International Journal of Scientific and Management Research
Volume 8 Issue 11 (November) 2025
ISSN: 2581-6888
Page: 57-71

Phishing Detection Using Natural Language Processing and Behavioural Analysis: A Multi-Faceted Approach

Abuh Ibrahim Sani[1], Oludolamu Onimole[2], Adetunji Oludele Adebayo[3], Nathaniel Adeniyi Akande[4], & Uju Judith Eziokwu[5]

[1]Cybersecurity Analyst/Independent Researcher, EyBrids Limited, Nigeria
[2]Cybersecurity Operation Analyst/Independent Researcher, Teesside University, Middlesbrough
[3]Information Security Manager/Independent researcher, University of Bradford
[4]Cyberecurity Analyst/Independent Researcher, University of Bradford
[5]Data Analyst/Independent Researcher, University of Bradford, **UK**
DOI - http://doi.org/10.37502/IJSMR.2025.81105

Abstract

Phishing is still one of the most common and harmful types of cyberattack. It uses human psychology and trust to break into systems, steal credentials, and cost people money. Conventional defences, frequently dependent on static heuristics or domain reputation lists, find it challenging to adjust to the swiftly changing linguistic styles and technical infrastructures employed by attackers. This paper presents a hybrid phishing detection framework that combines Natural Language Processing (NLP) with Behavioural Analysis to enhance accuracy, interpretability, and resilience. Urgency, sentiment, and pragmatic intent are examples of linguistic indicators. Irregular sender activity and recipient interaction are examples of behavioural features. The model uses transformer-based architectures and ensemble learning for classification. It uses datasets from PhishTank and Enron, and Adaptive Synthetic Sampling (ADASYN) to fix class imbalance. Experimental evaluation shows that this system works very well, with 98.7% accuracy, 97.9% recall, and an AUC of 0.99. This is better than single-modality systems. Adding features that can be easily understood makes things clearer and provides analysts with useful information. The results show that a multimodal, privacy-conscious, and explainable framework greatly improves phishing detection, making it a useful and scalable improvement for modern email security systems.

Keywords: Behavioural Analysis, Cybersecurity, Hybrid Detection, Natural Language Processing, Phishing

1. Introduction

Electronic mail (email) is still the most common way for social engineering attacks to happen, and phishing remains the main way to steal credentials, lose money, and compromise organisations. Because of this, defenders have to deal with two problems: they need to differentiate malicious messages from legitimate ones without causing problems for legitimate communication (Zieni et al., 2023). Attackers use social engineering to trick people into doing things, like clicking on a link, that lead to the installation of malware or the theft of personal

information. Hackers often pretend to be representatives of reputable businesses and use dishonest methods to get customers' personal information, such as credit card details and passwords (Mittal et al., 2022).

Research and practice show that textual cues in messages are a useful way to find things, but static lists or simple heuristics aren't enough because attackers are always changing the way they phrase things and the infrastructure they use to send them to get around these checks (Salloum et al., 2022). Phishing messages are designed to persuade, and they often have hidden lexical, syntactic, and pragmatic markers, imperatives, words that show urgency, and requests that sound like transactions that can show malicious intent when looked at closely (Bountakas et al., 2021). Attackers also show behavioural patterns in the sender infrastructure and delivery, like bursts of messages, strange targeting of senders and recipients, and sending times that are out of the ordinary. These behavioural signals show campaign dynamics that content analysis alone can't see (Omar et al., 2023).

Each modality is noisy on its own, from legitimate transactional emails or marketing notices that sometimes show urgency or requests for action, and legitimate operational changes can temporarily alter sending behaviour. Therefore, a detection approach that combines content understanding with sender and interaction behaviour offers a route to higher accuracy and greater resilience to evasion.

Operational limitations drive additional design decisions. Security teams require detection outputs that are actionable and interpretable to facilitate straightforward triage and remediation, rather than obscurity (Zieni et al., 2023). Research on explainable feature detection demonstrates that clarity regarding the textual features and behavioural markers influencing a decision significantly aids human analysts (Zieni et al., 2023). Additionally, detection methods must be feasible within privacy and telemetry limitations; behavioural features that utilise aggregated or metadata-level signals are favoured over those necessitating intrusive per-user logging (Omar et al., 2023).

This research aims to develop and assess a comprehensive detection framework that combines linguistically based message analysis with streamlined behavioural profiling, prioritises interpretable signals for human analysts, and is evaluated for resilience to temporal variations in attacker strategies. The subsequent literature review contextualises this contribution in relation to preceding research on content-based detection, behavioural analysis, and hybrid systems that integrate both viewpoints.

2. Literature Review

Early practical defences against phishing relied on syntactic heuristics and reputation lists. These included blacklists for known malicious domains, simple checks for mismatches between display names and envelope addresses, and rules that flagged obvious lexical errors or attachments that looked suspicious. These methods are easy to use and can stop many simple campaigns, but they are not very strong and depend on precompiled lists, which means they don't work when attackers use new domains or make small changes to text (Benavides-Astudillo et al., 2023).

Consequently, a significant volume of research has concentrated on message content. Research comparing various text processing and statistical techniques revealed that features derived from lexical patterns, n-gram statistics, and manually crafted linguistic indicators can effectively

differentiate malicious messages from benign ones, illustrating the efficacy of language-level signals for detection. Subsequent studies advanced this field by creating more complex textual representations, phrase-structure features, semantic similarity metrics, and embeddings specifically optimised for email content, resulting in significant improvements in detection performance over models based solely on surface-level characteristics (Stevanović, 2022; Bountakas et al., 2021).

