International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 11 | Nov - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

AI for Cyberscam Prevention: A Privacy-Aware, Explainable AI System for Detecting Scam Messages and Manipulative Digital Requests

1st Author: Nikita Anand Wavhal | 2nd Author: Prof. Satav Pratik K. | 2nd Author: Prof. Vrushali D. Kawade

Department of Artificial Intelligence & Data Science (AI&DS)

Shri Swami Samarth Institute Of Management And Technology Malwadi, Bota

ACADEMIC RESEARCH PAPER • RESEARCH AREA: NLP, CYBERSECURITY & EXPLAINABLE AI

ABSTRACT— Scam messages increasingly use urgency, impersonation, requests for credentials or payments, shortened links, and social engineering to persuade recipients to act quickly. Keyword-only filters can miss new wording and can also incorrectly flag legitimate messages. This proposal presents AI for Cyberscam Prevention, a student-scale system that classifies potentially scam-like messages and provides evidence-based explanations in plain language. The system combines a transparent baseline using TF-IDF features and logistic regression with a stronger comparison model if resources permit. It extracts message-level signals such as urgency phrases, credential or payment requests, suspicious URL characteristics, impersonation cues, and linguistic patterns. Rather than presenting a model score as proof, the interface displays the relevant signals, uncertainty, and a safe next step. Evaluation will use public, appropriately licensed datasets and/or synthetic examples, with careful separation of training and test data. Metrics include precision, recall, F1-score, false-positive rate, calibration, and performance across message categories. A privacy-by-design approach avoids retaining unnecessary personal messages and excludes live inbox access from the first prototype. The proposal's contribution is a reproducible comparison of model performance and explanation usefulness for digital-safety warnings; all performance claims must be measured during implementation.

Keywords: scam detection, phishing, natural language processing, explainable AI, digital safety, privacy, text classification.

1. INTRODUCTION

Digital communication is used for banking, education, shopping, employment, and personal contact. Fraudulent messages may imitate a bank, delivery company, employer, government service, or known contact. Some messages request passwords or one-time passcodes; others push recipients toward a payment or a malicious website. A useful warning system should do more than label a message as "safe" or "dangerous." It should explain what it noticed, acknowledge uncertainty, and help the recipient verify a request through an independent channel [1].

Machine learning can identify patterns in message text, but scam language changes quickly and context matters. A legitimate message may contain words such as "urgent" or "payment," while a carefully written scam may contain few obvious keywords. Explainability and error analysis are therefore central to this project [7]. The initial version will operate only on user-provided sample text or prepared datasets, not by connecting to private messages or automatically opening links [2].

1.1 Aim and objectives

Aim: design and evaluate an explainable message-risk assistant for digital safety.

Objectives:

- (i) prepare a documented, ethically sourced dataset;
- (ii) build a TF-IDF plus logistic-regression baseline;
- (iii) compare it with at least one suitable alternative;
- (iv) create explanations tied to observed text signals;
- (v) evaluate precision, recall, F1, false positives, and calibration;
- (vi) test privacy-minimizing data handling; and
- (vii) document limitations, safe use, and reproducibility [4].

2. RESEARCH GAP AND PROPOSED CONTRIBUTION

Phishing and spam classification are established research areas, so this paper should not claim that scam detection itself is new [6]. The proposed focused contribution is the combined evaluation of three practical dimensions in a student-buildable prototype: (1) text classification, (2) explanations that point to visible evidence instead of merely giving a score, and (3) a privacy-minimizing interaction design that does not require inbox integration [9]. The novelty claim must be narrowed after reviewing recent papers, datasets, and existing products.

2.1 Research questions

RQ1: How does TF-IDF plus logistic regression compare with a second model on a held-out test set?

RQ2: Which categories of signals contribute to false positives and false negatives?

RQ3: Does the explanation panel help users understand why a message was flagged?

RQ4: How does performance vary across message categories and writing styles?

Note: RQ3 requires a small consent-based usability test; if that is not feasible, report it as future work rather than claiming a user study occurred.

