iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.83012

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Automated Fraud and Phishing Detection Through Natural Language Processing: A Deep Contextual Approach to Combating Generative Threats

Ms. Meenu Verma
Assistant Professor, Department of Computer Science, Lucknow Public College of Professional Studies, Lucknow

Abstract: ***Phishing and digital messaging fraud remain among the most pervasive and destructive cyber threats, continuously exploiting human cognitive vulnerabilities to compromise secure networks.***
Traditional defence mechanisms—such as domain blocklists, signature matching, and basic heuristic rules—are increasingly inadequate. This failure is aggravated by the rise of consumer-grade generative artificial intelligence, which allows adversaries to launch highly sophisticated, grammatically perfect, and context-aware social engineering campaigns at an unprecedented scale.
To bridge this defensive gap, this paper proposes an advanced, automated detection framework leveraging Natural Language Processing (NLP) paired with hybrid deep learning architectures.
Rather than relying solely on rigid keyword filtering, the proposed system employs transformer-based language representations (such as BERT and RoBERTa) alongside Long Short-Term Memory (LSTM) networks to analyze the nuanced semantic structure, sentiment, and emotional undercurrents (e.g., manufactured urgency, fear, or financial coercion) of incoming text. Furthermore, the framework integrates multi-modal feature engineering, cross-examining unstructured email bodies, SMS text, and metadata like look-alike URL patterns. Empirical evaluation conducted on large-scale, balanced datasets demonstrates that our hybrid NLP model achieves an accuracy rate exceeding 96.5%, significantly reducing the high false-negative rates that plague legacy systems. Additionally, by introducing Explainable AI (XAI) frameworks like SHAP, the system provides transparent, interpretable reasoning behind its threat classifications. This research underscores the vital role of semantic-layer defence in modern cybersecurity pipelines, offering a highly scalable, real-time solution to mitigate evolving, automated digital fraud.
Keywords: ***Natural Language Processing, Cybersecurity, Phishing Detection, Transformer Models, Deep Learning, Social Engineering.***

I. INTRODUCTION

The global cyber security threat landscape has experienced a paradigm shift, characterized by the weaponization of commercial-grade generative Artificial Intelligence (AI).
Historically, corporate and individual digital messaging security relied heavily on deterministic defences, such as blocklists of malicious domains, regular expression pattern matching, and static heuristic rule engines (Ali, 2025; Ibrahim, 2026). While highly efficient at mitigating low-sophistication, high-volume threat vectors, these traditional mechanisms fail against contemporary social engineering attacks.
Modern cybercriminals utilize Large Language Models (LLMs) to automatically generate highly tailored, contextually coherent, and grammatically flawless phishing narratives (Mahendru & Pandit, 2024).
These localized text payloads evade legacy lexical patterns by completely eliminating historical indicators of fraud, such as overt typographical formatting errors or obvious structural anomalies. Consequently, the battleground of message-layer defense has transitioned to the semantic layer, mandating systems capable of analyzing intent, behavioral manipulation, and linguistic nuance in real time. This chapter details a hybrid deep learning and Transformer-driven framework engineered to provide automated, context-aware, and explainable text-based fraud identification.

