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

Machine Learning for Email Security: A Review Of Spam, Phishing And Malicious Email Detection Techniques

Gandhi Himani Dipak[1] Prof. C. A. Kolse[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

Abstract—Email remains one of the most widely used channels of communication in personal, academic and business environments, and for the same reason it is a favoured route for spam, phishing, sender spoofing, malicious links and malware attachments. Rule-based filters and blocklists depend on known signatures and struggle against new and evolving campaigns. This paper reviews Machine Learning (ML) approaches that analyse email content, header information, sender behaviour and embedded links to classify messages as legitimate, spam or phishing. The review traces the development from Bayesian spam filters and support vector machines to feature-rich phishing classifiers, URL-based detectors and transformer language models, and it explains how sender-authentication standards such as SPF, DKIM and DMARC complement learned models. Text preprocessing, feature extraction, algorithm selection with Naive Bayes, SVM, Random Forest and Logistic Regression, and evaluation with precision, recall and false-positive rate are discussed. A layered architecture for an ML-based email security gateway is proposed, together with a research gap analysis, objectives and a comparison of algorithms. The review concludes that combining authentication checks, multi-view ML classification and user-facing warnings provides a smarter and more reliable defence for digital communication, provided that models are regularly updated against emerging threats.

Index Terms—Email security, spam filtering, phishing detection, machine learning, natural language processing, Naive Bayes, support vector machine, SPF DKIM DMARC

I. INTRODUCTION

Billions of emails are exchanged every day, carrying everything from course announcements to financial approvals. Attackers exploit this trust. Spam wastes time and bandwidth, phishing emails impersonate banks, universities or colleagues to steal credentials, spoofed messages trick employees into transferring money, and attachments deliver malware. Reports from the Anti-Phishing Working Group [1] show that phishing volumes remain persistently high.

Early defences used manually written rules and lists of known bad senders. These approaches break easily, because attackers change wording, rotate domains and craft personalized messages. Machine learning offers an adaptive alternative: by learning from labelled examples, a classifier can recognize the statistical patterns of malicious email even when the exact message has never been seen before.

This paper reviews ML techniques for email security. It covers spam filtering, phishing detection and the analysis of links and headers, proposes a layered architecture for an ML-based email gateway and discusses challenges such as false positives and evolving attacks.

II. RELATED WORK

Sahami *et al.* [2] applied a Naive Bayes classifier to junk email filtering and showed that adding domain-specific features such as phrases and sender domains improved accuracy. Androutsopoulos *et al.* [3] evaluated Naive Bayes anti-spam filtering on a public corpus and highlighted the importance of cost-sensitive evaluation, since blocking a legitimate email is more harmful than letting a spam through. Drucker *et al.* [4] showed that Support Vector Machines achieve high accuracy on spam categorization with less training time than boosting. Cormack [5] reviewed this body of work and recommended realistic, time-ordered evaluation.

For phishing, Fette *et al.* [6] proposed PILFER, which used features such as IP-based URLs, the age of linked domains, mismatched link text and HTML content, and achieved high accuracy with a Random Forest. Abu-Nimeh *et al.* [7] compared Logistic Regression, classification trees, Bayesian networks, Random Forests, SVMs and neural networks on phishing email and found that no single method dominated on all metrics.

Links are often the decisive evidence. Ma *et al.* [8] learned to detect malicious websites from lexical and host-based URL

© 2026 IJSRET

features, and URLNet [9] used character- and word-level convolutional networks on URL strings. In enterprise settings, Ho *et al.* [10] detected credential spear phishing by modelling sender reputation and user interaction, and later characterized lateral phishing sent from compromised internal accounts [11], which bypasses authentication checks entirely.

Transformer language models such as BERT [12] capture context and intent in message text, helping to detect urgency and impersonation cues. Das *et al.* [13] reviewed phishing research and warned that many studies use small, outdated datasets that overstate accuracy. Table I summarizes the reviewed literature.

TABLE I Summary of Literature Survey

Ref.	Authors	Year	Technique Used	Merits	Demerits
[2]	Sahami et al.	1998	Naive Bayes junk-mail filter	Showed probabilistic filtering is effective	Simple word features only
[4]	Drucker et al.	1999	SVM for spam categorization	High accuracy with fast training	Binary spam vs. ham only
[3]	Androutsopoulos et al.	2000	Evaluation of Naive Bayes anti-spam	Introduced cost-sensitive evaluation	Small corpus
[6]	Fette et al.	2007	PILFER phishing features + Random Forest	High phishing detection accuracy	Features may be evaded
[7]	Abu-Nimeh et al.	2007	Comparison of six ML methods	No single best classifier on all metrics	Dated dataset
[8]	Ma et al.	2009	Lexical and host-based URL features	Detects malicious sites beyond blacklists	Host lookups add latency
[9]	Le et al.	2018	URLNet character/word CNN	Learns URL patterns automatically	URL only; no email context
[11]	Ho et al.	2019	Detection of lateral phishing	Characterized attacks from compromised accounts	Enterprise-specific data
[13]	Das et al.	2020	Systematization of phishing research	Identified evaluation pitfalls	Survey; no new detector

III. PRELIMINARIES

3.1 Email Threat Categories

Fig. 1 classifies the main email threats. Spam is unsolicited bulk mail. Phishing attempts to obtain credentials or money through deception; spear phishing targets specific individuals with personalized content. Spoofing forges the sender address, and malware emails carry harmful attachments or download links.

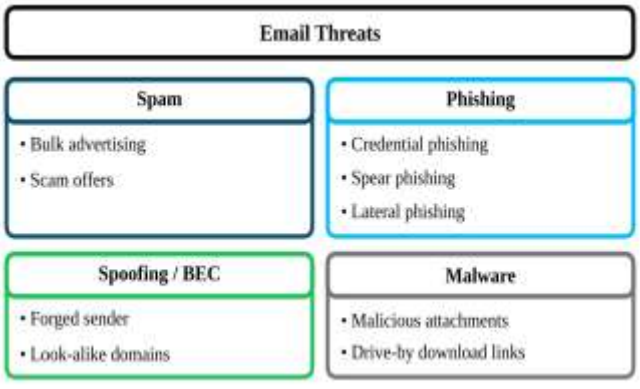


Fig. 1. Classification of email-borne threats.

3.2 Sender Authentication Standards

The Sender Policy Framework (SPF) [14] lets a domain list the servers allowed to send its mail. DomainKeys Identified Mail (DKIM) [15] attaches a cryptographic signature to messages, and DMARC [16] tells receivers how to handle mail that fails these checks. These standards reduce spoofing but do not stop phishing from look-alike domains, which is where ML is needed.

IV. PROBLEM STATEMENT AND OBJECTIVES

4.1 Identified Research Gaps

The following gaps are identified:

- Many reported results use old public corpora that do not reflect modern, personalized attacks.
- Models often analyse text, headers and URLs separately rather than combining them.
- Lateral phishing from compromised internal accounts is poorly handled by sender-reputation features.
- Explanations for why an email was flagged are rarely shown to users.

© 2026 IJSRET

- Regional-language and code-mixed phishing messages, common in India, are under-studied.

4.2 Objectives Of The Review

The objectives of this review are:

- To review ML and NLP techniques for spam and phishing email detection.
- To identify useful content, header and URL features for email classification.
- To propose a layered architecture for an ML-based email security gateway.
- To compare Naive Bayes, SVM, Random Forest, Logistic Regression and transformer models.
- To discuss challenges such as false positives, privacy and evolving attacks.

4.3 Proposed Framework

The proposed architecture in Fig. 2 analyses each incoming message from several viewpoints before delivery.

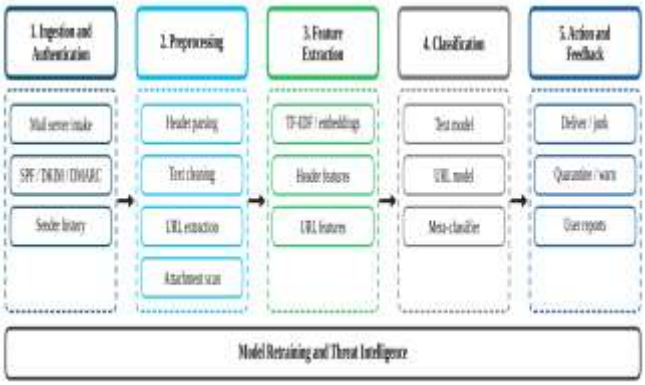


Fig. 2. Proposed architecture of an ML-based email security gateway.

4.4 Ingestion And Authentication Layer

Incoming messages are received from the mail server. SPF, DKIM and DMARC results are recorded, and the sender's history with the organization is looked up.

4.5 Preprocessing Layer

The message is parsed into headers, body text, HTML structure, links and attachments. Text is cleaned of markup, tokenized, lower-cased and stripped of stop words, and links are expanded and normalized.

C. Feature Extraction Layer

Text features include TF-IDF weights and transformer embeddings. Header features include authentication results, reply-to mismatches and routing anomalies. URL features include length, number of subdomains, presence of IP addresses, domain age and similarity to known brands.

4.6 Classification Layer

Separate models score text, URLs and attachments, and a meta-classifier combines their outputs into a final label: legitimate, spam or phishing.

4.7 Action And Feedback Layer

Legitimate mail is delivered; spam is moved to a junk folder; phishing is quarantined or delivered with a prominent warning banner. User reports of missed or wrongly blocked messages are used to retrain the models.

V. WORKING METHODOLOGY

The workflow is shown in Fig. 3. A Naive Bayes classifier estimates the probability that an email is malicious given its words $w_1,\ldots,w_n$ as

$$P(c \mid w) \propto P(c) \cdot \Pi_i P(w_i \mid c) \qquad (1)$$

and TF-IDF weights a term t in email d by

$$tfidf(t,d) = tf(t,d) \cdot log(N / df(t)) \qquad (2)$$

where N is the number of emails and df(t) the number containing t. Because blocking legitimate mail is costly, thresholds are tuned to keep the false-positive rate very low while maximizing phishing recall.

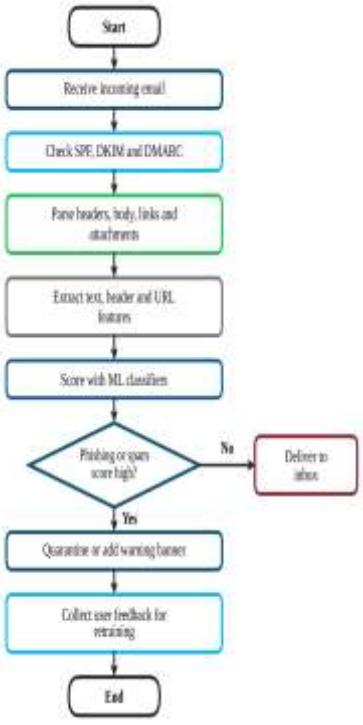


Fig. 3. Email classification workflow.

Models are retrained regularly on recent mail, and performance is tracked over time to detect drift as attackers change tactics.

© 2026 IJSRET

VI. DISCUSSION

6.1 Comparison Of Approaches

Table II compares common algorithms. Naive Bayes is fast and effective for spam; SVM and Logistic Regression perform well on high-dimensional text; Random Forest handles mixed header and URL features; and transformer models provide the best understanding of message context at higher computational cost.

TABLE II Comparison of Classification Algorithms

Algorithm	Accuracy	Speed	Notes
Naive Bayes	Good	Very fast	Strong spam baseline
Logistic Regression	Good–High	Fast	Interpretable weights
SVM	High	Medium	Effective on sparse text
Random Forest	High	Fast	Handles mixed features
BERT-based model	Very high	Slow (GPU)	Understands context and intent

6.2 Open Challenges And Future Work

Table III lists important challenges. Future directions include detecting AI-generated phishing text, analysing images and QR codes embedded in emails, using privacy-preserving training across organizations, and combining technical filters with user awareness training, since informed users remain the last line of defence.

TABLE III Challenges and Mitigation

Challenge	Mitigation
False positives	Low-FPR thresholds; warning banners instead of blocking
Evolving attacks	Frequent retraining; threat-intelligence feeds
Compromised accounts	Behavioural and lateral-phishing features
Privacy of email content	On-premise processing; minimal retention
Imbalanced data	Class weighting; recall-focused metrics

VII. CONCLUSION AND FUTURE WORK

Email will continue to be a primary target for attackers, and static rules alone cannot keep pace. The reviewed studies show that ML classifiers, from Naive Bayes to transformers, can detect spam and phishing with high accuracy when they combine content, header and link information. Sender authentication standards close some gaps, while ML handles look-alike domains and deceptive language. A layered gateway that authenticates, analyses from multiple views, explains its decisions and learns from user feedback offers a smarter and more reliable approach to protecting digital communication.

7.1 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 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. Anti-Phishing Working Group, “Phishing activity trends report,” APWG, quarterly reports. [Online]. Available: https://apwg.org/trendsreports/
2. M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, “A Bayesian approach to filtering junk e-mail,” in Proc. AAAI Workshop Learn. Text Categorization, 1998.
3. I. Androutsopoulos et al., “An evaluation of naive Bayesian anti-spam filtering,” in Proc. Workshop Mach. Learn. New Inf. Age, 2000, pp. 9–17.
4. H. Drucker, D. Wu, and V. N. Vapnik, “Support vector machines for spam categorization,” IEEE Trans. Neural Netw., vol. 10, no. 5, pp. 1048–1054, 1999.
5. G. V. Cormack, “Email spam filtering: A systematic review,” Found. Trends Inf. Retr., vol. 1, no. 4, pp. 335–455, 2008.
6. 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.
7. S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, “A comparison of machine learning techniques for phishing detection,” in Proc. Anti-Phishing Work. Groups eCrime Researchers Summit, 2007, pp. 60–69.
8. J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: Learning to detect malicious web sites from suspicious URLs,” in Proc. 15th ACM SIGKDD, 2009, pp. 1245–1254.
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. G. Ho et al., “Detecting credential spearphishing in enterprise settings,” in Proc. 26th USENIX Security Symp., 2017, pp. 469–485.
11. G. Ho et al., “Detecting and characterizing lateral phishing at scale,” in Proc. 28th USENIX Security Symp., 2019, pp. 1273–1290.

© 2026 IJSRET

12. 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.
13. 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.
14. S. Kitterman, "Sender Policy Framework (SPF) for authorizing use of domains in email, version 1," IETF RFC 7208, 2014.
15. D. Crocker, T. Hansen, and M. Kucherawy, "DomainKeys Identified Mail (DKIM) signatures," IETF RFC 6376, 2011.
16. M. Kucherawy and E. Zwicky, "Domain-based Message Authentication, Reporting, and Conformance (DMARC)," IETF RFC 7489, 2015.
17. Miryala, M., & Ramachandran, K. K. (2024). A framework for explainable adversarial resilience in deep neural networks against zero-day cyber threats leveraging causal inference and probabilistic model monitoring. International Journal of Artificial Intelligence & Machine Learning (IJAIML), 3(2), 257–270

© 2026 IJSRET