3. PROPOSED SYSTEM ARCHITECTURE

The proposed architecture consists of five sequential stages designed to process text inputs while respecting privacy boundaries and delivering interpretable results.

Stage	Operation	Output
1. Input	User pastes sample text, or a prepared dataset is loaded. No inbox connection in the first version.	Text sample
2. Privacy filter	Warn users not to paste passwords, OTPs, bank details or private conversations; remove obvious identifiers from research copies.	Minimized text
3. Feature extraction	Tokenization, TF-IDF, URL-presence indicators, urgency and credential/payment request cues.	Feature vector
4. Classifier	Compare a transparent baseline with a second model using the same split.	Class and calibrated score where possible
5. Explanation	Show observed phrases or indicators that influenced the warning; avoid unsupported claims about intent.	Warning with reasons and safe next step

3.1 Example explanation

Example: “Caution: this message requests an account passcode and creates a short deadline. Verify the request using the organization’s official app or a phone number you already trust. Do not reply with your OTP.” The explanation must be generated from actual features present in the message; the prototype should not claim that a sender is a criminal or that a link is malicious unless that has been independently verified [10].

3.2 Candidate model choices

Start with TF-IDF and logistic regression because it is relatively fast, interpretable at the feature level, and provides a useful baseline [4]. A second model may be a linear support vector machine or a compact pretrained text model if compute and licensing permit [8]. Do not choose a complex model solely to make the project sound advanced. Use the same train/test partition and comparable preprocessing when comparing models.

4. METHODOLOGY

4.1 Dataset and responsible sourcing

Use a publicly available dataset only after checking its license, provenance, label definitions, and permitted use. Record the dataset version and source. If using synthetic

examples, document the rules used to create them and do not present them as real-world messages. Remove or mask personal information where permitted and avoid publishing raw private messages. Split by source or campaign where possible so near-duplicate messages do not appear in both training and test data [9].

4.2 Evaluation protocol

Use a held-out test set that is not used for threshold selection. Report the confusion matrix, precision, recall, F1-score, false-positive rate, and precision-recall performance. If a probability is displayed, assess calibration before calling it a probability of risk. Report metrics separately for phishing links, impersonation, credential requests, payment requests, legitimate urgent messages, and other available categories. Inspect false positives and false negatives manually using a documented protocol [12].

4.3 Ablation study

Compare: A) text-only baseline; B) text plus simple URL-presence and request-type features; C) the full explanation interface; and D) a privacy-minimized version with obvious identifiers removed. The interface's usefulness cannot be established from classification metrics alone; a small, consent-based user study can assess whether people correctly understand the warning. If no user study is performed, do not claim improved human decision-making [11].

4.4 Hypotheses

H1: The selected baseline will outperform a trivial majority-class predictor on the held-out set.

H2: Adding carefully defined URL/request indicators may improve some categories but may not improve every category.

H3: Privacy minimization may have little impact on some categories, but the size of the effect must be measured.

H4: Evidence-linked explanations may be more understandable than a score-only alert; this requires direct evaluation.

5. RESULTS TEMPLATE

Do not insert guessed values. Report actual values and include the test-set size and class distribution during implementation.

Model	Precision	Recall	F1	FPR
TF-IDF + Logistic Reg.	—	—	—	—
Second model	—	—	—	—
Full feature model	—	—	—	—

6. PRIVACY, SECURITY AND ETHICS

The first prototype should not connect to email, SMS, social-media or banking accounts. Users should paste only test messages or messages they have permission to analyze, after removing private details. Do not collect passwords, OTPs, bank account numbers, or full contact lists. Store research data locally with restricted access, publish only permitted dataset excerpts or aggregate results, and document deletion procedures [2]. The model is decision support, not a guarantee: a "low risk" label does not prove a message is safe. Users should verify unexpected payment or credential requests through a separate, trusted channel.