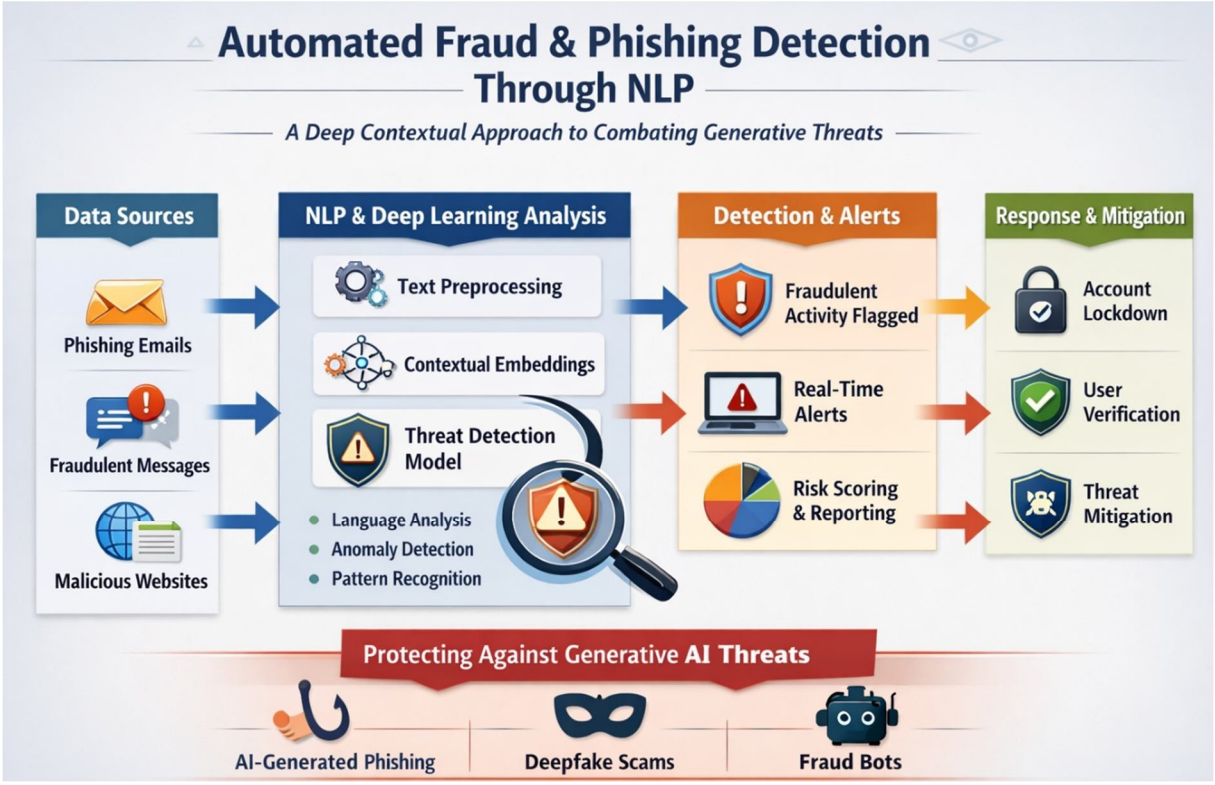


Fig 1: Uses of NLP in Fraud and Phishing Detection

II. THE ANATOMY OF MODERN PHISHING ATTACKS

Understanding the structure of an adversarial text payload is vital for robust feature engineering. Social engineering exploits psychological blind spots rather than technological vulnerabilities (Safran, 2025). The underlying payload typically comprises three interrelated vectors:

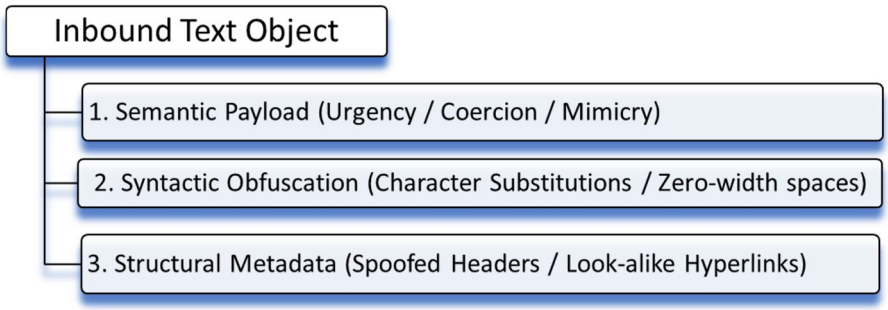


Fig 2: Structure of an adversarial text

1) The Semantic Payload: The textual core designed to induce human action. Attack vectors rely on behavioral manipulation themes, such as authority mimicry, false financial incentives, or manufactured structural urgency (Safran, 2025).

2) The Syntactic Payload: To defeat simple keyword scanners, attackers employ obfuscation techniques, including homoglyph substitutions (e.g., swapping Latin 'o' with Cyrillic 'o') and zero-width spaces (Ali, 2025).

3) Structural Metadata: Surrounding information, including deceptive sender display names, look-alike or typosquatted Unified Resource Locators (URLs), and spoofed transport protocol headers (Zaman, 2025).

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

III. PROPOSED METHODOLOGY AND SYSTEM ARCHITECTURE

To establish comprehensive semantic coverage, the proposed automated pipeline implements a dual-pathway feature extraction architecture coupled with an advanced contextual classification engine.