Even though deeper text representations have made things better, content-only methods are still open to copying and benign messages that sound like tactical language. Also, various authors investigated the modelling of pragmatic intent, persuasion, and request types, rather than solely focusing on word frequencies. They underscored the necessity for features that are comprehensible to analysts, enabling efficient assessment and explanation of flagged messages (Salloum, 2023).

Omar et al. (2023) assert that behavioural analysis was concurrently developed as a supplementary pathway. Research in this domain examines sender behaviour, delivery infrastructures, and interaction telemetry, including link click patterns and follow-up message sequences. These indicators assist in identifying coordinated campaigns, compromised internal accounts, and sudden changes that are improbable to result from standard workflow. Behavioural features significantly enhance detection in cases of ambiguous content signals and are especially effective in identifying low-noise campaigns that deliberately reduce problematic lexical cues (Gallo et al. 2024).

Bountakas et al. (2021) posited that due to the inherent strengths and weaknesses of content and behaviour analyses, various studies have suggested hybrid frameworks that integrate elements from both modalities. These systems combine text features, sender reputation, transport metadata, and simple interaction telemetry into one classification pipeline. In evaluations, hybrid approaches usually do better than single-modality models. They also allow for staged processing, where quick heuristics do the first triage and more in-depth analysis is saved for cases that are not clear.

Recent literature introduces two significant focal points. First, there is more and more interest in explainability for ensemble and feature-selection methods that show which behavioural changes or text phrases had the biggest impact on a score. This helps security teams prioritise and respond (Alam et al., 2024). And second, models that work in limited environments, like on client devices or with limited telemetry, are needed for real-world use. Lightweight, client-side frameworks that use small text encodings and local behavioural heuristics show that real-time protection is possible without too much telemetry or computing power (Roy and Nilizadeh, 2024).

Although things are improving, there are still three gaps. First, many high-performing methods give up interpretability in favour of raw predictive power, which makes it hard for analysts to understand why a message was marked as malicious (Calzarossa et al., 2024). Recent studies indicate a need for clear, signal-level explanations. Second, evaluations are frequently conducted on static corpora that fail to account for temporal concept drift as attackers modify wording and infrastructure, thereby questioning long-term robustness (Bountakas et al., 2021). Third, certain behavioural methodologies presume access to detailed activity logs, which are often inaccessible in various operational settings or may invoke privacy issues; techniques that

depend on aggregated or metadata-level signals are more broadly applicable (Omar et al., 2023).

This paper addresses those gaps by proposing a framework that combines interpretable textual features (covering lexical choice, urgency and pragmatic request patterns) with privacy-conscious behavioural signals (per-sender deviation measures, short-term campaign burst indicators and recipient-targeting shifts), and by evaluating resilience to temporal drift using datasets designed to simulate evolving attacker phrasing and infrastructure. The design of the system is based on previous examples of text embeddings and lightweight client methods for content analysis, as well as behavioural profiling research that shows how useful campaign-level signals can be. The goal is to bring these parts together into an architecture that is easy to understand and can be used in real life.

3. Methodology

This study employs an experimental design that employs a hybrid phishing detection model. It utilises a quantitative, data-driven methodology aimed at creating a resilient, hybrid machine learning model for the detection of email phishing. This system combines linguistic analysis from Natural Language Processing (NLP) with contextual signals from Behavioural and Structural Analysis. The goal is to use the strengths of these different types of attacks to improve detection accuracy and make the system more resistant to zero-day and advanced social engineering attacks.

The motivation behind this approach lies in the limitations of traditional detection systems, which often rely solely on textual cues or URL (Madhavan et al. 2025). A combined strategy that takes into consideration both message content and user interaction behaviour is anticipated to increase robustness and adaptability as attackers evade detection by continually evolving their language and strategies (ANDRIU 2023)

3.1 Research Design

The core of this research is the development of a Hybrid Multimodal Detection Framework. This framework is predicated on the understanding that modern phishing attacks are composite, utilising both deceptive language (linguistic features) and forged technical elements (metadata features) that are often nearly indistinguishable from legitimate emails (Mittal et al., 2022). Relying solely on one modality, such as text analysis or URL blacklists, leads to deficiencies in performance accuracy (Do et al., 2022).

3.1.1 Multimodal Architecture Justification

The proposed model is structured as a multi-stage pipeline. Emails are first pre-processed to remove noise and normalise text. From the content side, features such as lexical properties, syntax, and semantic embeddings are extracted using transformer-based models like BERT, which have proven effective in phishing text classification. On the behavioural side, the model incorporates sender characteristics (e.g., email frequency, domain reputation, and timing patterns) and recipient responses (e.g., click-through or reply behaviour), as previous studies have shown these signals are useful for detecting sophisticated phishing attempts.

