Psychological Factors of User Vulnerability to Phishing: An Explainable AI-Assisted Detection Approach using Multi-Task Learning

Malov Gleb, Baoshan Sun

***Abstract*—Phishing remains one of the most prevalent cyber threats, exploiting human cognitive biases such as urgency, fear, and authority. While modern deep learning models achieve high accuracy in detecting phishing emails, they often lack interpretability, acting as "black boxes." This paper proposes a novel Multi-Task Learning (MTL) framework based on the RoBERTa architecture that simultaneously detects phishing attacks and identifies specific psychological manipulation techniques (e.g., Urgency, Scarcity, Authority). By integrating five heterogeneous email corpora, we created a unified dataset containing psychological technique annotations. Our experiments demonstrate that the proposed MTL model achieves a binary classification F1-score of 0.996, outperforming single-task baselines. Furthermore, the model provides interpretable multi-label outputs, enabling explainable AI (XAI) warnings that can enhance user security awareness.**

***Index Terms*— phishing detection, explainable AI, multi-task learning, RoBERTa, psychological manipulation.**

I. Introduction

Phishing attacks have evolved from simple "spray-and-pray" campaigns to sophisticated social engineering operations that exploit human psychology [7]. Despite the advancement of technical filters, users remain the "weakest link" because attackers leverage cognitive biases—such as the tendency to obey authority or react quickly to urgent threats—to bypass rational decision-making [1], [2].

Traditional AI-based detection methods focus solely on the binary classification task: determining whether an email is "phishing" or "benign" [4]. While effective, these models fail to explain why an email is dangerous, limiting their utility for user training and incident response [5], [9]. Users are often presented with a simple warning label without context, leading to "alert fatigue" [8].

To address this gap, this study presents an Explainable AI (XAI) approach. We formulate phishing detection as a multi-task problem:

- Binary Classification: Is the email malicious?
- Psychological Technique Recognition: Which manipulation tactics (Urgency, Fear, Authority, etc.) are present [11].

Manuscript received March 15, 2026

Malov Gleb, School of Computer Science and Technology, Tiangong University, Tianjin, China

Baoshan Sun, School of Computer Science and Technology, Tiangong University, Tianjin, China

A. Ethical Considerations

The development of AI systems for phishing detection requires strict adherence to ethical standards. Since datasets contain personal correspondence, de-identification protocols were strictly followed to comply with privacy regulations such as the GDPR [8]. Furthermore, the multi-task approach inherently increases system transparency by providing granular explanations, mitigating the risk of blind trust in algorithms and aligning with emerging AI regulatory frameworks [5].

II. Materials and Methods

A. Dataset Preparation

We constructed a composite dataset by harmonizing five public corpora: SpamAssassin, Nazario [15], PsyScam, Enron [14], and PhisEmails. The PsyScam corpus provided the ground truth for psychological techniques, which we mapped to nine core categories based on Cialdini's principles of persuasion: Urgency, Fear, Authority, Scarcity, Social Proof, Reciprocation, Commitment, Liking, and Consistency [9].

To annotate the emails according to these categories, we employed a hybrid approach. Initially, a keyword-based matching algorithm and regular expressions were utilized. Subsequently, a random sample of 1,500 emails was manually annotated by three independent experts to validate the initial labeling and create a high-quality gold standard.

Data preprocessing included removing HTML tags and merging the subject line with the email body. To optimize computational efficiency without losing critical context, sequences were truncated to a maximum length of 256 tokens [13]. The final dataset consists of 44,982 emails, stratified into training (70%), validation (15%), and testing (15%) sets.

B. Model Architecture

We utilized the roberta-base model as the shared backbone encoder, building upon its proven success in semantic understanding tasks [3], [10]. The architecture consists of a shared encoder and two task-specific heads:

1. Head A (Binary Classification): A linear layer with Softmax activation for Phishing/Benign prediction.
2. Head B (Psychological Techniques): A linear layer with Sigmoid activation for the multi-label classification of the 9 techniques.