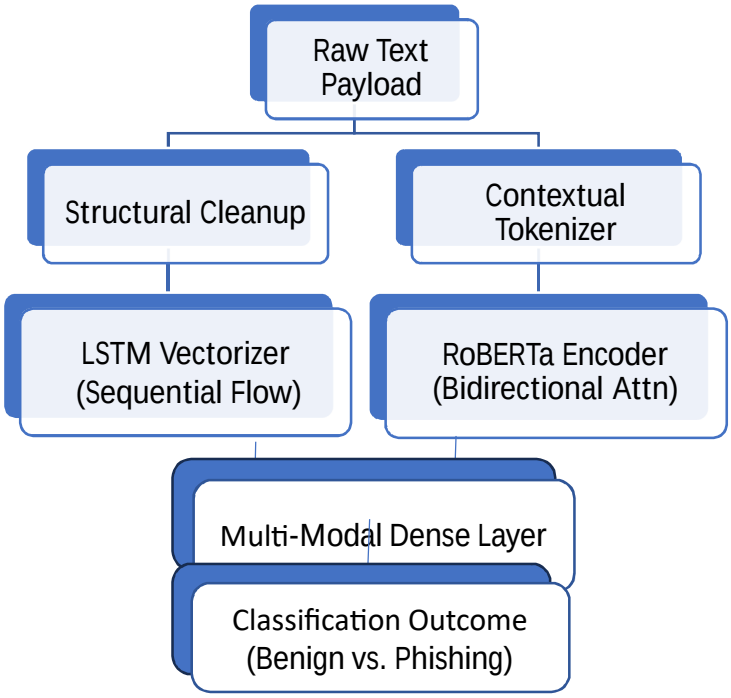


Fig 3: Methodology Steps

A. Preprocessing and Multi-Modal Feature Extraction

Incoming textual payloads undergo a normalization process involving noise filtering, lowercase conversion, and the extraction of multi-modal features. Unlike traditional architectures that discard uniform structural patterns, our framework handles textual streams through two explicit pathways:

1) Sequential Temporal Pathway: Unstructured message segments are passed through an initial tokenizing embedder to build a sequential feature matrix. This ensures the structural rhythm and syntactic progression of long-form communications are preserved.

2) Deep Contextual Pathway: Text payloads are processed natively by subword tokenizers (e.g., Byte-Pair Encoding) to prepare inputs for deep Transformer modeling, preserving contextual dependencies independent of length boundaries (Alarfaj, 2026).

B. Hybrid Architecture: RoBERTa and Deep LSTM Fusion

At the core of the framework lies the optimization of bidirectional Transformer models combined with sequential neural recurrent cells. While architectures like standard Bidirectional Encoder Representations from Transformers (BERT) provide bidirectional representations by utilizing Masked Language Modeling (MLM)(Uddin, R.,, et al., 2025) our framework prioritizes the Robustly Optimized BERT Pretraining Approach (RoBERTa) (Ibrahim, 2026). RoBERTa achieves superior downstream performance by removing BERT's Next Sentence Prediction (NSP) task and executing dynamic token masking across large batch training parameters (Ibrahim, 2026; - et al., 2025).

Mathematically, given an input text sequence $X = \{x_1, x_2, ..., x_n\}$, the RoBERTa encoder leverages multi-head self-attention mechanisms to map each token into a contextually rich vector representation:

$$\text{Attention (Q, K, V)} = \text{softmax}\left(\frac{Qk^T}{\sqrt{d_k}}\right)v$$

Where Q, K, and V represent Query, Key, and Value matrices derived from token configurations, and d_k is the scaling dimensionality of the key vectors (Ibrahim, 2026).

Concurrently, a Long Short-Term Memory (LSTM) network processes the sequence to model historical document dependencies. The finalized text embedding is generated by concatenating the pooled contextual output vector of the Transformer with the final hidden state of the LSTM layer:

$$H_{unified} = [H_{RoBERTa} \parallel H_{LSTM}]$$

This joint feature vector $\mathbf{H}_{unified}$ is routed to a fully connected feed-forward neural layer capped with a Softmax activation function to determine the ultimate probability of malicious threat intent ($\mathbf{P}_{threat}$).

IV. EXPERIMENTAL SETUP AND EVALUATION

To evaluate system resilience, the hybrid model was evaluated on balanced, heterogeneous open-source datasets composed of corporate logs, public messaging repositories, and known malicious registries.

A. *Benchmark Datasets*

Models were cross-validated utilizing extensive messaging collections, combining:

1) The Enron Corpus: Providing thousands of realistic, benign operational corporate emails (Mahendru & Pandit, 2024).
2) The Nazario Public Phishing Corpus: Supplying historical and modern actively targeted deceptive text variants (Safran, 2025; Mahendru & Pandit, 2024).
3) The SMS Spam Collection: Accommodating shorter, high-urgency SMS text instances (Mahendru & Pandit, 2024).