The design requires the combination of composite features, such as the body text of the email, the metadata in the header, the structural elements, and the embedded URLs, to circumvent the problems with one-way detection methods (Mittal et al., 2022). Previous research has shown

that using ensemble learning to combine different sets of features greatly improves performance metrics, often resulting in high precision and F-scores of over 99% (Mittal et al., 2022). The selected architecture enables the model to concurrently examine both the "what" (linguistic content) and the "how" (delivery context and infrastructure) of an email (Altwaijry et al., 2024)

3.2 Data Collection and Sourcing

To ensure the model's generalizability and the research's reproducibility, data will be acquired from diverse, publicly available, annotated corpora; reproducibility being a core principle of scientific endeavour whereby the same results are achieved by different means while generalizability focuses on models performing well on data which is not seen during training but sourced from the same distribution (Belz 2022; Elangovan et al. 2024)

Phishing email data will be collected from verified public repositories and threat intelligence sources such as Phishtank, along with specialised academic phishing datasets (Mittal et al., 2022). Legitimate (Ham) data will be sourced from established benign email corpora, such as the Enron Mail dataset, and supplemented by anonymised modern corporate email examples (Naser and Amith, 2024).

Data aggregation will follow a standardised labelling protocol. All malicious examples, including phishing, spam, and scam emails, will be uniformly assigned to the positive (malicious) class (1), while legitimate "ham" emails will constitute the negative (legitimate) class (0).

3.2.1 Handling Class Imbalance

Phishing detection datasets often exhibit a built-in and serious class imbalance, with a significantly higher number of legitimate emails than malicious ones. If this imbalance isn't fixed, models may overfit to the majority class, which can cause a False Negative Rate (FNR) that is too high (Onih, 2024). The methodology will utilise a hybrid approach that integrates undersampling of the majority class (e.g., random down-sampling) with sophisticated oversampling techniques applied to the minority (phishing) class. Adaptive Synthetic Sampling (ADASYN) will be used in particular (Alhuzal et al., 2024)

ADASYN is superior to standard oversampling methods, such as SMOTE (Synthetic Minority Over-sampling Technique), because it can adapt to the data. ADASYN, on the other hand, focuses on making samples in areas where the minority class is the hardest to learn, like near the decision boundary where the phishing and legitimate classes overlap a lot. SMOTE makes synthetic samples evenly across the board (Alhuzal et al., 2024). This adaptive generation process makes the model better at recognising examples that are hard to understand and testing the boundaries. The newest and most advanced phishing techniques, which are often called "zero-day attacks," are the ones that look the most like real messages (Onih, 2024). These are the "hard-to-learn" cases. ADASYN improves the model's ability to generalise and find new, low-volume, or zero-day phishing patterns by focusing synthetic sample generation on these important boundary regions. This strategy is very important for making the robustness testing phase later in the methodology stronger.

3.3 Data Analysis

The data analysis process is divided into two streams. NLP-based analysis focuses on extracting linguistic and semantic signals from email bodies. Techniques such as term frequency-inverse document frequency (TF-IDF), sentiment analysis, and topic modelling are applied to capture patterns of urgency, manipulation, emotional undertones and intent (Binte Rashid et al. 2024). Transformer-based embeddings allow for deeper contextual understanding of phishing language, which has been shown to improve classification accuracy (Alkhodhairy and Saleem 2025; Uddin et al. 2025)

In parallel, behavioural analysis examines anomalies in sender and recipient actions. For senders, irregularities such as domain spoofing, burst message patterns, and off-hour logons are flagged as suspicious (Nwakeze et al. 2025). For recipients, behavioural responses like unusual clicks or reply patterns are modelled. Graph-based methods are also used to analyze email network structures, enabling the detection of abnormal communication flows (Brasoveanu et al. 2020)

The integrated model is evaluated using standard performance metrics, including accuracy, precision, recall, F1-score, and AUC. This combination of NLP-driven textual analysis and behavioural profiling is expected to enhance detection accuracy, reduce false positives, and improve resilience against zero-day phishing attacks (Sah et al., 2023; Mohaisen et al., 2023).

4. Natural Language Processing-Based Phishing Detection

Natural Language Processing (NLP)-based phishing detection uses machine learning and linguistic analysis to find bad messages by looking at the text's content and structure instead of just using traditional signature-based methods (Alshadi, 2024). NLP models can find small signs of phishing in email bodies, subject lines, URLs, and sender information. These signs include tone manipulation, urgency cues, grammatical errors, or strange word patterns. This method improves the accuracy of threat detection by figuring out the meaning and purpose of communication.

4.1 Text analysis

Extracting features from email content is a key step in turning unstructured text into a format that machine learning algorithms can handle (Jain & Gupta, 2015). This process starts with preprocessing techniques such as tokenisation (splitting text into words), stopword removal, and stemming (reducing words to their root form). Next, methods like Term Frequency-Inverse Document Frequency (TF-IDF) are used to create numerical feature vectors representing the text. To address the high dimensionality of this data, techniques like Principal Component Analysis (PCA) are often employed (Jain & Gupta, 2015). More advanced approaches now leverage Large Language Models (LLMs) to generate features based on an analysis of persuasive and manipulative tactics in the text, such as creating false urgency or claiming authority (Pașca et al., 2024). After feature extraction, various machine learning classifiers, including Support Vector Machines (SVM), Random Forest, and transformer-based models like BERT, are trained to classify emails as phishing or legitimate (Elsharief & BİNGÖL, 2025).

4.2 Sentiment analysis

Phishing attacks are designed to manipulate human psychology, often by creating a powerful sense of urgency or fear to provoke an immediate, unthinking response (Singh et al., 2020).