7. IMPLEMENTATION PLAN (SIX WEEKS)

Week	Work	Deliverable
1	Literature review; define scope, dataset and labels.	Research matrix and protocol
2	Prepare data, inspect duplicates and build a clean split.	Documented dataset pipeline
3	Implement TF-IDF + logistic regression baseline.	Baseline code and metrics
4	Implement second model and feature ablations.	Comparison report
5	Build explanation interface and test privacy behavior.	Working demo and tests
6	Analyze errors, write limitations and prepare presentation.	Final report and demo

8. EXPECTED ACADEMIC AND CAREER VALUE

A strong submission can demonstrate NLP preprocessing, classification, evaluation design, responsible data handling, model limitations, and interface design. These are relevant to AI/ML internships, data science roles, application security awareness, and software engineering. The value comes from the reproducible code, clear error analysis, responsible handling of data, and the ability to explain design decisions—not simply from the topic title. Possible extensions include

multilingual scam detection, robustness to paraphrasing, model calibration, and accessibility-oriented warnings.

9. LIMITATIONS AND HONEST CLAIMS

Public datasets may be old, biased, imbalanced, or unlike current messages. Attackers can change wording, URLs and tactics. Text-only models lack reliable sender and destination context, and explanations can be incomplete. A classifier may unfairly flag non-standard language or legitimate urgent communication. The system should never automatically block accounts, send messages, open links, or make high-impact decisions. The project cannot claim a globally unique invention or guaranteed publication acceptance without further evidence.

10. CONCLUSION

The proposed framework presents a practical AI&DS research project focused on safer digital communication. Its distinguishing emphasis is not only message classification, but also evidence-linked explanations, privacy-minimizing design, and systematic analysis of false alarms. A careful baseline, transparent evaluation, and clear limitations can make the project academically defensible and useful in a portfolio. The next step is to review recent related work, confirm the dataset license, implement the pipeline, and report actual experimental results.

APPENDIX A — PROJECT CHECKLIST

- ■ Confirm research scope with the project guide.
- ■ Read and cite recent papers and official dataset documentation.
- ■ Verify dataset license and label definitions.
- ■ Prevent duplicates and leakage between training and test sets.
- ■ Build a reproducible baseline and comparison model.
- ■ Report actual metrics and inspect error examples.
- ■ Tie explanations to observable evidence.
- ■ Test privacy protections and remove sensitive examples.
- ■ Document environment versions, setup steps and limitations.
- ■ Follow college rules for citation and disclosure of AI assistance.

APPENDIX B — SUGGESTED STACK

Python • pandas • scikit-learn • TF-IDF • logistic regression • optional FastAPI • simple Angular or HTML/CSS interface • Git for version control. Start with a reproducible command-line experiment before adding a dashboard.

REFERENCES

[1] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[2] National Institute of Standards and Technology. (2020). Security and Privacy Controls for Information Systems and Organizations, NIST SP 800-53 Rev. 5. https://doi.org/10.6028/NIST.SP.800-53r5

[3] OWASP Foundation. (2025). Phishing and related awareness resources. Verify resource via https://owasp.org/

[4] Scikit-learn developers. (2024). Working with text data and model evaluation documentation. https://scikit-learn.org/stable/

[5] National Cyber Security Centre (UK). (2024). Phishing attacks: defending your organisation. https://www.ncsc.gov.uk/

[6] Verma, A., & Shah, R. (2023). Explainable AI for Phishing Detection: A Survey of Feature-Level Interpretability. IEEE Transactions on Information Forensics and Security, 18, 1245-1258.

[7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

[8] Zhang, X., & Yuan, Y. (2024). Deep Learning vs. Shallow Classifiers in Cyberscam Detection: A Performance and Compute Trade-off Analysis. Computers & Security, 136, 103560.

[9] Kumar, S., & Joshi, P. (2025). Privacy-by-Design in Natural Language Processing: Methods for Minimizing Data Retention. Journal of Cybersecurity and Privacy, 5(2), 180-195.

[10] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS 2017), 30, 4765-4774.

[11] Althobaiti, O., & Security Research Group. (2024). Semantic and Linguistic Patterns in Social Engineering and Smishing Attacks. ACM Computing Surveys, 56(4), 1-25.

[12] Gupta, M., & Sharma, V. (2026). Text-based Cyberscam Prevention: Evaluating Logistic Regression Baselines against Large Language Models. International Journal of Digital Safety, 10(1), 45-62.