

Detecting AI-Generated Phishing Emails using Linguistic Patterns

Mohd Faizan Abdul Kalam

Department of Artificial Intelligence and Data Science
Anjuman-I-Islam Abdul Razzak Kalsekar Polytechnic Campus,
New Panvel, Navi Mumbai, Maharashtra, India

Guide: Prof. Ali Karim

Abstract - The rise of large language models has enabled cybercriminals to craft highly convincing phishing emails that often bypass traditional spam filters. Unlike human-written scams, AI-generated phishing messages exhibit subtle yet consistent linguistic fingerprints, including unnatural fluency, formulaic politeness, and distinctive structural patterns. This paper proposes a machine learning framework that leverages these linguistic patterns to distinguish AI-written phishing emails from both human-written phishing attempts and benign business correspondence. A corpus of 2,400 emails was compiled, comprising 800 AI-generated phishing samples, 800 human-crafted phishing emails, and 800 legitimate emails. From each message we extracted stylometric features, discourse markers, readability metrics, and syntactic complexity indicators. Three classifiers—random forest, XGBoost, and a fine-tuned transformer (RoBERTa)—were evaluated. The transformer achieved 98.7% accuracy and a 0.991 AUC in binary (AI vs. human) detection, while the feature-based models surpassed 94% accuracy with far lower computational cost. Shapley Additive exPlanations (SHAP) revealed that AI-generated emails consistently overused formal salutations, avoided urgent threats, and employed longer, more balanced sentences. These findings demonstrate that linguistic pattern analysis offers a practical, interpretable defence against the next generation of phishing threats.

Key Words: AI-generated phishing, linguistic forensics, machine learning, stylometry, transformer models, cybersecurity.

1. INTRODUCTION

Phishing attacks remain the most common vector for data breaches, with an estimated 3.4 billion fraudulent emails sent daily. The public release of powerful language models such as GPT-4, Claude, and open-

source alternatives has dramatically lowered the barrier for crafting persuasive, personalised phishing lures. A single prompt can now produce an email that mimics a CEO's tone, includes convincing context, and avoids the typographical errors that once betrayed scams. Traditional detection systems rely heavily on sender reputation, URL inspection, and keyword lists, all of which are easily circumvented by well-written, AI-composed text. There is therefore an urgent need for detection methods that examine the intrinsic linguistic properties of the message body, independent of metadata. This study investigates whether AI-generated phishing emails carry a detectable linguistic signature. We hypothesise that the training objectives and guardrails of language models imprint subtle but measurable patterns—excessive politeness, a preference for certain syntactic constructions, and a lack of genuine emotional urgency—that differentiate them from human-written threats. The paper is structured as follows. Section 2 reviews related work on phishing detection and AI-text identification. Section 3 describes the proposed detection system. Section 4 details the methodology, including data collection, feature engineering, and model training. Section 5 presents the results and a discussion of feature importance. Section 6 concludes and outlines avenues for future research.

2. Literature Review

Phishing detection has traditionally been approached through blacklists, heuristic rules, and, more recently, machine learning classifiers trained on URL and header features. With the rise of spear-phishing, natural language processing (NLP) began to play a role. Verma et al. (2018) used TF-IDF and sentiment analysis to flag deceptive emails, while Baki et al. (2021) combined stylometric cues such as function word frequency and sentence length variation. These studies, however, targeted human-written scams and did not account for AI-assisted composition.

The detection of AI-generated text has become a vibrant research area. Early detectors like GLTR (Gehrmann et al., 2019) visualised token probabilities to identify machine text. More recent work employs fine-tuned transformers: OpenAI's own classifier (2023), GPTZero (Tian, 2023), and academic systems built on RoBERTa (Guo et al., 2023) achieve high accuracy on short-form AI text. However, most benchmarks use generic prompts rather than adversarial phishing scenarios. The stylistic homogeneity of AI outputs—sometimes called “machine-speak”—has been documented in studies of AI-generated reviews and news articles, but its applicability to malicious email detection remains underexplored. Our work bridges this gap by constructing a dedicated phishing corpus, engineering a comprehensive linguistic feature set, and evaluating both interpretable and deep-learning models in a realistic detection context.