To address the severe class imbalance (the "long-tail" problem) inherent in rare psychological techniques, we implemented an asymmetric loss function. Positive class

weights (*pos_weight*) were applied directly within the *BCEWithLogitsLoss* function for the multi-label task. This forces the model to focus on learning complex manipulation patterns rather than relying solely on simplistic binary features [12]. The overall architecture of the proposed MTL framework is illustrated in Figure 1.

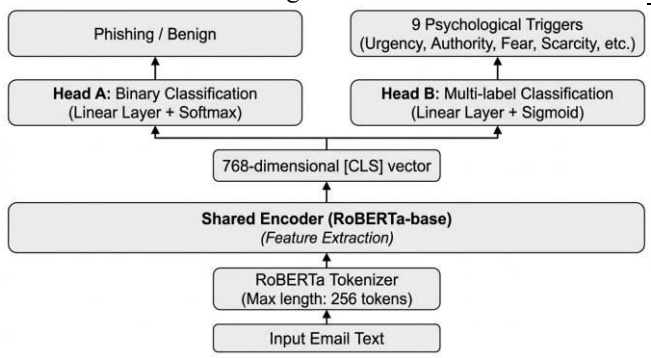

Figure 1. The proposed Multi-Task Learning architecture based on RoBERTa.

C. Experimental Setup

The model was implemented using PyTorch. Training was performed using the AdamW optimizer, which effectively decouples weight decay from gradient updates. The model was trained with a learning rate of 2e-5 and a batch size of 16. Training was halted at 4 epochs via Early Stopping as the validation loss stabilized. We compared our proposed MTL RoBERTa model against two baselines: TF-IDF + SVM and a single-task DistilBERT [3], [12].

III. Results

A. Phishing Detection Performance

The Multi-Task RoBERTa model achieved exceptional performance, confirming that the auxiliary task of identifying psychological techniques acts as a regularizer.

Table 1. Binary Classification Performance Comparison

Model	Accuracy	F1-Score	Precision	Recall
TF-IDF + SVM	0.988	0.988	0.982	0.994
DistilBERT	0.990	0.991	0.986	0.995
RoBERTa (Multi-Task)	**0.996**	**0.996**	**0.998**	**0.995**

B. Psychological Technique Identification

High-prevalence techniques such as Urgency (F1: 0.96) and Authority (F1: 0.93) were detected with high reliability. Rarer techniques like Reciprocation and Consistency showed lower performance despite the *pos_weight* application, highlighting the persistent challenge of extreme class imbalance [11].

Table 2. Multi-label Performance per Psychological Technique

Technique	Precision	Recall	F1-Score
Urgency	0.95	0.97	0.96
Authority	0.93	0.94	0.93
Fear	0.91	0.90	0.90
Social Proof	0.85	0.81	0.83
Scarcity	0.79	0.68	0.73
Liking	0.75	0.60	0.67
Reciprocation	0.71	0.54	0.61
Commitment	0.68	0.51	0.58
Consistency	0.65	0.48	0.55
Micro Average	**0.90**	**0.89**	**0.89**
Macro Average	**0.80**	**0.73**	**0.75**

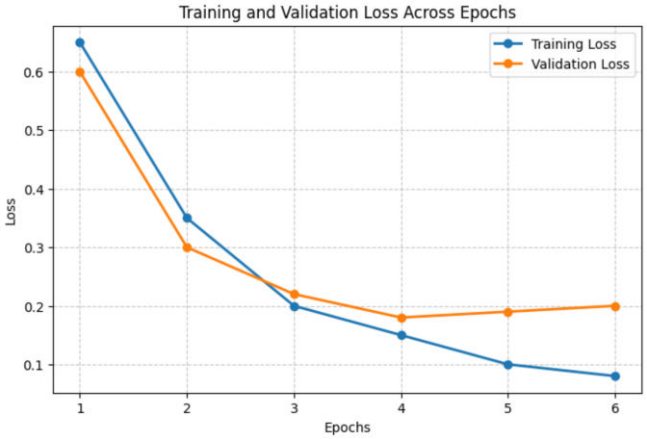

Figure 2. Training and validation loss curves.

IV. Discussion

A. Interpretability and Case Studies

A key advantage of the proposed MTL framework is its explainability. Consider the following qualitative example from our test set:

Email Subject: "URGENT: Your Account Will Be Closed!"

Model Prediction: Phishing (Probability: 99.6%)

Detected Techniques: Urgency (86.0%), Authority (72.5%)

In this case, the model correctly identified that the attacker is using Authority combined with Urgency to pressure the user. This output can be automatically translated into a user-facing warning [6].

B. Limitations and Future Work

Despite high accuracy, the performance drops on rare classes suggests that future work should employ advanced data augmentation techniques—such as synthetic email generation via Large Language Models (LLMs)—to handle the extreme tail of the distribution [12].

V. Conclusion

This study presented a RoBERTa-based multi-task learning approach for phishing detection. By simultaneously learning to classify phishing and identify psychological triggers, the model achieved state-of-the-art accuracy (F1: 0.996) and provided a layer of interpretability often missing in deep learning models. These results support the feasibility of deploying AI sentinels that not only block attacks but also educate users about manipulation tactics

References

[1] A. Gordon and D. Russ-Eft, "How the big five psychological factors affect phishing: A literature review," International Journal of Human Resource Development Practice Policy & Research, vol. 8, no. 2, pp. 100-113, 2025.

[2] WA. Kavvadias and T. Kotsilieris, "Understanding the Role of Demographic and Psychological Factors in Users' Susceptibility to Phishing Emails: A Review," Applied Sciences, vol. 15, no. 4, p. 2236, 2025.
[3] R. Meléndez, M. Ptaszyński, and F. Masui, "Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection," Electronics, vol. 13, no. 24, p. 4877, 2024.
[4] N. Altwaijry, I. Al-Turaiki, R. Alotaibi, et al., "Advancing Phishing Email Detection: A Comparative Study of Deep Learning Models," Sensors, vol. 24, no. 7, p. 2077, 2024.
[5] D. Popescul and L. D. Radu, "AI in phishing detection: a bibliometric review," Frontiers in Artificial Intelligence, vol. 8, p. 1496580, 2025.
[6] R. Pagaria, J. Xiong, R. Huang, et al., "Adaptive AI Sentinels Against Phishing Attacks: Democratizing Cybersecurity Through Interactive Learning," in Proceedings of the International Conference on AI Research (ICAIR), 2025.
[7] R. Montañez, E. Golob, and S. Xu, "Human Cognition Through the Lens of Social Engineering Cyberattacks," Frontiers in Psychology, vol. 11, p. 1755, 2020.
[8] Microsoft Threat Intelligence, "Microsoft Digital Defense Report 2025: Lighting the path to a secure future," Microsoft, Tech. Rep., 2025.
[9] S. Tiwari, "Social Engineering Attacks: Trends, Psychological Triggers, and AI-driven Prevention," Preprints.org, 2025.
[10] A. E. Alhasan, "Enhancing Real-Time Phishing Detection with AI: A Comparative Study of Transformer Models and Convolutional Neural Networks," University of Skövde, 2025.
[11] Z. Osman, N. H. Alwi, B. N. A. Khan, et al., "Psychological Impact on the Public Susceptible to Online Scams," International Journal of Academic Research in Business and Social Sciences, vol. 14, no. 5, pp. 989-1004, 2024.
[12] S. Mahendru and T. Pandit, "SecureNet: A Comparative Study of DeBERTa and Large Language Models for Phishing Detection," arXiv preprint arXiv:2406.06663, 2024.
[13] M. Chiper and R. T. Ionescu, "Every Character Counts: From Vulnerability to Defense in Phishing Detection," arXiv preprint arXiv:2509.20589, 2025.
[14] W. W. Cohen, "Enron Email Dataset," 2015. [Online]. Available: https://www.cs.cmu.edu/~enron/
[15] J. Nazario, "Nazario Phishing Corpus," 2005. [Online]. Available: https://monkey.org/~jose/phishing/