Issues in Information Systems
Volume 27, Issue 4, pp.657-669, 2026

DOI: https://doi.org/10.48009/4_iis_2026_156

Persuasion Principles and Machine Learning for Phishing Detection

Raul Diaz-Parra, *Universidad Esan*, jdiazp@esan.edu.pe

Abstract

Phishing remains one of the most significant cybersecurity threats because it increasingly relies on psychological manipulation to deceive users. Although recent detection methods have achieved high accuracy, they provide limited insight into the persuasion strategies underlying phishing attacks. This study proposes an approach that integrates persuasion theory, natural language processing (NLP), and transformer-based learning to identify multiple persuasion techniques in phishing text. A fine-tuned DistilRoBERTa model was evaluated on a balanced dataset, achieving a micro-F1 score of 0.84, a macro-F1 score of 0.84, and a Hamming Loss of 0.026. The results demonstrate that transformer-based NLP models can effectively recognize linguistic persuasion patterns, contributing to more explainable and psychologically informed phishing detection systems.

Keywords: Phishing, Natural language processing, Machine learning, Persuasion techniques, Transformed-based classification

Introduction

Phishing remains one of the most critical cybersecurity threats for organizations worldwide, driven by its high frequency, financial impact, and increasing sophistication enabled by artificial intelligence. According to the IBM (2025), phishing was identified as the most common initial attack vector, accounting for 16% of the analyzed security breaches, with an average breach cost of USD 4.8 million. Furthermore, phishing-related incidents required an average of 254 days to identify and contain significantly increasing operational and financial consequences for organizations. The report also highlights that cybercriminals are leveraging generative artificial intelligence to enhance phishing and social engineering campaigns, enabling the rapid creation of highly convincing fraudulent communications. As a result, phishing attacks not only compromise sensitive information, such as customer personally identifiable information (PII), but also cause operational disruptions, reputational damage, and substantial recovery costs, making it one of the most significant cybersecurity challenges faced by modern organizations (IBM, 2025).

Considering persuasion in phishing attacks is essential because phishing does not primarily exploit technological vulnerabilities, but rather human cognitive and psychological weaknesses. Research has shown that phishing attacks rely heavily on persuasion principles such as authority, scarcity, reciprocity, and social proof to manipulate users into making impulsive decisions and disclosing sensitive information. These persuasion techniques reduce critical evaluation and encourage heuristic thinking, especially under time pressure, thereby increasing users' susceptibility to deception. Studies indicate that attackers strategically use psychological triggers to elicit trust, create urgency, or foster social conformity, thereby significantly increasing the effectiveness of phishing campaigns. Consequently, understanding persuasion mechanisms is critical for developing more effective cybersecurity awareness programs, phishing detection systems, and user-centered defensive strategies that address not only technical threats but also the behavioral factors exploited by cybercriminals (Auton & Sturman, 2025; Stylianou et al., 2025).

Machine learning has become a fundamental technology for phishing detection due to its ability to identify complex patterns, adapt to emerging threats, and reduce classification errors in malicious emails. Several studies emphasize that traditional rule-based methods and blacklist approaches are no longer sufficient against modern social engineering techniques and sophisticated phishing attacks. In this context, Alkhodhairy and Saleem (2025) argue that machine learning models combined with Natural Language Processing (NLP) techniques can analyze linguistic, semantic, and contextual patterns associated with malicious emails, achieving 98.65% accuracy while significantly reducing false positives and false negatives. Likewise, Jerjes et al. (2026) state that targeted phishing attacks, such as whaling, employ highly personalized content capable of bypassing traditional security filters, making hybrid models based on NLP and Support Vector Machine (SVM) essential for detecting malicious intent in real time, achieving precision rates of 99.2% and recall rates of 98%. Furthermore, Ibrahim and Elhafiz (2026) highlight that transformer-based models such as BERT and RoBERTa substantially improve phishing detection by understanding contextual and sociolinguistic characteristics of messages, overcoming the limitations of static approaches and enabling the identification of disguised URLs, urgent language, and advanced text manipulation techniques. Collectively, these studies demonstrate that machine learning provides intelligent, adaptive, and scalable mechanisms to address the continuous evolution of phishing attacks, thereby strengthening organizational cybersecurity and user protection.