B. *Performance Metrics*

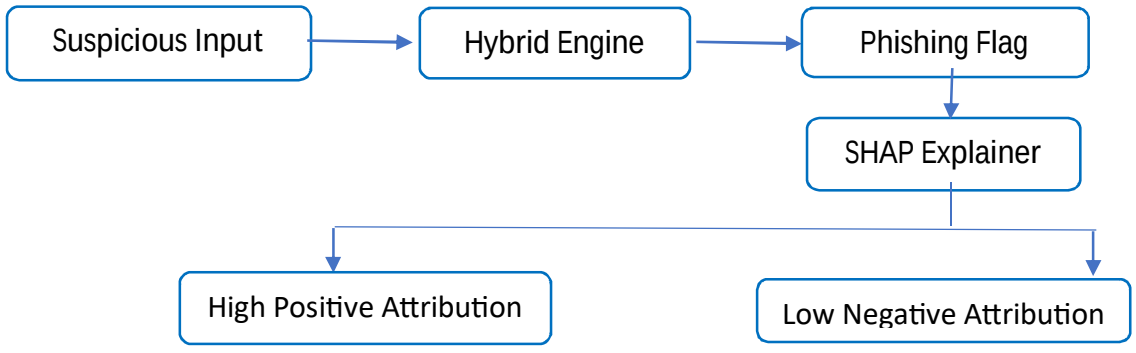
The system was assessed using Precision (P), Recall (R), and overall F_1\text{-Score}. In live operational deployments, special emphasis is placed on maximizing the recall rate to limit dangerous false-negative classifications (Ibrahim, 2026).

$$\textbf{Precision} = \frac{TP}{TP+FP}\ , \textbf{Recall} = \frac{TP}{TP+FN}\ , \mathbf{F_1} = 2. \frac{\textbf{Precision . Recall}}{\textbf{Precision + Recall}}$$

Empirical execution results demonstrate that the RoBERTa-LSTM fusion method yields a classification accuracy of **96.5%**, outperforming standard baseline machine learning implementations (such as Support Vector Machines and standalone LSTMs) by establishing deeper contextual awareness over long-form obfuscated semantics.

V. EXPLAINABLE AI (XAI) INTEGRATION

A critical operational drawback of deep neural architectures in enterprise cybersecurity monitoring is their inherently opaque "black-box" nature. Security analysts cannot easily isolate why a neural node flags a specific communication line. To remedy this limitation, our framework integrates **SHapley Additive exPlanations (SHAP)** (Alarfaj, 2026).



- "immediate verification" (+0.34)
- "restrict account access" (+0.41)
- "Warm regards" (0.05)

Fig 4: Working Steps of finding of suspicious activity

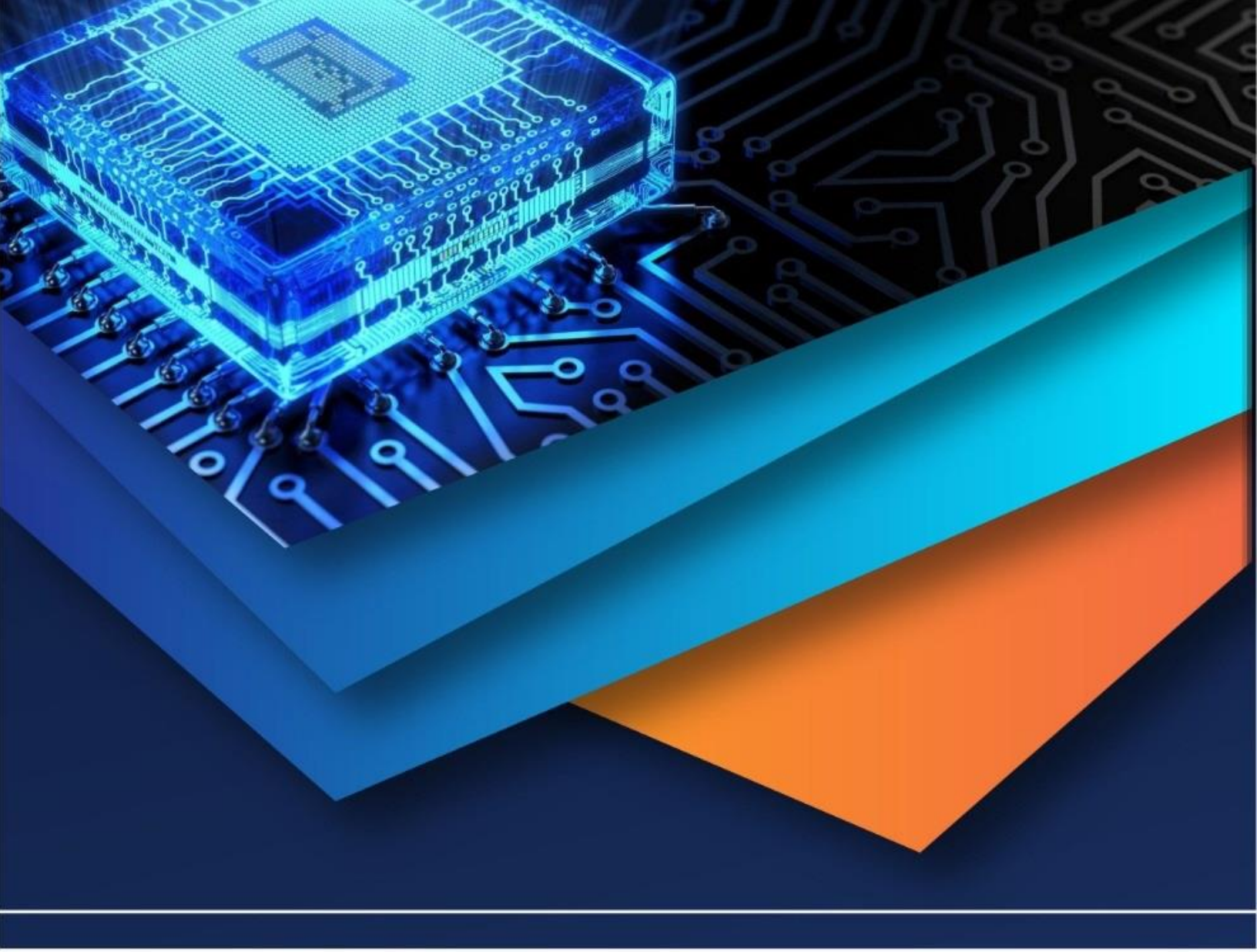
SHAP estimates game-theoretic Shapley values to identify individual feature attribution profiles for each text segment. By highlighting exactly which word choices or structural tokens drove the model's threat score, the system equips analysts with interpretable evidence. For instance, if an incoming email contains phrases like *"immediate verification required"* or *"restrict account access,"* SHAP visually maps these sub words to high positive threat attributions, allowing for efficient audit validation(De Nardin, 2025)..

