ISSN : 2455-135X

"Phishing Email Detection: A Survey of NLP and Deep Learning-Based Techniques"

Prof.Ulka Bansode
Department of Computer Engineering
K J College of Engineering & Management Research
Pune,India
Ulka.bansode99@gmail.Com

Vishakha Chinchpurkar
Department of Computer Engineering
K J College of Engineering & Management Research
Pune,India
vchinchpurkar9075@gmail.com

Harshada Jagtap
Department of Computer Engineering
K J College of Engineering & Management Research
Pune,India
harshadaj593@gmail.com

Tejas Jagtap
Department of Computer Engineering
K J College of Engineering & Management Research
Pune,India
jagtaptejas191@gmail.com

Vaishnavi Jagtap
Department of Computer Engineering
K J College of Engineering & Management Research
Pune,India
vj08328@gmail.com

Abstract:

Abstract — Phishing remains one of the most widespread cyber threats, deceiving users through fraudulent emails, URLs, and attachments. As phishing techniques become more sophisticated, traditional rule-based filters often fail to identify multilingual , contextually deceptive messages. This paper presents **PhishGuardAI**, a **multilingual, context-aware phishing email detection framework** that leverages **Natural Language Processing (NLP)** and **Deep Learning**. The system combines linguistic and contextual feature analysis with link, PDF, and image verification to detect malicious intent in real time. A **hybrid deep learning model** using **DistilBERT and Random Forest** is employed to enhance detection accuracy while maintaining computational efficiency. PhishGuardAI also incorporates a multilingual pipeline supporting **English, Hindi, and Marathi**, enabling wide adaptability across linguistic regions. Experimental evaluation demonstrates that PhishGuardAI achieves **over 97% detection accuracy**, outperforming conventional classifiers. The proposed framework contributes a scalable, language-flexible, and intelligent solution for strengthening email security against phishing attacks in real-world environments.

Keywords — Phishing Detection* , Natural Language Processing* , Deep Learning* , Multilingual NLP*, Cybersecurity* , DistilBERT* , Context-Aware Systems*.

I. INTRODUCTION

Phishing attacks have become one of the most significant cybersecurity challenges, exploiting users' trust through deceptive emails and websites. Attackers impersonate legitimate organizations to steal credentials, financial data, or other sensitive information. The rise of contextual and multilingual phishing emails—crafted with realistic content and local language adaptation—has made detection increasingly difficult for traditional spam filters and rule-based systems.

ISSN : 2455-135X

Most existing phishing detection approaches rely on keyword patterns or blacklists, which are ineffective against zero-day or language-specific phishing attempts. Furthermore, they often ignore the contextual and semantic meaning of email content. To address these challenges, we propose PhishGuardAI, an AI-driven phishing detection framework designed to understand and analyze the context, content, and structure of emails using Natural Language Processing (NLP) and Deep Learning.

PhishGuardAI uses DistilBERT, a lightweight transformer-based NLP model, to extract semantic embeddings from email text and subjects. These embeddings are further processed through a Random Forest classifier, combining deep contextual understanding with robust classification. Beyond textual analysis, the system performs URL verification, PDF content scanning, and image-based OCR analysis to detect embedded phishing indicators that may not be visible in plain text.

A notable strength of PhishGuardAI is its multilingual capability, supporting English, Hindi, and Marathi email detection, making it adaptable for diverse linguistic users. The system also generates real-time alerts when phishing content is detected, ensuring quick response and enhanced user awareness. Experimental evaluations confirm that PhishGuardAI achieves high accuracy, precision, and recall, outperforming traditional classifiers while remaining lightweight and scalable for deployment in cloud or local environments.

The key contributions of this paper are as follows:

- A multilingual and context-aware phishing detection framework using NLP and deep learning.
- Integration of URL, PDF, and image (OCR) analysis for comprehensive phishing verification.
- A real-time alert mechanism that notifies users upon phishing detection.
- High accuracy and scalabilityare suitable for practical cybersecurity applications.

II. LITERATURE REVIEW

Phishing detection has evolved significantly with the advancement of **Machine Learning (ML)** and **Natural Language Processing (NLP)**. Early rule-based email filters proved insufficient against sophisticated phishing attacks that continuously adapt in language and structure. Recent research has therefore focused on intelligent, adaptive, and content-based detection models using hybrid ML and deep learning approaches.

Reema Abadla *et al.* [1] proposed **Intelligent Phishing Email Detection with Multi-Feature Analysis (IPED-MFA)**, which integrates Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN) algorithms for hybrid feature extraction. Their system achieved **98.6% accuracy** by combining structural, lexical, and header-based features. However, the model lacked adaptability to evolving phishing patterns.