Literature Review

Recent literature on phishing detection reveals a common trend toward maximizing detection accuracy through machine learning, deep learning, transformers, and hybrid models. However, studies differ in the type of data analyzed, the extracted features, and the methodological approaches employed.

Rao et al. (2025) developed Phish-Jam, a hybrid super learner ensemble model designed for mobile devices. The study combined handcrafted URL features, transformer-based embeddings, and deep learning models to classify phishing websites. The authors employed Machine Learning (ML) and Deep Learning (DL) algorithms integrated into an ensemble framework and achieved 98.93% accuracy, 99.15% precision, and 99.07% F1-score, while also highlighting low latency and adaptability in mobile environments.

Similarly, Al Tawil et al. (2024) compared different textual representation techniques for phishing email detection using Term Frequency–Inverse Document Frequency (TF-IDF), Word2Vec, and BERT. The study evaluated algorithms such as Logistic Regression, Decision Tree, Random Forest, and Multilayer Perceptron (MLP). The results demonstrated that BERT achieved the best performance, with precision, recall, F1-score, and accuracy values close to 0.99, outperforming traditional textual representation approaches.

On the other hand, Zhang et al. (2025) proposed a combined feature selection approach for malicious email detection. The authors integrated 79 static features extracted from email headers and bodies together with NLP-based textual feature extraction. Additionally, they constructed the EPVME dataset containing 49,136 malicious email samples and consolidated a final dataset of 660,985 samples. Using Random Forest and other ML algorithms, the proposed model achieved 99.968% accuracy and a false-positive rate of 0.099%.

In contrast, Bustio-Martínez et al. (2024) focused their research on the psychological dimension of phishing through an analysis of persuasion principles. The study employed textual representations and machine learning classifiers to identify persuasion principles such as authority, urgency, and scarcity within phishing messages. The models achieved Area Under the Receiver Operating Characteristic Curve (AUC-ROC) values ranging from 0.7306 to 0.8191 for persuasion principle detection, while Random Forest achieved an AUC-ROC of 0.859842 for phishing classification.

Pan et al. (2026) expanded this research line by incorporating Large Language Models (LLMs) as cognitive proxies. The study operationalized psychological persuasion principles and analyzed a corpus of 340,912 emails using psychometric techniques, NLP, and LLM prompting strategies. The results showed that shifting from heuristic to systematic processing increased phishing detection accuracy from 83.09% to 88.20%, demonstrating the potential of LLMs to model human cognitive processes in this task.

In the context of real-time web detection, Dandotiya et al. (2026) proposed a browser extension based on machine learning and visual URL analysis. The study employed SVM, Decision Tree, and Random Forest, combined with the Grey Wolf Optimizer (GWO), to select lexical, structural, and visual features. The optimized Random Forest model achieved 98.7% accuracy and a Matthews Correlation Coefficient (MCC) score of 0.96, while also demonstrating the ability to identify obfuscated URLs in real time.

Finally, Goldenits et al. (2025) explored the use of Small Language Models (SLMs) for phishing website detection using HTML and URLs as input. The study evaluated 15 SLMs ranging from 1B to 70B parameters and compared their performance, computational cost, and privacy implications against proprietary models such as GPT-5.2. The results indicated that Llama3.3:70B achieved an F1-score of 0.893, compared to 0.929 for GPT-5.2, suggesting that open-source models may represent viable alternatives for local deployments with lower privacy risks.

Based on these studies, important differences emerge in the literature. Some approaches prioritize technical characteristics such as URLs, headers, or HTML, while others employ semantic representations based on NLP and transformers. More recent studies incorporate psychological variables and advanced language models. Additionally, some works emphasize accuracy and robustness, whereas others focus on privacy, interpretability, local deployment, or real-time efficiency.