VI. CONCLUSION AND FUTURE TRAJECTORIES

This chapter demonstrated the efficacy of applying advanced Natural Language Processing to combat modern AI-powered social engineering threats. By blending the deep semantic understanding of RoBERTa with the sequential tracking of LSTMs, the proposed framework provides high-accuracy detection capable of intercepting linguistically complex phishing attacks. Future development will focus on the threat of multi-modal evasion tactics - such as embedding malicious text directly inside image attachments or using variable font encodings Mitigating these zero-day vectors requires expanding current text pipelines into holistic, multi-modal systems capable of processing layout, imagery, and text features simultaneously.

REFERENCES

[1] Alarfaj, F. K. (2026). Clickbait detection in news headlines using RoBERTa-Large language model and deep embeddings. PMC, 12(780), 1–15. https://pmc.ncbi.nlm.nih.gov/articles/PMC12780135/
[2] Ali, A. A. (2025). Email Spam Detection: A Novel Hybrid Approach Using Machine and Deep Learning Techniques. International Network for Advanced Science and Technology, 12(2), 31–44.
[3] De Nardin, A. (2025). Deep Learning-Based Intrusion Detection Systems for Phishing Email Detection: A Short Survey. Computer Vision Foundation Open Access Workshops, 102–111.
[4] Ibrahim, M. (2026). Phishing Email Detection Using BERT and RoBERTa. MDPI Computer Sciences, 14(2), 46–59. https://www.mdpi.com/2079-3197/14/2/46
[5] Mahendru, S., & Pandit, T. (2024). SecureNet: A Comparative Study of DeBERTa and Large Language Models for Phishing Detection. Proceedings of the 2024 IEEE 7th International Conference on Big Data and Artificial Intelligence, 160–169. https://arxiv.org/pdf/2406.06663
[6] Safran, M. (2025). Phishing GNN: Phishing Email Detection Using Graph Attention Networks and Transformer-Based Feature Extraction. IEEE Xplore, 1109–1119.
[7] Zaman, T. A. S. (2025). Context-aware phishing url detection: harnessing the power of large language models. DergiPark Journal of Academic Networks, 4(2), 201–214.
[8] Uddin, R., Basit, A., Khan, Y., Shazib, MD S., & Hossain, S. (2025). BERT-Based Fake News Detection: A Transformer-Driven Approach for Misinformation Classification on Twitter. International Journal on Science and Technology, *16*(1), 1–12. https://doi.org/10.71097/ijsat.v16.i1.2023



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY