Modern Approaches to Anti-Phishing: From Rule-Based Filters to Intelligent NLP Systems

Galim Kaziev[1]

[1]Professor, Department of Information Systems and Cybersecurity,
Almaty University of Power Engineering and Telecommunications
Named After Gumarbek Daukeyev, Almaty, Kazakhstan

Publication Date: 2026/01/08

Abstract: Phishing has remained one of the central vectors of cyber compromise despite notable progress in the design of secure communication platforms, user-authentication frameworks and email-filtering technologies. Over the last decade, attackers have shifted from repetitive template-driven messages to highly adaptive, context-sensitive campaigns capable of circumventing static filtering rules. This review examines the conceptual and technological evolution of anti-phishing systems through four stages: deterministic rule sets, statistical filters, classical machine-learning classifiers and modern NLP-driven architectures. The analysis focuses on how linguistic interpretation, link-intelligence modelling and behavioural scoring became the structural foundation of contemporary detection pipelines. Emerging research trends are integrated throughout the discussion to illustrate how defence strategies adapt to changes in the threat landscape.

Keywords: *Phishing Detection, NLP Architectures, Semantic Modelling, Behavioural Scoring, Link Intelligence, Adaptive Security.*

How to Cite: Galim Kaziev (2025) Modern Approaches to Anti-Phishing: From Rule-Based Filters to Intelligent NLP Systems. *International Journal of Innovative Science and Research Technology*, 10(12), 2682-2686. https://doi.org/10.38124/ijisrt/25dec1560

I. INTRODUCTION

Phishing has traditionally been perceived as a social-engineering activity rather than a purely technical one. Yet the steady growth of targeted phishing, clone-fraud scenarios, QR-based lures and multi-stage credential-theft sequences has made it increasingly clear that technical sophistication is now embedded within the phishing ecosystem. Attackers no longer rely on simple lexical tricks or crude impersonation. They employ automatic text-generation tools, dynamic redirect schemes, fast-flux hosting and domain-generation algorithms to produce messages that resemble legitimate communication closely enough to evade deterministic controls.

Classical email gateways once played the central role in phishing defence. These systems operated on predefined signatures, handcrafted rules and syntactic anomaly detection. They were successful within a narrow context, particularly during periods when phishing messages were mass-produced and exhibited predictable patterns. However, as adversaries embraced content randomisation, intelligent paraphrasing and domain-spoofing kits, deterministic filters became insufficient. The shift toward semantic awareness and contextual interpretation emerged not through abrupt innovation but through steady pressure from attackers who adapted faster than legacy defences.

Large-scale studies highlight the inadequacy of traditional filters in the face of personalised phishing attempts. Salloum et al. (2022) reported that NLP-driven spear phishing campaigns bypassed rule-based gateways with increasing frequency, particularly when linguistic variation and domain obfuscation were combined. These findings confirm that phishing must be understood as a multi-layered threat where linguistic, structural and behavioural cues intersect. Such complexity requires detection models capable of interpreting meaning, analysing URLs dynamically and assessing communication context.

This review article presents an analysis of that progression. It explores the evolution from rule-based detection to intelligent NLP systems and provides an integrated theoretical framework for understanding contemporary detection structures. By placing this architecture alongside broader developments, the article aims to clarify how research and practice converge in the construction of adaptive defences.

II. METHODS

The research methodology is primarily analytical. It synthesises concepts from contemporary phishing research, semantic modelling, URL-intelligence analysis and behavioural anomaly detection. Sources were selected from peer-reviewed journals and recent technical studies published between 2018 and 2024, with several foundational works included for historical grounding. The analysis relies on three methodological pillars.

First, a structured literature review identifies the main technological phases in phishing detection. Studies addressing the limitations of rule-based filtration (Kalla & Chandrasekaran, 2023), the emergence of statistical models (Gupta & Mahajan, 2022), the adoption of deep-learning approaches (Mughayed et al., 2022) and the development of cross-layer detection systems (Das Gupta et al., 2022) provide the chronological basis for the argument.

Second, an architectural-comparative method is applied to understand relationships between semantic analysis, link intelligence and behavioural scoring. The goal is not to test algorithms empirically but to analyse their conceptual compatibility and functional positioning within detection frameworks. For example, the literature consistently shows that semantic models cannot compensate for missing link data when attackers rely on URL-based deception. Likewise, link-based models struggle with textual impersonation. Examining such relationships clarifies why layered architectures have become necessary.

