EXPLAINABLE AI FOR PHISHING DETECTION IN MULTILINGUAL AND LOW-RESOURCE CONTEXTS: A HUMAN-CENTERED CYBERSECURITY APPROACH

Abdukhalimov Usmonbek Eshberdievich
Presidential School in Termez

Annotation
Phishing attacks remain one of the most pervasive and effective forms of cybercrime worldwide. By exploiting human trust through deceptive emails, messages, and websites, attackers successfully steal sensitive information such as login credentials, financial data, and personal identities. Despite advances in technical defenses, phishing continues to evolve, becoming more sophisticated, personalized, and linguistically diverse.

Keywords: credentials, financial data, and personal identities, sophisticated, personalized, and linguistically diverse.

In recent years, Artificial Intelligence (AI) and Machine Learning (ML) techniques have significantly improved phishing detection systems. However, most existing solutions focus on high-resource languages such as English and operate as “black-box” models, offering little insight into how decisions are made. This lack of transparency poses serious challenges for trust, usability, and ethical deployment—especially in multilingual and low-resource contexts where linguistic diversity, limited datasets, and cultural nuances complicate detection.
Explainable Artificial Intelligence (XAI) has emerged as a promising solution to these challenges. XAI aims to make AI decision-making processes understandable to humans, enabling users and security professionals to trust, validate, and effectively interact with AI systems. This article explores the role of explainable AI in phishing detection, with a particular focus on multilingual and low-resource environments, through a human-centered cybersecurity perspective.

Phishing as a Human-Centered Cybersecurity Problem

Phishing is not merely a technical issue; it is fundamentally a human-centered cybersecurity problem. Unlike malware that exploits software vulnerabilities, phishing targets human cognition, emotions, and behavior. Attackers use urgency, authority, fear, or curiosity to manipulate users into taking harmful actions.
Human vulnerability is further amplified in multilingual contexts. Users may receive phishing messages in their native or second language, where grammatical errors, cultural references, or translation issues make detection more difficult. In low-resource languages, users often lack

access to reliable cybersecurity education, localized detection tools, or language-aware defenses.

Traditional phishing detection systems focus primarily on technical indicators such as URLs, sender domains, and lexical patterns. While effective to some extent, these approaches often fail to address the human factors involved in decision-making. A human-centered approach emphasizes usability, transparency, and user empowerment—areas where explainable AI plays a critical role.

AI-Based Phishing Detection: Current Approaches

Modern phishing detection systems increasingly rely on AI and ML models, including:

- Supervised learning classifiers (e.g., SVM, Random Forest, Neural Networks)
- Deep learning models (e.g., CNNs, RNNs, Transformers)
- Natural Language Processing (NLP) techniques for text analysis

These models analyze features such as email content, metadata, URLs, and user behavior to classify messages as phishing or legitimate. Deep learning models, in particular, achieve high accuracy but suffer from a lack of interpretability.

Most AI-based phishing detectors are trained on large datasets in high-resource languages, especially English. As a result, their performance drops significantly when applied to low-resource languages or multilingual environments. This bias not only reduces effectiveness but also increases false positives and false negatives, undermining user trust.

Challenges in Multilingual and Low-Resource Contexts

Data Scarcity

Low-resource languages often lack large, labeled phishing datasets. This limits the ability to train robust AI models and increases dependence on transfer learning or synthetic data generation. However, models trained on other languages may fail to capture language-specific phishing cues.

Linguistic and Cultural Diversity

Phishing messages vary across cultures and languages. Certain persuasive strategies, idioms, or honorifics may be specific to a language or region. AI systems that ignore these nuances risk misclassification.

Limited User Awareness

In many low-resource contexts, users have limited exposure to cybersecurity education. Black-box AI systems that provide no explanation leave users confused and less likely to trust warnings, reducing the effectiveness of detection systems.

Explainable AI: Concept and Importance

Explainable AI refers to methods and techniques that make AI systems' decisions understandable to humans. Instead of simply labeling a message as "phishing," an XAI system explains *why* the message was classified as such.