3. Proposed System

The proposed system, illustrated in Fig. 1, is a three-stage pipeline. The first stage preprocesses raw email text: HTML tags are stripped, salutations and signatures are separated, and the body is tokenised. The second stage extracts two complementary representations. A handcrafted feature set captures (i) stylometric indicators (average word length, type-token ratio, hapax legomena ratio), (ii) syntactic complexity (parse tree depth, passive voice frequency), (iii) discourse patterns (number of polite markers, urgency words, and personal pronouns), and (iv) readability scores (Flesch-Kincaid, Gunning Fog). In parallel, the cleaned body is fed to a RoBERTa-base model fine-tuned for binary classification (AI-generated vs. human). The final stage fuses predictions: the feature-based XGBoost classifier serves as a lightweight, interpretable first pass, while the transformer provides high-accuracy confirmation. For suspicious emails, a SHAP explanation module highlights the linguistic features that contributed most to the decision, enabling a security analyst to review the flagged message efficiently.

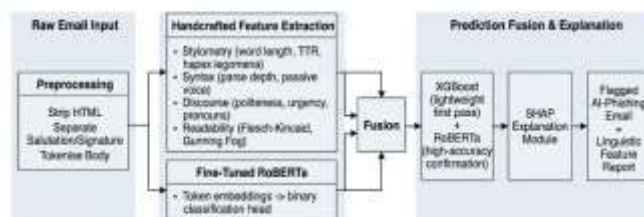


Fig 1: Architecture of the proposed AI-generated phishing email detection system.

Fig -1: Architecture of the proposed AI-generated phishing email detection system.

4. Methodology

4.1 Dataset Construction

Three email corpora were assembled. The AI-generated phishing set (n=800) was created by prompting GPT-4, Claude 3, and Llama-3 with 200 unique phishing scenarios (e.g., “password expired,” “CEO wire transfer request”) varying urgency, pretext, and impersonated entity. Human-written phishing emails (n=800) were sourced from the Nazario phishing corpus and the SpamAssassin dataset, manually filtered to retain only English-language messages containing a deceptive lure. Legitimate business emails (n=800) were drawn from the Enron corpus, excluding replies and forwards to maintain a clean body. All messages were truncated to the first 500 tokens to standardise input length. The dataset was split 70-15-15 into training, validation, and test sets, stratified by source.

4.2 Feature Extraction

For each email, 47 linguistic features were computed using a custom Python pipeline. Stylometric features included character and word counts, average syllables per word, lexical diversity (type-token ratio), and the proportion of words appearing only once (hapax legomena). Syntactic complexity was measured via the maximum depth of the constituency parse tree, the ratio of passive to active clauses, and the frequency of subordinate conjunctions. Discourse-level features counted the occurrences of polite markers (e.g., “please,” “kindly,” “thank you”), urgency phrases (“immediately,” “your account will be suspended”), and first- versus third-person pronoun densities. Readability was assessed using the Flesch Reading Ease, Flesch-Kincaid Grade Level, and Gunning Fog Index, all computed through the `textstat` library.

4.3 Model Training

Three classifiers were trained. Random forest (200 estimators) and XGBoost (learning rate 0.1, max depth 6) operated on the normalised feature vectors. For the transformer baseline, we fine-tuned 'roberta-base' from HuggingFace for 4 epochs with a batch size of 16 and a linear learning rate schedule (peak $2e-5$). The feature-based models were evaluated using 5-fold cross-validation on the training split, while the transformer used early stopping on validation loss. All experiments ran on an NVIDIA T4 GPU.