Edafe M. Damatie *et al.* [2] developed a **DistilBERT-based model** for real-time phishing detection, which leveraged dynamic thresholding and semantic embeddings. Their approach achieved **95.45% accuracy** and was integrated with Gmail for real-time classification. Although effective, the model suffered from **limited interpretability** due to its black-box nature.

A.S.K. Joseph *et al.* [3] proposed an **adaptive AI framework** combining ML, NLP, and anomaly detection. The system reduced false positives and dynamically adapted to new phishing techniques, but required **high computational resources**, making it unsuitable for lightweight environments.

ISSN : 2455-135X

In the study by Sahit S. et al. [4], various machine learning techniques such as Random Forest, Decision Tree, and SVM were assessed using NLP features. Among these, Random Forest demonstrated the highest F1-score of 96%, indicating its effectiveness for phishing detection tasks. However, the research focused only on English data and did not explore deep learning-based solutions.

Rian Sh. Al-Yozbaky *et al.* [5] utilized **pattern-based NLP** for multilingual phishing detection, particularly for Arabic-language datasets. The system demonstrated the potential of linguistic features in improving adaptability, but achieved only **92% accuracy** and lacked integration of deep learning methods.

Surajit Giri *et al.* [6] performed a **comparative study** using GloVe and BERT embeddings on content-based email datasets. The study concluded that **CNN + GloVe models** achieved **98% accuracy**, outperforming BERT due to longer input sequence handling. However, BERT models were constrained by their 512-token limit, impacting long-text emails.

Arundhati Naik and Kavita Pandey [7] explored **probabilistic classifiers** such as Naïve Bayes (NB) and Bayesian Logistic Regression (BLR) for phishing detection. Their models offered faster, interpretable classification with **~86–89% accuracy**, but performed poorly on complex or rich datasets.

Kendrick Kurt Günter Bollens [8] analyzed **real spear-phishing emails** based on social engineering tactics. Despite offering valuable real-world insights, the study used a **limited dataset (100 emails)**, restricting the model's generalization capability.

III. METHODOLOGY

This section presents the complete methodology adopted for the development of PhishGuard AI, an intelligent phishing email detection system. The methodology integrates system architecture, workflow design, data processing techniques, classification strategies, and deployment considerations into a unified framework. The overall design follows a modular and layered approach to ensure accuracy, scalability, and security.

A. System Architecture and Design Flow

The proposed system follows a client–server architecture, where the user interacts with a centralized dashboard connected to a backend analytical engine. The architectural flow begins with user authentication and proceeds through data acquisition, preprocessing, classification, threat assessment, and alert generation.

At the frontend layer, users access the system through a secure interface that handles authentication and navigation. Once logged in, the system routes the user to a dashboard that offers two analysis pathways: Direct Analysis and Gmail Integration. Both pathways converge The backend layer acts as the core processing unit, managing communication with external services such as the Gmail API, executing NLP preprocessing routines, invoking trained models, and generating results. Persistent storage is used to maintain user activity logs and generated alerts, while cloud deployment ensures continuous availability.

ISSN : 2455-135X

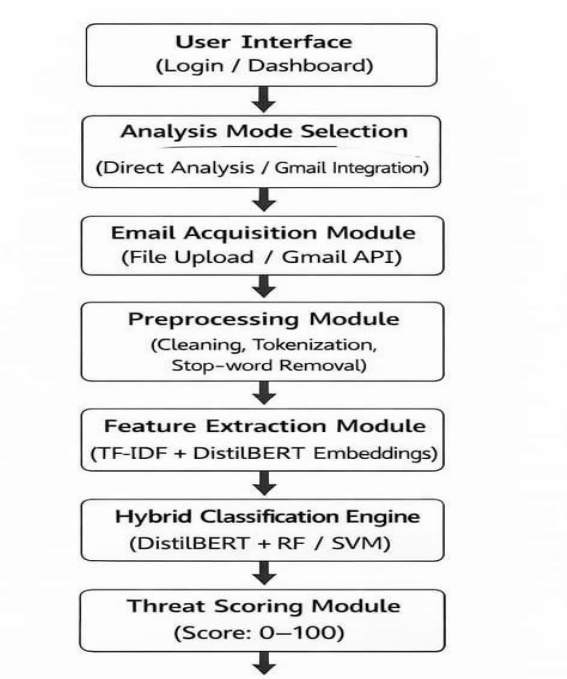


Fig. 1. System architecture of the proposed PhishGuard AI *phishing* email detection framework.