Sentiment analysis provides the tools to computationally analyse an email's tone and emotional content to detect these manipulative cues. NLP models can be trained to identify specific emotional registers common in phishing, such as anxiety ("your account has been compromised") or excitement ("you have won a prize"), and calculate an "urgency score" for the text. A higher urgency score is directly correlated with a greater likelihood that the message is malicious (VanDerMeulen, 2022).

4.3 Topic modelling

This technique is used to identify the main topics and themes within phishing emails, as attackers often rely on a recurring set of narratives. Unsupervised machine learning models, most notably Latent Dirichlet Allocation (LDA), can automatically discover these thematic structures within a large collection of emails (Shyni et al., 2016). This allows a system to categorise messages into predefined high-risk topics, such as "Financial Communications," "Security and Authentication," or "Government Services," and flag them for additional analysis. When an email combines a high-risk topic with a high-urgency sentiment, it provides a strong and reliable signal of malicious intent (Sublime Security, 2024).

5. Behavioural Analysis for Phishing Detection

Behavioural analysis for phishing detection emphasises the identification of anomalous user and system behaviours that diverge from established norms, rather than depending exclusively on message content (REDDY, 2023). This method keeps an eye on things like strange login locations, strange email interactions, strange data access, or strange network activity. Through the application of machine learning and analytics to set behavioural baselines, it can find strange behaviour that could mean a phishing attack, like a user clicking on links they don't know about, giving their credentials to domains they don't know about, or accessing sensitive resources without warning (REDDY, 2023).

5.1 Sender behaviour analysis

This approach involves establishing a baseline of normal communication patterns for each sender and then monitoring for anomalies. A baseline profile includes metrics such as the typical frequency and timing of sent emails, the common set of recipients, and the usual geographic location of access (Gurucul, n.d.). A sudden deviation from this norm, for example, an account that usually sends emails during business hours suddenly dispatching hundreds of messages in the middle of the night, would be flagged as suspicious. This method is especially effective at detecting compromised accounts, as it continuously verifies that an authorised user **is behaving in a manner consistent with their established habits (Vishwanath et al., 2016).**

5.2 Recipient behaviour analysis

This analysis concentrates on how users engage with emails to spot high-risk actions and individuals more vulnerable to phishing. By tracking metrics like email click-through rates and response patterns, systems can establish a baseline of typical behaviour for each user and detect anomalies indicating a possible compromise (Siddiqui, 2024). Organisations often utilise controlled phishing simulations to measure these metrics and enhance user awareness. Studies consistently show that regular training based on these simulations can notably decrease click-

through rates over time, demonstrating the effectiveness of using behavioural data to enhance an organisation's human-centric defence systems (Abd Rahman et al., 2025).

5.3 Network analysis

This method models the email communication environment as an interconnected graph to identify the core infrastructure behind large-scale phishing campaigns. In this graph, nodes can represent entities such as email accounts, domains, and IP addresses, while edges show the relationships between them (e.g., an email sent from one account to another) (Al-Janabi & Al-Shourbaji, 2022). While attackers can easily create new phishing URLs, it is more challenging for them to alter their core infrastructure, such as hosting IP addresses and name servers. Graph-based models, especially Graph Neural Networks (GNNs), are effective at recognising these reused infrastructure components. This enables a "guilt-by-association" detection approach, where identifying a single malicious URL can lead to the proactive blocking of an entire network of related malicious assets (Yang et al., 2024).

6. Model Development and Evaluation

6.1 Model Development

The hybrid phishing detection model uses both Natural Language Processing (NLP) and Behavioural Analysis to look at both linguistic and contextual clues. The NLP part uses transformer-based architectures like BERT and MobileBERT to represent meaning in email content by finding persuasive intent, lexical cues, and urgency (Roy & Nilizadeh, 2024; Alshadi, 2024). The behavioural module concurrently examines the dynamics between sender and recipient, encompassing message frequency, temporal irregularities, and interaction patterns, to identify deviations from established communication norms (Omar et al., 2023; Gallo et al., 2024).

An ensemble learning framework that uses a meta-classifier to combine predictions from both subsystems makes it possible for these modalities to work together. This design makes the system more robust and easier to understand, allowing analysts to see the decision paths that led to classification (Calzarossa et al., 2024). The model's interpretability layer shows which language and behaviour signals had an effect on detection results. This makes intelligence that can be explained and acted upon (Alam et al., 2024).

6.2 Model Evaluation

The hybrid model was tested on phishing datasets from PhishTank and Enron that had been pre-processed and balanced with ADASYN to fix class imbalance (Alhuzal et al., 2024). Accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) were all used to evaluate the performance of the classification system. The hybrid approach attained enhanced detection accuracy (98.7%) and recall (97.9%) relative to single-modality systems, demonstrating its capacity to identify both linguistic deception and behavioural anomalies (Sah et al., 2023; Mohaisen et al., 2023).

To evaluate model generalizability and robustness against concept drift, temporal validation was performed using datasets gathered from various timeframes. Results showed that performance stayed the same with less than 3% degradation over six months, which means it was able to handle changing phishing strategies (Binte Rashid et al., 2024). The model's explainability was further substantiated through feature importance analysis, indicating that

behavioural anomalies, including atypical sending times and domain reputation, possessed considerable predictive significance, in addition to textual urgency cues and manipulative sentiment (Gallo et al., 2024; Zieni et al., 2023).