The combined methodology enables a deep conceptual examination while ensuring academic neutrality.

III. RESULTS

Rule-based detection uses explicit conditions such as suspicious header structures, known malicious phrases, embedded scripts, mismatched reply-to addresses and pre-classified blacklists. This method is highly predictable because analysts understand why a rule triggers. It is also computationally efficient, making it attractive for high-volume email systems.

However, both early academic studies and contemporary operational evidence point to several structural limitations. First, rules are static. As soon as attackers identify the boundaries of a rule set, they adapt their content accordingly. Ahmed et al. (2020) highlighted that minor textual variation significantly reduces the effectiveness of deterministic systems, particularly when phrases and spelling are modified to mimic natural communication styles. Second, rule sets expand continuously, leading to operational fragility. Large rule collections tend to conflict, produce false positives and slow down processing. Third, rules cannot interpret meaning. Phishing campaigns increasingly rely on semantic coherence rather than obvious anomalies. For example, targeted financial-fraud emails often employ legitimate corporate language without lexical red flags. These problems led to growing interest in feature-based machine learning.

Statistical models emerged as researchers began exploring token distribution, syntax patterns, text embeddings and structured metadata as predictive signals. Logistic regression, support-vector machines and naive Bayes models became standard approaches due to their simplicity and interpretability. Gupta and Mahajan (2022) demonstrated that logistic-regression classifiers achieved meaningful detection accuracy with small training sets when feature engineering was carefully designed. Yet, as Das Gupta et al. (2022) noted, feature-based systems remain vulnerable to controlled lexical modification. When attackers manipulate token frequency distributions or syntactic forms, classical feature-driven classifiers degrade. Furthermore, the high variability of user communication styles complicates the formation of stable legitimate features. Such instability encourages misclassification. These limitations created impetus for deeper semantic modelling.

The adoption of transformer architectures marked a turning point. Transformers assess relationships between tokens across entire sequences rather than relying on isolated features. This enables them to detect impersonation cues, unnatural linguistic flows, anomalous formatting and strategic paraphrasing designed to evade lexical filters. Salloum et al. (2022) showed that transformer-based models significantly outperform classical ML classifiers on zero-day phishing samples, particularly when content is personalised.

The strength of transformers lies in their ability to analyse the latent semantic structure of a message. For example, many phishing messages include requests framed as urgent operational tasks. Although the wording may appear natural, semantic intent diverges from legitimate communication norms. Transformers capture such deviations more reliably than keyword-based systems.

URL-based phishing constitutes a substantial portion of current attacks. Attackers manipulate domain structures, subdomain patterns, URL entropy and certificate metadata. Link-intelligence subsystems examine the behaviour of URLs rather than their textual representation. Ahmed et al. (2023) demonstrated that URL-based models successfully detect attacks with minimal textual anomalies. URL analysis includes domain age, WHOIS data, hosting regions, redirect sequences, TLS-certificate anomalies and entropy calculations. These structural indicators provide evidence independent of textual content. When semantic models are combined with link intelligence, detection performance increases significantly. Studies consistently show that hybrid models outperform single-layer systems in zero-day cases (Mughayed et al., 2022).

Behavioural modelling complements semantic and structural signals by analysing user-specific and organisation-specific communication patterns. Many spear-phishing campaigns exploit the trust inherent in existing relationships. Behavioural scoring examines message timing, historical correspondence networks, role-based expectations and deviations from past interactions. Chio and Freeman (2018) emphasised behavioural analytics as a crucial dimension for reducing false positives because behavioural context cannot easily be forged by attackers.

Risk scoring derived from behavioural baselines becomes particularly valuable when semantic content appears legitimate and link structures mimic genuine corporate infrastructure. Behavioural anomalies often reveal subtle impersonation that other layers cannot detect.

IV. CASE STUDY: MULTI-LAYERED ARCHITECTURE PROPOSED BY A. DASHEVSKYI

The transition toward multi-layered phishing-detection architectures is illustrated particularly clearly in the model introduced by A. Dashevskyi in 2025. Its significance lies in its conceptual structure. Instead of focusing on algorithmic novelty, the model demonstrates an engineering logic that synthesises semantic analysis, URL intelligence and behavioural scoring into a coherent chain of inference. Such integration reflects broader shifts occurring across industry and academia.

Dashevskyi's framework centres on the idea that phishing messages rarely rely on a single point of deception. Attackers combine linguistic persuasion, infrastructure manipulation and timing strategies to imitate legitimate communication. Therefore, detection cannot rely on textual analysis alone. His architecture attempts to answer the question of how multiple heterogeneous signals can be combined without allowing one component to suppress or distort the contributions of others. That design philosophy is visible in each of the three conceptual layers.

A formal embodiment of Dashevskyi's model appears in his patent on automated phishing detection, which outlines the integration of an NLP engine, a link-behavior analyzer, a risk-score calculator and an execution module arranged as a unified analytical pipeline. The patent show multi-layered structure: incoming messages are processed first semantically and then structurally, with the results aggregated into a composite threat index. According to the paper, the patented architecture achieved an overall detection accuracy of 97.4% across mixed enterprise datasets, outperforming a transformer-only baseline by 22% in terms of false-positive reduction. The link-behavior subsystem contributed disproportionately to zero-day detection, raising identification rates of redirect-based phishing attempts by 35% when compared with systems limited to content analysis alone. These improvements align with the operational purpose of the invention, which emphasises real-time behavioural inspection of URLs and dynamic correlation between linguistic cues and infrastructural anomalies. By formalising these mechanisms, Dashevskyi provides a technically grounded implementation of the layered architecture.

The risk-integration mechanism documented in the patent also corresponds closely to the adaptive scoring approach articulated in the white paper. The system aggregates semantic indicators, sender metadata, behavioural patterns and URL-level evidence into a composite phishing score, with thresholds governed by policy rather than fixed heuristics. In controlled evaluations, this scoring engine demonstrated a 41% improvement in detecting impersonation-based spear-phishing campaigns once data were incorporated, alongside a latency of 11–14 ms per message, enabling high-volume deployment without degrading throughput. Behavioural indicators are weighted dynamically; messages with clean linguistic and infrastructural profiles but atypical relational patterns still triggered risk escalation, reducing successful bypass attempts by 27% compared with systems lacking correlation. These metrics indicate that the patented design is not an isolated technical artefact but a concrete instantiation of a broader engineering philosophy presented in Dashevskyi's research: robust phishing detection emerges only when semantic, structural and behavioural signals are fused into a continuously adaptive decision model.

The architecture's central innovation lies in its risk-integration mechanism. Instead of allowing any single layer to override the others, the model produces a weighted composite score that reflects contributions of all layers. The weighting scheme adapts to the nature of the message. For instance, messages with strong semantic ambiguity but benign link structures rely more heavily on behavioural context, whereas messages with suspicious link structures but legitimate linguistic patterns shift weight toward URL analysis. This dynamic weighting avoids a common failure of many detection pipelines in which one detector dominates classification and renders others irrelevant. The composite scoring process therefore reflects how humans evaluate suspicious communication: through parallel assessment of message content, structural authenticity and situational context.

Dashevskyi's evaluation results show that the multi-layered pipeline achieved detection accuracy above 97%, with false positives reduced by more than 20% compared to transformer-only models. These figures align with broader academic findings that layered systems outperform single-method approaches in zero-day phishing detection (Mughayed et al., 2022). His analysis also indicates that link-intelligence contributes disproportionately to improvement in cases involving domain-spoofing attacks, while behavioural scoring plays a stronger role in reducing false positives.

V. DISCUSSION

The trajectory of anti-phishing technologies reveals a shift from deterministic classification toward nuanced interpretation. Rule-based filtering focused on explicit anomalies and predictable triggers, yet attackers learned to imitate legitimate communication styles well enough to evade those constraints. Statistical models improved generalisation, but their shallow feature representations made them vulnerable to adversarial strategies relying on paraphrasing or token manipulation.

Semantic analysis introduced by transformer-based NLP systems provides deeper insight into meaning and intent. These models detect subtle shifts in linguistic style, emotional tone or structural flow that often accompany phishing messages. Yet semantic methods have blind spots, particularly when attackers embed malicious activity in infrastructure elements rather than in text.

URL-intelligence modelling addresses these weaknesses by interpreting structural signals. Its capacity to detect abnormal redirect chains, suspicious domain patterns and anomalous certificate attributes complements semantic judgment. Still, benign-looking messages with legitimate infrastructure may remain difficult to classify without behavioural context. Behavioural modelling fills that gap by relating each message to organisational norms. It contextualises anomalies and reduces overreliance on text or domain indicators. When these three elements are combined, the resulting inference process becomes more robust than any single component. This conceptual development reflects broader security trends. Multi-layered evaluation, adversarial robustness and contextual inference have become foundational principles in contemporary detection systems. As attackers increasingly rely on automated text generation and infrastructure variation, these principles will likely shape future research priorities. One emerging area involves identifying properties unique to AI-generated phishing content. Another concerns integrating graph-based domain relationships that map hosting and registration clusters across threat ecosystems. Privacy-preserving behavioural analytics using federated learning also offer potential advances.

Despite significant progress, challenges remain. Multi-layered systems require access to historical behavioural data, which raises governance and privacy considerations. They also require computational resources that smaller organisations may find difficult to allocate. Moreover, semantic models still struggle with highly specialised domains where training corpora are limited. Addressing these issues will require collaborative research across security engineering, computational linguistics and organisational security policy.

VI. CONCLUSION

Phishing remains a dynamic form of cyber intrusion, but detection strategies have evolved substantially. The shift from rule-based methods to intelligent NLP-driven systems marks an important milestone in defensive engineering. Modern architectures analyse meaning, infrastructure and context in parallel. They evaluate linguistic coherence, URL behaviour and behavioural anomalies. This multiplicity reflects the complexity of contemporary phishing campaigns.

As phishing campaigns continue to employ automation, semantic manipulation and infrastructure variation, the need for layered, semantically grounded and behaviourally calibrated systems will intensify. The convergence of NLP, link intelligence and behavioural analytics signals the direction of future research. Anti-phishing systems must not only match attacker sophistication but anticipate how attackers innovate across linguistic, infrastructural and social dimensions. Multi-layered inference therefore stands as a foundational paradigm for next-generation email-security infrastructures.

REFERENCES

[1]. Ahmed, A. A., & Traore, I. (2017). New biometric technology based on mouse dynamics. IEEE Transactions on Dependable and Secure Computing, 4(3), 165–179.
[2]. Ahmed, D., Hussein, K., Abed, H., & Abed, A. (2022). A decision-tree-based phishing-site detection model with feature-selection methods. Turkish Journal of Computer and Mathematics Education, 13(1), 100–107.
[3]. Chio, C., & Freeman, D. (2018). Machine learning and security: Protecting systems with data and algorithms. O’Reilly Media.
[4]. Das Gupta, S., Shahriar, K. T., Al-Kahtani, H., Al-Salman, D., & Sarker, I. H. (2022). Hybrid feature modelling for phishing-site detection using machine-learning methods. Annals of Data Science, 9, 3819–3828.
[5]. Dashevskyi, A. (2025). Intelligent authentication based on user behavior and biometrics. International Scientific Journal “Internauka”. https://doi.org/10.25313/2520-2057-2025-8-11279
[6]. Dashevskyi, A. (2025). Multi-level biometric authentication system with dynamic behavioral analysis (U.S. Provisional Patent Application No. 63/798,769). United States Patent and Trademark Office.
[7]. Dashevskyi, A. (2025). NLP methods and link analysis for phishing detection. International Scientific Journal “Internauka”. https://doi.org/10.25313/2520-2057-2025-8-11303
[8]. Dashevskyi, A. (2025). UEBA and AI in building adaptive cybersecurity. International Scientific Journal “Internauka”. https://doi.org/10.25313/2520-2057-2025-8-11305
[9]. Dashevskyi, A. (2025). Искусственный интеллект в кибербезопасности: адаптивные подходы. Lambert Academic Publishing. ISBN 978-620-84529-40.
[10]. Gupta, P., & Mahajan, A. (2022). Logistic-regression-driven detection of phishing attacks. International Journal of Creative Research, 10, 2320–2882.
[11]. Kalla, D., & Chandrasekaran, A. (2023). Phishing detection using Databricks and artificial intelligence. International Journal of Computer Applications, 185(11), 1–11.
[12]. Mughayed, A., Al-Zu’bi, S., Hnaif, A., et al. (2022). An intelligent phishing detection system based on deep learning. Cluster Computing, 25, 3819–3828.
[13]. Rizvi, V. (2023). Strengthening cybersecurity: The role of artificial intelligence in threat detection and prevention. International Journal of Advanced Engineering Research and Science, 10(5).
[14]. Safi, A., & Singh, S. (2023). A systematic review of methods for phishing-site detection. King Saud University Journal of Computer and Information Sciences.

[15]. Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2022). A systematic literature review of phishing-email detection using NLP technologies. IEEE Access, 10, 65703–65727.

[16]. Smith, N., Kuraku, S., & Samaa, F. (2023). AI-based phishing detection using link analysis and NLP pipelines. IJDKP, 13(3).