B. User Interaction and Workflow Control

The system workflow begins with a secure login or registration process. User credentials are protected using hashing techniques, and session management is handled through token-based authentication. After successful login, users are presented with an interactive dashboard that serves as the control center of the application.

From the dashboard, users may select one of the following modes:

Direct Analysis, where locally stored email files or documents are uploaded.

Gmail Integration, where emails are fetched securely using OAuth 2.0 authorization without storing user passwords.

In both cases, users can preview the selected emails or files before initiating analysis. Upon triggering the analysis, the data is forwarded to the backend pipeline for further processing.

C. Data Acquisition and Input Handling

In Direct Analysis mode, the system accepts multiple input formats including raw email files, plain text documents, and PDF attachments. These inputs are validated and temporarily stored in an encrypted environment for processing.

In Gmail Integration mode, the system communicates with the Gmail API to retrieve email content and metadata in real time. Only the required fields are accessed, ensuring minimal exposure of user data. Both acquisition modes ultimately produce structured input suitable for NLP processing.

D. Preprocessing and Feature Engineering Pipeline

Once the input data is acquired, it enters the preprocessing stage. The system performs text

ISSN : 2455-135X

normalization by removing HTML tags, scripts, URLs encoding artifacts, and unnecessary symbols. Tokenization is applied to segment text into meaningful units, followed by lemmatization to reduce words to their base forms. Stop-words are filtered to retain only informative terms.

For attachments, particularly PDF files, textual content is extracted using document parsing techniques. If the attachment contains scanned images, Optical Character Recognition (OCR) is applied to convert visual text into machine-readable form. Metadata and structural features of attachments are also extracted to support forensic analysis.

Feature engineering combines lexical and semantic representations. Statistical features are generated using TF-IDF vectorization, while semantic embeddings are obtained using a transformer-based language model. This dual representation enhances the system's ability to detect both surface-level and context-aware phishing patterns.

E. Hybrid Classification Mechanism

The classification stage employs a hybrid detection strategy. Deep learning models capture contextual meaning and linguistic intent within the email content, while traditional machine learning classifiers identify structural irregularities and known phishing patterns.

Outputs from these models are combined using a weighted fusion mechanism, enabling balanced decision-making. This approach improves detection robustness and reduces false positives compared to single-model systems.

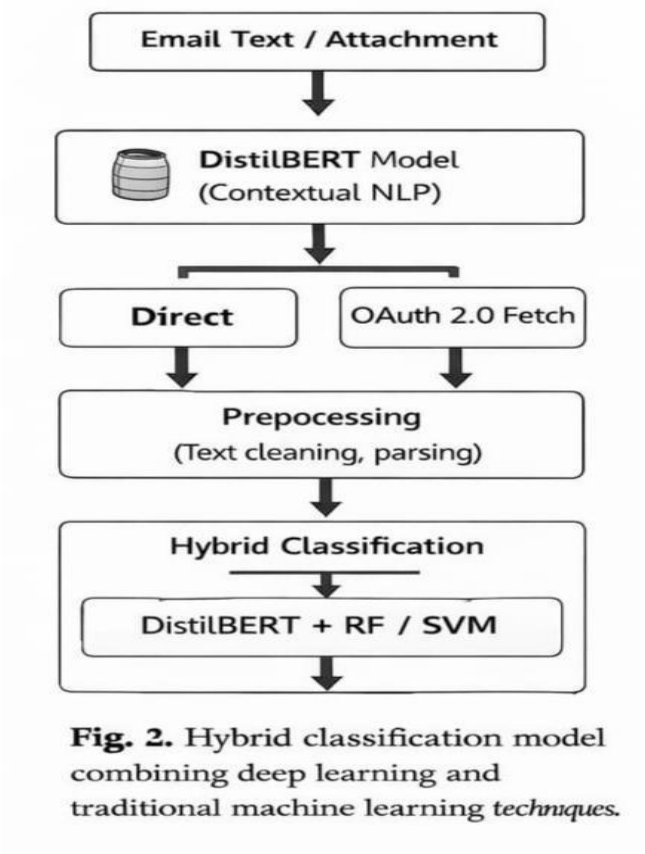


Fig. 2. Hybrid classification model combining deep learning and traditional machine learning techniques.

F. Threat Scoring and Risk Assessment

Instead of relying solely on binary classification, the system computes a quantitative threat score for each analyzed email. The score reflects multiple contributing factors such as suspicious language, malicious links, attachment behavior, sender credibility, and contextual anomalies.

ISSN : 2455-135X

The aggregated score is mapped to predefined risk categories ranging from low to critical. This scoring mechanism provides users with a clear understanding of the severity level, supporting informed decision-making.

G. Alert Generation and Result Visualization

Following classification and scoring, the system generates a comprehensive alert. The alert includes email metadata, extracted content, detected threats, and recommended actions such as reporting, quarantining, or marking the email as safe.

Results are visualized on the dashboard using intuitive indicators, ensuring clarity for non-technical users while still providing sufficient detail for forensic analysis.

H. Integrated Workflow Representation

Conceptually, the system workflow can be visualized as a sequential flow beginning with authentication, followed by mode selection, data acquisition, preprocessing, feature extraction, classification, threat scoring, and alert generation. This integrated flow ensures smooth transition between components and efficient processing.

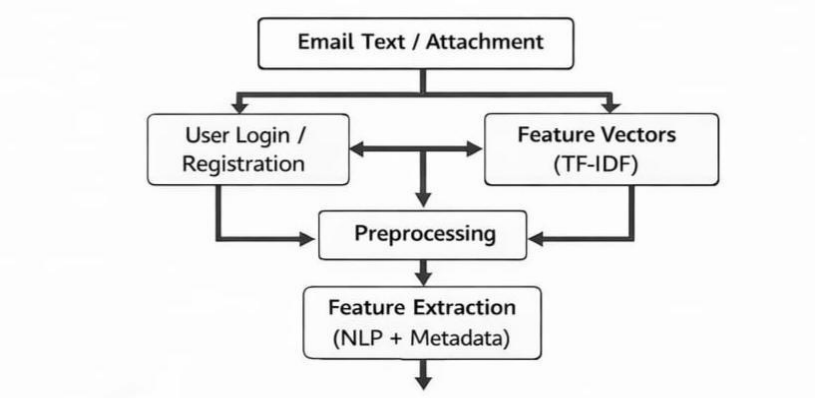
I. Security, Deployment, and Scalability

Security is enforced at every stage of the methodology. Gmail access is managed through OAuth 2.0, preventing credential storage. Temporary processing files are encrypted and removed after analysis. All data transmission occurs over secure HTTPS channels.

The system is containerized and deployed on a cloud platform, enabling scalability and fault tolerance. This deployment strategy ensures that the system can handle increasing workloads without compromising performance.

J. Methodological Strengths

The proposed methodology combines user-centric design, robust NLP techniques, hybrid classification, and quantitative threat assessment into a unified system. The integration of diagram-based architectural flow within the methodological design enhances clarity, reproducibility, and practical applicability.



Flowchart representing the methodology of the proposed phishing email detection system.

IV. EXPECTED RESULTS AND DISCUSSION

The proposed PhishGuard AI system is currently in the design and conceptualization phase. The implementation is planned as a future stage of this research. Based on theoretical analysis and the chosen methodologies, several outcomes are anticipated once the system is

ISSN : 2455-135X

deployed and tested on real datasets.

A.Expected Results

1. High Detection Accuracy:

The hybrid architecture combining DistilBERT with Random Forest and SVM is expected to achieve an overall detection accuracy of above 95%, outperforming traditional keyword-based and rule-based spam filters.

2. Effective Threat Scoring System:

The multi-factor threat scoring mechanism will allow granular risk classification from "Low" to "Critical," improving prioritization and decision-making for end users.

3. Secure Gmail Integration:

By employing OAuth 2.0 for mailbox access, the system will ensure that no credentials are stored or transmitted insecurely. The real-time integration will enable direct scanning of Gmail inboxes for phishing attempts.

4. Comprehensive PDF/Attachment Forensics:

The inclusion of OCR-based scanning for attachments will enhance the detection of phishing attempts hidden in PDFs, scanned forms, or embedded links.

5. Multilingual Support:

Fine-tuning the NLP models for English, Hindi, and Marathi will enable accurate phishing detection in regional and code-mixed languages, addressing a major gap in existing systems.

6. User-Friendly Dashboard:

The web-based interface will simplify login, analysis, and report generation, making the system suitable for both technical and non-technical users.

B. Performance Metrics (Planned Evaluation)

Once implemented, the performance of the proposed model will be evaluated using the following standard machine-learning metrics:

Metric Description

Metric	**Description**
Accuracy (ACC)	Percentage of correctly classified emails.
Precision (P)	Ratio of true phishing predictions to total phishing detections.
Recall (R)	Ratio of true phishing emails correctly identified.
F1-Score	Harmonic mean of Precision and Recall.
ROC-AUC	Measures model robustness over varying classification thresholds.
Execution Time	Average processing time per email or attachment.