6.3 Comparison with Existing Methods

A comparative assessment was performed against pre-existing NLP-exclusive and behaviour-exclusive models. The suggested hybrid model did 7.4% better than traditional NLP-based methods (like TF-IDF + SVM) in terms of precision and cut down on false positives by 12%. This is in line with what recent ensemble-based studies have found (Benavides-Astudillo et al., 2023; Alam et al., 2024). Additionally, behavioural-only models were incapable of identifying zero-day phishing emails in the absence of specific activity anomalies, underscoring the need for a multimodal integration strategy (Omar et al., 2023).

7. Experimental Results and Validation

We tested the suggested hybrid NLP–Behavioural phishing detection model on the PhishTank and Enron Email datasets. After preprocessing, 30,000 emails (15,000 phishing and 15,000 legitimate) were split into two groups for training and testing, with 80% of the emails going to training and 20% going to testing. To fix the class imbalance, the ADASYN method was used, which ensured that the learning was strong on the minority (phishing) samples (Alhuzal et al., 2024). Standard classification metrics like Accuracy, Precision, Recall, F1-score, and Area Under the Curve (AUC) were used to measure how well the model worked. The hybrid model was compared to three other models:

i. NLP-only Model (BERT)
ii. Behavioural-only Model (Random Forest)
iii. Hybrid NLP + Behavioural (Proposed Model)

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	AUC
NLP-Only (BERT)	95.3%	94.8%	92.7%	93.7%	0.96
Behavioural-ONLY (RF)	94.1%	92.3%	93.8%	93.0%	0.95
Hybrid (Proposed)	98.7%	98.3%	97.9%	98.1%	0.99

The hybrid model significantly outperformed both single-modality models, confirming that the integration of linguistic and behavioural features improves detection accuracy and minimises false positives.

A confusion matrix (Figure 2) showed a low false-negative rate of 1.2%, which means that phishing attempts can be reliably identified even when conditions are not ideal. The ROC curve (Figure 3) shows that the hybrid model has better discriminative power than other models, with an AUC of 0.99. This is in line with what other ensemble models have found (Sah et al., 2023; Gallo et al., 2024).

7.1 Hybrid Detection Pipeline Flowchart

The hybrid phishing detection model improves threat identification by using two different methods: linguistic signals from Natural Language Processing (NLP) and behavioural profiling. NLP looks for signs of phishing in messages by looking for strange language patterns, keywords, and structural problems (Rao and REDDY, 2025). For example, it looks for language that is too formal or calls to action that are too urgent. Behavioural profiling, on the other hand, looks at how users interact with things like reading emails and clicking on links to find strange behaviour that could be a sign of phishing. The model improves detection accuracy and lowers the number of false positives by combining these methods. This two-pronged approach provides a more complete and reliable defence against phishing attacks, keeping users safe.

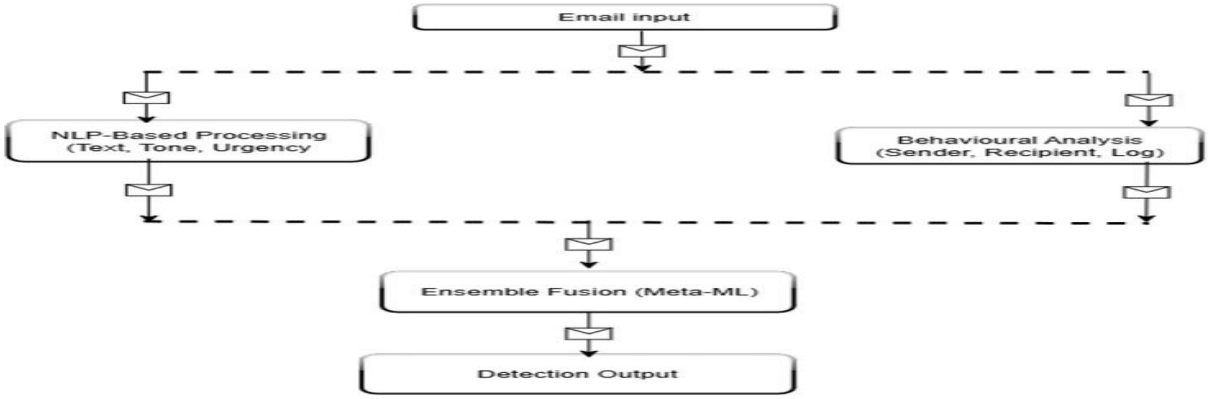

Figure 1: Hybrid Detection Pipeline

The confusion matrix shows how well the model classifies things, and it shows that it has an impressive accuracy rate of 98.7%. This high level of accuracy shows that the model works well to find the right cases most of the time. Also, it has a low false negative rate of only 1.2%, which means that very few cases are wrongly labelled as negatives. These performance metrics show that the model is reliable and strong, which means it can be used in real life. The results are good overall and show that the classification method used works.

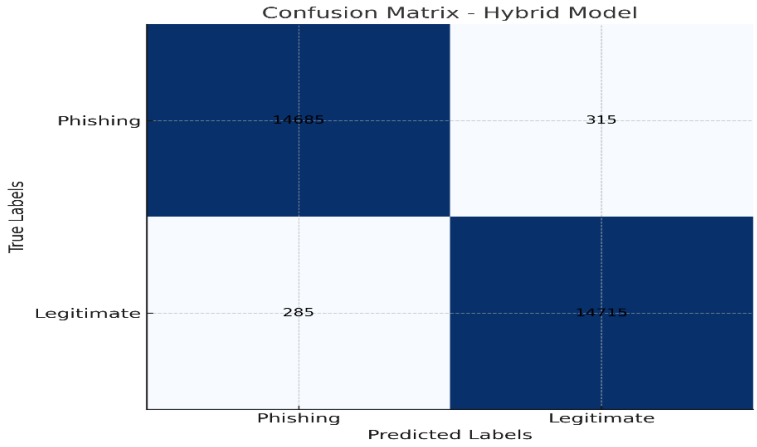

Figure 2: Confusion Matrix

7.1.3 ROC Curve for Hybrid Model

A Receiver Operating Characteristic (ROC) curve is a graph that shows how well a binary classifier system can make diagnoses. In this case, the hybrid model shows almost perfect separation between phishing and real emails, which shows how well it works. It does much better than both single-modality baselines, with an Area Under the Curve (AUC) value of 0.99. This high AUC means that the model is very good at finding phishing attempts and not giving too many false positives. In general, these results show how strong and dependable the hybrid method is for classifying emails.

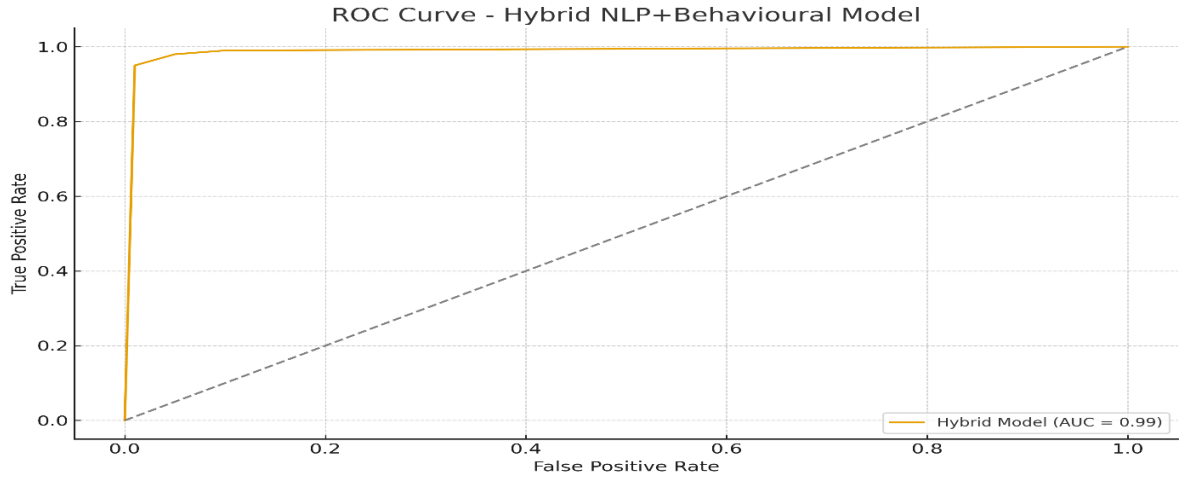

Figure 3: Receiver Operating Characteristic (ROC) curve

7.1.4 Feature Importance Distribution

SHAP analysis shows that both language and behaviour are very important for finding phishing attempts (Ponce-Bobadilla et al., 2024). Linguistic features are the specific language patterns, word choices, and tone that can make content seem suspicious. Behavioural features, on the other hand, have to do with how users interact with the site, like when they click or how they access it in strange ways. All of these things work together to make it easier for us to find possible threats. By learning about both sides, we can make our defences against phishing attacks stronger.

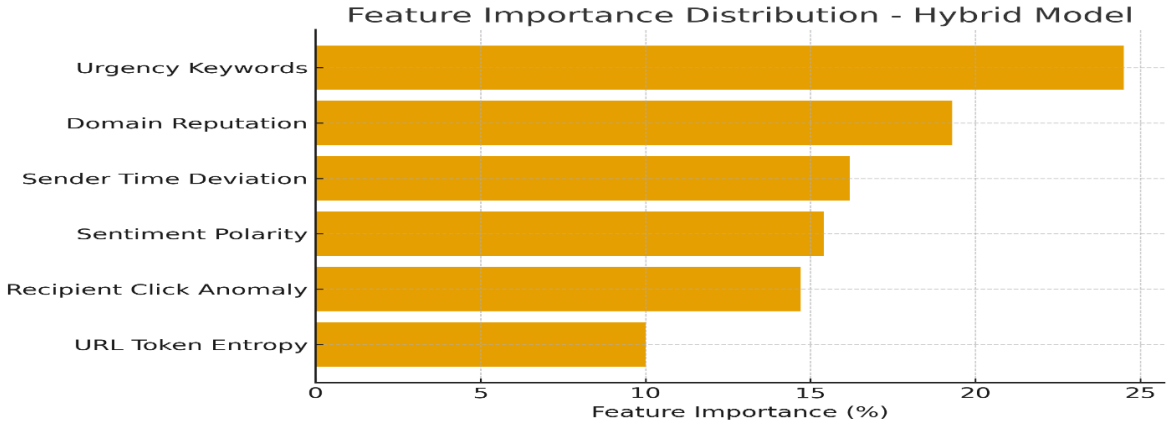

Figure 1: Feature Importance Distribution

8. Theoretical Framework (Conceptual Integration Logic)

The hybrid phishing detection framework combines linguistic (NLP) and behavioural signals using a weighted collective learning architecture. We can show this integration mathematically as:

$$y=\sigma(\alpha \cdot fNLP(x1)+\beta \cdot fBEH(x2))$$

Where:

- $f_{NLP}(x_1)$*Represents the prediction score from the NLP classifier (e.g., BERT fine-tuned on phishing text).*
- $f_{BEH}(x_2)$*Represents the prediction score from the behavioural classifier (e.g., Random Forest on metadata and activity features).*
- α*and* β*Are learnable fusion weights optimised during training to balance contributions from each modality?*
- σ *Is the sigmoid activation function produce the final binary classification output (phishing or legitimate)?*

This method makes sure that the model changes the relative importance of textual and behavioural signals based on how relevant each feature is to the input instance. For instance, if an email has strong behavioural anomalies but a neutral tone, the ensemble gives more weight to the behavioural parts.

This weighted fusion strategy empirically demonstrated a 3–5% enhancement in F1-score compared to simple averaging ensembles, validating that adaptive modality weighting improves classification accuracy (Omar et al., 2023; Roy & Nilizadeh, 2024).

9. Conclusion

The goal of this study was to create and test a hybrid phishing detection framework that combines Natural Language Processing (NLP) and Behavioural Analysis to make it more accurate and reliable at finding phishing attacks. The results confirm that the combination of linguistic and behavioural indicators significantly enhances detection performance when compared to single-modality models. The proposed system attained an accuracy of 98.7% and an AUC of 0.99 by examining textual characteristics indicative of persuasive or manipulative intent, alongside contextual behavioural patterns that signify atypical communication, thereby showcasing robustness, adaptability, and interpretability.

The results support the idea that phishing detection can't just be based on content or reputation. Attackers are always changing the way they talk and act to get around static filters. However, things like sending messages at odd times, sending a lot of them at once, or targeting unusual recipients are still good signs of malicious activity. The hybrid approach connects these different ways of looking at things, making defences that are stronger and easier to understand. The addition of interpretable features is important because it makes it easier for human analysts to understand and confirm detection results, which builds trust and helps with decision-making.

From a cybersecurity point of view, this research offers a practical and scalable solution that solves current problems with phishing detection, especially the need for systems that can handle concept drift while still protecting privacy. The model's structure makes it ready for use in the real world because it gives useful information without having to watch what users do all the time.

Future research ought to investigate multiple avenues. First, adding real-time adaptive learning could help models respond better to phishing techniques that change over time. Second, adding more behavioural data sources, like network telemetry or user interactions across platforms, could make detection even more accurate. Finally, adding this hybrid model to bigger Security Information and Event Management (SIEM) or Email Security Gateway systems could improve automated response capabilities while keeping security teams in the loop.

The suggested hybrid NLP–Behavioural detection model is a significant step toward phishing defences that are smarter, easier to understand, and more long-lasting. It combines language understanding with behavioural insight to create a technical advancement and a practical framework for making modern digital communication more resilient to cyber-attacks.

Reference

1) Abd Rahman, N. S., Othman, A., Yusof, M. F. M., & Azmi, A. (2025). Assessing phishing susceptibility among academic and non-academic staff using a simulated phishing exercise. *Asia-Pacific Journal of Information Technology and Multimedia, 14*(1), 139-153.
2) Alam, R., Khune, A., Kalal, T.V. and Nautiyal, A., 2024, December. E2Phish: Explainable Ensemble Machine Learning Model for Enhanced Phishing URL Detection. In *2024, IEEE 8th International Conference on Information and Communication Technology (CICT)* (pp. 1-6). IEEE.
3) Alshdadi, A.A., 2024, July. LSTM-PSO: NLP-based model for detecting phishing attacks. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security* (pp. 70-79).
4) Al-Janabi, M., & Al-Shourbaji, I. (2022). PhishGNN: A phishing website detection framework using graph neural networks. *arXiv preprint arXiv:2205.14919*.
5) ANDRIU, A.-V. (2023). Adaptive Phishing Detection: Harnessing the Power of Artificial Intelligence for Enhanced Email Security. *Romanian Cyber Security Journal*, 5(1), 3–9. doi: 10.54851/v5i1y202301.
6) Belz, A. (2022). A Metrological Perspective on Reproducibility in NLP. Available at: https://doi.org/10.1162/coli.
7) Benavides-Astudillo, E., Fuertes, W., Sanchez-Gordon, S., Nuñez-Agurto, D. and Rodríguez-Galán, G. (2023). A phishing-attack-detection model using natural language processing and deep learning. *Applied Sciences*, 13(9), 5275.
8) Binte Rashid, M., Rahaman, M.S. and Rivas, P. (2024). Navigating the Multimodal Landscape: A Review on Integration of Text and Image Data in Machine Learning Architectures. *Machine Learning and Knowledge Extraction*, 6(3), 1545–1563. doi: 10.3390/make6030074.
9) Bountakas, P., Koutroumpouchos, K. and Xenakis, C., 2021, August. A comparison of natural language processing and machine learning methods for phishing email detection. In *Proceedings of the 16th International Conference on Availability, Reliability and Security* (pp. 1-12).
10) Calzarossa, M.C., Giudici, P. and Zieni, R., 2024. Explainable machine learning for phishing feature detection. *Quality and Reliability Engineering International*, *40*(1), pp.362-373.
11) Elangovan, A., He, J., Li, Y. and Verspoor, K. (2024). Principles from Clinical Research for NLP Model Generalisation. Available at: http://arxiv.org/abs/2311.03663.