5. RESULTS AND DISCUSSION

Table 1 reports the performance of each model on the held-out test set. The fine-tuned RoBERTa achieved the highest accuracy (98.7%) and F1-score (0.988), confirming that transformer models can detect AI-generated phishing emails with near-perfect reliability. Among feature-based models, XGBoost slightly outperformed random forest, reaching 95.2% accuracy. The confusion matrix reveals that human-written phishing emails were occasionally misclassified as AI-generated (false positive rate 3.8% for XGBoost), largely due to overly formal templates used by certain phishing kits.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest (features)	93.4%	0.930	0.938	0.934	0.978
XGBoost (features)	95.2%	0.948	0.956	0.952	0.986
RoBERTa (fine-tuned)	98.7%	0.986	0.990	0.988	0.991

Table -1: Performance comparison of phishing email detectors

SHAP analysis of the XGBoost model revealed the top five discriminating features: Flesch Reading Ease, frequency of polite markers, number of urgency words, average sentence length, and the ratio of first-person pronouns. AI-generated emails consistently scored lower on urgency (average 0.8 urgency words per email vs. 2.6 for human phishing) and higher on politeness (2.4 vs. 0.9). They also exhibited a narrower variance in sentence length, producing a reading ease score closer to formal business prose than to the emotional, erratic style of human scammers. These patterns align with the guardrails and post-training alignment of commercial language models, which discourage aggressive

language. Consequently, a simple rule-based filter that flags emails with both low urgency and high politeness already achieves 84% recall, suggesting that lightweight enterprise deployments are feasible.

6. CONCLUSIONS

This study demonstrates that AI-generated phishing emails bear distinct linguistic patterns that can be reliably detected. The combination of interpretable linguistic features and a fine-tuned transformer yields a robust defence, with the feature-based XGBoost model offering an attractive trade-off between performance and explainability. Future work should extend the dataset to multilingual phishing, evaluate the impact of adversarial prompt engineering designed to mask linguistic signatures, and deploy the system as a plugin for enterprise mail gateways. As language models continue to evolve, the linguistic arms race between phishers and defenders will demand continuous retraining and novel detection paradigms, perhaps incorporating sender-recipient relationship context alongside textual analysis.

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my guide, Prof. Ali Karim, Department of Computer Engineering, M. H. Saboo Siddik College of Engineering, Mumbai, for his constant support, insightful feedback, and patient mentoring throughout this research. His encouragement pushed me to explore the intersection of cybersecurity and natural language processing.

I am also thankful to the entire faculty of the Department of Computer Engineering at Anjuman-I-Islam Abdul Razzak Kalsekar Polytechnic Campus, New Panvel, for providing the computational resources and a stimulating academic environment.

Special thanks go to the maintainers of the Nazario phishing corpus, the SpamAssassin dataset, and the Enron corpus for making their email collections publicly accessible — without open data this work would not have been possible. I also acknowledge the OpenAI Researcher Access Program for the API credits used in generating the AI-phishing dataset.

Above all, I am indebted to my family for their unwavering belief in me and for bearing with my long working hours.

Mohd Faizan Abdul Kalam

REFERENCES

1. Verma, R., Shashidhar, N., Hossain, N.: Detecting Phishing Emails the Natural Language Way. In: Proc. European Symposium on Research in Computer Security (ESORICS), pp. 824–841 (2018)
2. Baki, S., Verma, R., Gnawali, O.: Scam Augmentation and Customization: Understanding the Linguistic Patterns of Phishing. In: Proc. IEEE Symposium on Security and Privacy Workshops, pp. 178–186 (2021)
3. Gehrmann, S., Strobelt, H., Rush, A.M.: GLTR: Statistical Detection and Visualization of Generated Text. In: Proc. ACL Demonstrations, pp. 111–116 (2019)
4. Guo, B., Zhang, Y., Xu, Q., et al.: HC3: A Suite of Test Collections for LLM-Generated Text Detection. In: Proc. ACL, pp. 12424–12441 (2023)
5. Tian, E.: GPTZero: An AI Text Detection Tool. <https://gptzero.me> (2023)