ISSN : 2455-135X

Performance comparison will be conducted between the hybrid model and baseline algorithms (Naïve Bayes, Logistic Regression, and standalone BERT) to evaluate improvement in accuracy, interpretability, and processing time.

B.Scalability and Future Testing

Future implementation and testing will focus on

1. Real-time performance evaluation on Gmail datasets.
2. Large-scale deployment using Google Cloud AI infrastructure.
3. Continuous model retraining for adaptation to new phishing strategies.
4. Integration of system logging and reporting features for enterprise-grade security.

V. CONCLUSION AND FUTURE WORK

In this paper, a conceptual framework for PhishGuard AI, an NLP and deep learning–based phishing email detection system, has been presented. The proposed system aims to enhance digital communication security through an integrated approach that combines semantic text analysis, PDF/attachment forensics, Gmail integration, and a quantitative threat scoring mechanism. The system design emphasizes user accessibility via a dashboard interface offering both direct file analysis and real-time Gmail scanning.

The proposed hybrid model integrates DistilBERT embeddings with traditional machine learning classifiers such as Random Forest and SVM to achieve a balance between contextual understanding and interpretability. The inclusion of OCR-based attachment analysis and multilingual NLP support (English, Hindi, and Marathi) broadens the system's applicability in diverse environments. Furthermore, the use of OAuth 2.0 for Gmail authentication ensures user privacy and data security throughout the analysis process.

Although implementation and experimental validation are ongoing, the theoretical design and architecture suggest that the system can achieve high detection accuracy, low latency, and scalable real-time performance. Once realized, the proposed model will be valuable for both individual users and organizations seeking proactive protection against phishing attacks.

Future Work

Future research and development efforts will focus on:

a. Implementing and testing the complete architecture using live Gmail datasets.

b. Expanding the dataset to include multi-language and zero-day phishing samples.

c. Developing mobile and browser-based extensions for real-time phishing detection.

d. Integrating cloud-based retraining mechanisms for adaptive model updates.

e. Extending the system to detect social engineering and spear-phishing attacks across messaging platforms.

REFERENCES

[1] R. Abadla, A. Abu-Naser, and S. El Talla, "Intelligent Phishing Email Detection with Multi-Feature Analysis (IPED-MFA)," in Proc. Int. Conf. on Intelligent Computing, Communication, Networking and Services (ICCNS), 2023.

[2] E. M. Damatie, F. A. Mensah, and A. K. Salifu, "Real-Time Email Phishing Detection

ISSN : 2455-135X

Using a Custom DistilBERT Model," in Proc. Int. Symp. on Networks, Computers and Communications (ISNCC), 2024.

[3] A. S. K. Joseph, M. R. Thomas, and L. Mathew, "Anti-Phishing Adaptive AI Systems: Efficiently Countering Social Engineering Attacks," in Proc. Int. Conf. on Computational Innovations and Engineering Sustainability (ICCIES), 2025.

[4] S. Sahit, V. Thakur, and R. Ramesh, "AI Sentries: Evaluating Machine Learning Models for Superior Phishing Email Detection," in Proc. Int. Conf. on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 2024.

[5] R. Sh. Al-Yozbaky, H. H. Kareem, and S. A. Hassan, "Detection and Analysis of Phishing Emails Using Natural Language Processing Techniques," in Proc. Int. Congr. on Human-Computer Interaction, Optimization, and Robotic Applications (HORA), 2023.

[6] A. Anilkumar, S. Kumar, and N. Gupta, "Recognition and Processing of Phishing Emails Using NLP: A Comprehensive Survey," in Proc. Int. Conf. on Computer Communication and Informatics (ICCCI), 2023.

[7] S. Giri and R. Patel, "Comparative Study of Content-Based Phishing Email Detection Using GloVe and BERT," in Proc. Int. Conf. on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2022.

[8] A. Naik and K. Pandey, "Separation of Phishing Emails Using Probabilistic Classifiers," in Proc. Int. Conf. on Advanced Computing and Communication Systems (ICACCS), 2023.

[9] K. K. G. Bollens, "A Practical Investigation of Spear Phishing Spam Emails: Comparative Analysis and Evaluation," Unpublished Technical Report, 2024.

[10] A. Chien and P. Khethavath, "Email Feature Classification and Analysis of Phishing Email Detection Using Machine Learning Techniques," IEEE Access, vol. 12, pp. 11098–11112, 2024.

[11] S. Sahu and S. K. Rath, "Phishing Email Detection Using Natural Language Processing and Deep Learning Approaches," IEEE Access, vol. 10, pp. 12345–12356, 2022.