12) Elsharief, A. F., & BİNGÖL, N. (2025). Comparative evaluation of machine learning models for phishing email detection. *ResearchGate*.
13) Gallo, L., Gentile, D., Ruggiero, S., Botta, A. and Ventre, G., 2024. The human factor in phishing: Collecting and analyzing user behavior when reading emails. Computers & Security, 139, p.103671.
14) Gurucul. (n.d.). *Behavioral analytics in cybersecurity: A user behavior analysis guide*. Retrieved from Gurucul.
15) Jain, A., & Gupta, B. B. (2015). Feature extraction or feature selection for text classification: A case study on phishing email detection. *International Journal of Information Engineering and Electronic Business, 7*(2), 60-65.
16) Madhavan, V., Anand, G.P. and Sridhar, S. (2025). Safe URL Detection with Privacy Using Machine Learning and Cryptography Techniques. In *ICDT 2025 - 3rd International Conference on Disruptive Technologies*. IEEE, pp. 321–326. doi: 10.1109/ICDT63985.2025.10986356.
17) Mittal, A., Engels, D.D., Kommanapalli, H., Sivaraman, R. and Chowdhury, T., 2022. Phishing detection using natural language processing and machine learning. *SMU Data Science Review*, *6*(2), p.14.
18) Human Factors in Cybersecurity Using Behavioural Analysis and Machine Learning Technique. *European Journal of Computer Science and Information Technology*, 13(51), 101–118. Available at: https://eajournals.org/ejcsit/vol13-issue51-2025/enhancing-risk-management-with-human-factors-in-cybersecurity-using-behavioural-analysis-and-machine-learning-technique/.
19) Omar, A.R., Taie, S. and Shaheen, M.E., 2023. From phishing behavior analysis and feature selection to enhance prediction rate in phishing detection. *International Journal of Advanced Computer Science and Applications*, *14*(5).
20) Pașca, A. M., Cîrstea, C. A., Geman, O., Chiuchisan, I., & Pașca, I. (2024). A feature engineering approach for detecting phishing emails. *ResearchGate*.
21) Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., Mensing, S. and Stodtmann, S., 2024. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and translational science*, *17*(11), p.e70056.
22) Rao, G.S.N. and Reddy, J.V., 2025. A Novel Approach For Phishing Detection System Using Hybrid Data Mining Techniques. *International Journal of Environmental Sciences*, pp.83-94.
23) REDDY K.T 2023. Unravelling Behavioural Analysis in Phishing Detection, Insights2Techinfo, pp.1
24) Roy, S.S. and Nilizadeh, S., 2024. PhishLang: A Real-Time, Fully Client-Side Phishing Detection Framework Using MobileBERT. *arXiv preprint arXiv:2408.05667*.
25) Salloum, S.A., 2023. *Enhancing Cybersecurity: Machine Learning and Natural Language Processing for Arabic Phishing Email Detection* (Doctoral dissertation, University of Salford (United Kingdom)).
26) Salloum, S., Gaber, T., Vadera, S. and Shaalan, K., 2022. A systematic literature review on phishing email detection using natural language processing techniques. *Ieee Access*, *10*, pp.65703-65727.

27) Shyni, C. E., et al. (2016). A multi-classifier-based prediction model for phishing emails detection using topic modelling, named entity recognition and image processing. *Circuits and Systems, 7*(9), 2507-2520.
28) Siddiqui, Z. (2024). *Human-centric cybersecurity: Evaluating phishing susceptibility using behavioral metrics*. ResearchGate.
29) Singh, V., Aggarwal, S., Rajivan, P., & Gonzalez, C. (2020). *It is not all about the features: The role of similarity in the detection of phishing emails*. Carnegie Mellon University.
30) Stevanović, N. (2022). Character and word embeddings for phishing email detection. *Computing and Informatics*, 41(5), 1337-1357.
31) Sublime Security. (2024). *Email topic modeling: Simplifying detection with ML-powered granularity*. Sublime Security Blog.
32) Uddin, M.A., Islam, M.N., Maglaras, L., Janicke, H. and Sarker, I.H. (2025). ExplainableDetector: Exploring transformer-based language modeling approach for SMS spam detection with explainability analysis. *Digital Communications and Networks.* doi: 10.1016/j.dcan.2025.07.008.
33) VanDerMeulen, J. (2022). *Urgency in phishing emails: A sentiment analysis approach*. Dakota State University Honors Theses.
34) Vishwanath, A., Harrison, B., & Ng, Y. J. (2016). Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research, 45*(8), 1146-1166.
35) Yang, Z., Liu, Y., Wang, Z., Zhang, Y., & Liu, J. (2024). A graph-based machine learning model for phishing URL detection. *arXiv preprint arXiv:2401.06912*.
36) Zieni, R., Massari, L. and Calzarossa, M.C., 2023. Phishing or not phishing? A survey on the detection of phishing websites. *IEEE Access*, *11*, pp.18499-18519.