PILFER and achieved high accuracy on earlier phishing corpora. Verma *et al.* [6] applied natural language analysis to detect phishing intent, such as requests for action combined with urgency, rather than relying only on links.

Salloum *et al.* [7] systematically reviewed NLP techniques for phishing email detection and found a shift from bag-of-words features toward word embeddings and deep models, while noting that most datasets are old and English-only. Gutierrez *et al.* [8] studied phishing emails that evaded a university's filters and built a learning system to detect such new forms, showing the value of continuously updated models

For URLs, Le *et al.* [9] learned representations with character- and word-level CNNs, and Sahingoz *et al.* [10] showed that NLP features from URLs work well with Random Forest. In enterprises, Ho *et al.* [11] detected credential spear phishing using sender reputation and anomaly scoring with a small alert budget. Das *et al.* [12] systematized phishing research and recommended realistic, time-aware evaluation. Table I summarizes these works.

TABLE I Summary of Literature Survey

Ref.	Method	Key Findings	Limitations
[5]	PILFER structural + link features	High accuracy on classic phishing	Pre-LLM attacks
[6]	Natural-language intent analysis	Detects action + urgency cues	Rule-heavy
[7]	Review of NLP phishing detection	Shift to embeddings and DL	Old datasets
[8]	Learning from evaded phishing	Detects new forms	Single organization
[9]	URLNet char/word CNN	Automatic URL features	URL only
[11]	Enterprise spear-phishing detection	Low alert budget, high precision	Enterprise data needed
[1]	LLM-generated spear phishing	Showed cheap, convincing attacks	Attack demonstration
[2]	AI-generated phishing and LLM detection	LLMs can help detect intent	Small user study

II. COMPARATIVE STUDY

Table II compares models. Keyword and rule filters are fast but easily evaded, classical ML on TF-IDF is a strong baseline, transformers best capture manipulative language, and multi-view fusion is most robust.

TABLE II Comparison of Detection Models

Model	Handles AI-Written Text	Speed	Explainability
Rules / keywords	Poor	Very fast	High
TF-IDF + SVM / LR	Medium	Fast	Medium
BERT / RoBERTa	Good	Medium	Medium (attention)
Multi-view fusion	Very good	Medium	High (cues)

III. RESEARCH GAPS

- **G1:** Most training data predate AI-generated phishing and contain obvious errors absent from new attacks.
- **G2:** Detectors focus on English; Indian-language and code-mixed phishing is under-studied.
- **G3:** Explanations shown to users are often missing or vague.
- **G4:** Models degrade as attackers adapt, but continuous retraining is rarely evaluated.
- **G5:** Privacy constraints limit sharing of real email data for research.

IV. AIMS AND OBJECTIVES

- **O1:** To review content, sender and URL-based phishing detection methods.
- **O2:** To examine transformer models for phishing email classification.
- **O3:** To analyse the threat of AI-generated phishing and possible defences.
- **O4:** To propose an adaptive, explainable detection architecture.
- **O5:** To identify challenges and future directions.

© 2026 IJSRET

V. PROPOSED ARCHITECTURE

The proposed architecture is shown in Fig. 2

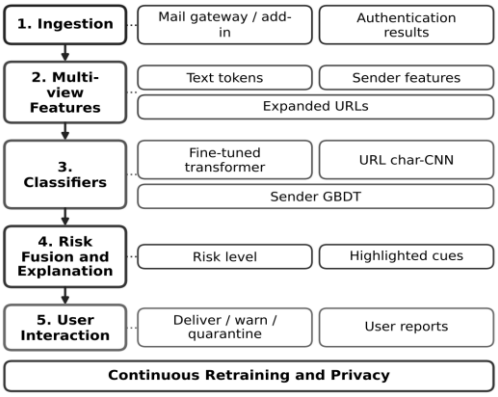


Fig. 2. Proposed adaptive AI phishing email detection architecture.

1. Ingestion Layer

Incoming emails are captured at the mail gateway or through an email-client add-in, and SPF, DKIM and DMARC [13] results are recorded.

B. Multi-view Feature Layer

The body is cleaned and tokenized; sender features include display-name similarity to known contacts and domain age; URLs are extracted and expanded.

2. Classifier Layer

A fine-tuned transformer classifies the text, a character-level URL model scores links and a gradient-boosted model scores sender and context features.

3. Risk Fusion And Explanation Layer

Scores are fused into a risk level; the explanation highlights suspicious phrases, mismatched links and sender anomalies.

4. User Interaction Layer

Low-risk emails are delivered, medium-risk emails receive a warning banner and high-risk emails are quarantined. User reports feed retraining.

5. Process Flow And Evaluation

Fig. 3 outlines the workflow. The transformer outputs a phishing probability from the representation of the classification token,

$$P(phish \mid email) = softmax(W \cdot h_{[CLS]} + b) \quad (1)$$

and the final risk combines component scores,

$$R = w_t \cdot S_{text} + w_u \cdot S_{url} + w_s \cdot S_{sender} \quad (2)$$

with thresholds chosen to keep false positives low while maximizing recall for phishing.

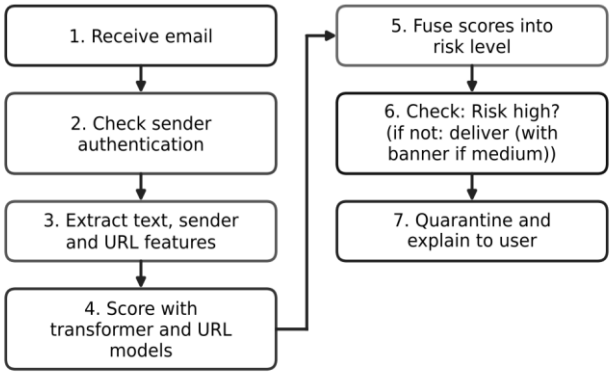


Fig. 3. Phishing email detection workflow.

Evaluation uses time-split data, so that the model is tested on emails newer than those used in training, and includes AI-generated phishing samples.

6. Limitations And Future Scope

Table III summarizes challenges. Future work includes adversarial training with AI-generated phishing, multilingual detectors, on-device privacy- preserving models and integration with phishing-resistant authentication.

TABLE III Challenges and Mitigation

Challenge	Mitigation
AI-generated phishing	Train on synthetic attacks; intent focus
Adversarial adaptation	Continuous retraining; ensembles
False positives	Warning banners; user feedback
Privacy of email	On-premise or on-device models
Multilingual phishing	Multilingual transformers

VI. CONCLUSION

Phishing is becoming AI-assisted, so detection must become AI-driven too. The reviewed literature shows that natural-language and transformer models capture manipulative intent, URL and sender analysis add independent evidence, and continuous learning is essential as attacks evolve. An adaptive, explainable, multi-view detector provides a practical defence that protects users' credentials and privacy.

1. Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance

© 2026 IJSRET

and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

REFERENCES

1. J. Hazell, "Spear phishing with large language models," arXiv:2305.06972, 2023.
2. F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing emails using large language models," IEEE Access, vol. 12, pp. 42131–42146, 2024.
3. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
4. Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.
5. I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in Proc. 16th Int. Conf. World Wide Web (WWW), 2007, pp. 649–656.
6. R. Verma, N. Shashidhar, and N. Hossain, "Detecting phishing emails the natural language way," in Proc. Eur. Symp. Res. Comput. Secur. (ESORICS), 2012, pp. 824–841.
7. S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using natural language processing techniques," IEEE Access, vol. 10, pp. 65703–65727, 2022.
8. C. N. Gutierrez et al., "Learning from the ones that got away: Detecting new forms of phishing attacks," IEEE Trans. Dependable Secure Comput., vol. 15, no. 6, pp. 988–1001, 2018.
9. H. Le, Q. Pham, D. Sahoo, and S. C. H. Hoi, "URLNet: Learning a URL representation with deep learning for malicious URL detection," arXiv:1802.03162, 2018.
10. O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Expert Syst. Appl., vol. 117, pp. 345–357, 2019.
11. G. Ho et al., "Detecting credential spearphishing in enterprise settings," in Proc. 26th USENIX Security Symp., 2017, pp. 469–485.
12. A. Das, S. Baki, A. El Aassal, R. Verma, and A. Dunbar, "SoK: A comprehensive reexamination of phishing research from the security perspective," IEEE Commun. Surveys Tuts., vol. 22, no. 1, pp. 671–708, 2020.
13. M. Kucherawy and E. Zwicky, "Domain-based Message Authentication, Reporting, and Conformance (DMARC)," IETF RFC 7489, 2015.

© 2026 IJSRET