Common XAI techniques include:

- Feature importance analysis
- Rule-based explanations
- Attention visualization in neural networks
- Post-hoc explanation tools such as LIME and SHAP

In phishing detection, explainability helps users understand which elements—such as suspicious links, urgent language, or sender inconsistencies—triggered the alert.

Explainable AI for Phishing Detection

Applying XAI to phishing detection offers several advantages:

Transparency and Trust

When users understand why a message is flagged as phishing, they are more likely to trust the system and follow its recommendations. Transparency is especially important in multilingual settings, where misclassification can occur due to linguistic ambiguity.

User Education

Explanations serve as real-time educational tools. By highlighting phishing indicators, XAI systems help users learn how to recognize future attacks independently, strengthening human defenses.

Error Analysis and System Improvement

Explainability allows cybersecurity professionals to analyze errors, detect bias, and improve models. In low-resource languages, this feedback loop is essential for refining detection systems with limited data.

Human-Centered Design in Explainable Phishing Detection

A human-centered cybersecurity approach places users at the core of system design. For XAI-based phishing detection, this involves:

- Designing explanations that are clear, simple, and language-appropriate
- Avoiding technical jargon that non-expert users cannot understand
- Supporting multilingual explanations in users' native languages
- Considering cultural communication styles and norms

For example, instead of a generic warning, an explainable system might state: "This message is suspicious because it asks for urgent action and contains a link that does not match the sender's domain."
Such explanations empower users and reduce cognitive overload.

Explainable AI in Low-Resource Language Environments

In low-resource contexts, XAI can compensate for limited datasets by strengthening human understanding. Even if a model's accuracy is slightly lower, transparent explanations allow users to make informed decisions.
Hybrid approaches combining AI predictions with rule-based explanations and human feedback are particularly effective. Community-driven data collection, user reporting, and localized explanation templates can further enhance system performance.
Additionally, explainable models can support cybersecurity training programs by providing real examples of phishing indicators in local languages.

Ethical and Social Implications

Explainable AI addresses several ethical concerns in cybersecurity:

- **Fairness:** Transparent models help identify and reduce linguistic or cultural bias
- **Accountability:** Clear explanations support responsibility in automated decision-making
- **Accessibility:** Human-centered explanations make cybersecurity tools usable for diverse populations

In multilingual and low-resource settings, ethical deployment of AI requires not only technical accuracy but also social sensitivity and inclusiveness.

Future Directions

Future research in explainable phishing detection should focus on:

- Developing multilingual XAI frameworks
- Creating benchmark datasets for low-resource languages
- Evaluating explanation quality from a user-centered perspective
- Integrating explainable AI into cybersecurity education

Combining explainability with adaptive learning systems may further personalize protection based on user behavior and linguistic background.

Conclusion

Explainable AI represents a crucial advancement in phishing detection, particularly in multilingual and low-resource contexts where traditional AI systems fall short. By making AI

decisions transparent and understandable, XAI enhances trust, user awareness, and system effectiveness.

A human-centered cybersecurity approach recognizes that users are not merely the weakest link but a vital part of defense. Empowering users through explainable, culturally aware, and language-sensitive AI systems is essential for combating phishing in a globally connected digital world.

In conclusion, integrating explainable AI into phishing detection not only improves technical performance but also strengthens human resilience, making cybersecurity more inclusive, ethical, and effective.

Used Literature

1. Bishop, C. M. Pattern Recognition and Machine Learning. New York: Springer, 2006.
2. Goodfellow, I., Bengio, Y., & Courville, A. Deep Learning. Cambridge, MA: MIT Press, 2016.
3. Mitchell, T. M. Machine Learning. New York: McGraw-Hill, 1997.
4. Buczak, A. L., & Guven, E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials, 2016, Vol. 18(2), pp. 1153–1176.
5. Sommer, R., & Paxson, V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. IEEE Symposium on Security and Privacy, 2010, pp. 305–316.