Despite significant advances in phishing detection through machine learning, deep learning, transformers, and language models, most studies continue to focus primarily on technical features such as URLs, headers, HTML, textual embeddings, or lexical patterns. Studies such as those by Rao et al. (2025), Zhang et al. (2025), and Al Tawil et al. (2024) prioritize structural and semantic variables to maximize classification accuracy, achieving performance levels above 98%. However, these approaches provide a limited understanding of the psychological and persuasive components that constitute one of the primary mechanisms behind phishing attack success.

Although recent studies have begun incorporating persuasion principles and psychometric variables into phishing detection, this research line remains in an early stage. Bustio-Martínez et al. (2024) analyzed persuasion principles using machine learning classifiers and demonstrated that different combinations of textual representations and models can identify persuasive signals with AUC-ROC values ranging from 0.7306 to 0.8191. Similarly, Pan et al. (2026) operationalized psychological principles such as authority, urgency, and scarcity using LLMs and psycholinguistic processing, showing that systematic reasoning increased detection accuracy from 83.09% to 88.20%. Nevertheless, both studies focus primarily on the linguistic and psychological analysis of email textual content.

Consequently, a significant gap in literature persists regarding the deep integration of persuasion principles into intelligent phishing detection models. Current literature does not yet fully explain how persuasion mechanisms interact with technical and contextual signals in sophisticated, multichannel, and adaptive phishing attacks. Furthermore, there is limited empirical evidence regarding models that simultaneously combine psychological, linguistic, and technical variables to identify emerging persuasive tactics generated through artificial intelligence and LLMs. Therefore, there is a need to develop integrated, explainable approaches that understand how persuasion principles influence phishing effectiveness and how these

signals can be leveraged to improve the predictive and adaptive capabilities of phishing detection systems in real-world scenarios.

Theorical Framework

Persuasion Techniques in Phishing Attacks

Phishing is one of the most significant social engineering threats because it manipulates victims through fraudulent messages designed to appear legitimate. These attacks aim to obtain sensitive information, such as credentials, banking information, or personal data, via email, instant messaging, or fake websites (Salloum et al., 2021). The effectiveness of phishing largely depends on psychological persuasion mechanisms that exploit human emotions such as fear, urgency, trust, and curiosity.

Previous studies indicate that persuasion plays a central role in phishing success because attackers intentionally design messages to influence users' beliefs and behaviors (Ferreira & Teles, 2019). One of the most widely adopted theoretical models is Cialdini's persuasion principles, which include authority, scarcity, reciprocity, social proof, liking, and commitment/consistency. According to Cialdini (2009) persuasion is influenced by several psychological principles that affect human decision-making and behavior. The principle of authority suggests that individuals tend to trust experts or legitimate institutions, while scarcity exploits fear of loss and urgency. Reciprocity refers to the tendency to return favors or respond positively after receiving perceived benefits. Social proof occurs when people follow the behavior or opinions of others under uncertain situations. The liking principle explains that individuals are more easily persuaded by familiar or attractive entities, whereas commitment and consistency describe the tendency to act consistently with previous actions or beliefs. These persuasion principles are commonly exploited in phishing and social engineering attacks to manipulate users into performing malicious actions.

Research also suggests that susceptibility to persuasion techniques varies across industries. For example, finance organizations are more vulnerable to authority and reciprocity-based attacks, while non-financial sectors are more influenced by liking and social proof techniques (Tian et al., 2023). These findings demonstrate that organizational context influences phishing susceptibility and should be taken into account in cybersecurity awareness strategies.

Although the persuasion techniques included in the PerSent dataset were originally developed for computational propaganda and persuasion analysis, several conceptual similarities can be identified with the psychological persuasion principles proposed by Robert Cialdini. These persuasion techniques represent observable linguistic manifestations of psychological influence mechanisms commonly exploited in phishing and social engineering attacks. Therefore, this study proposes a conceptual mapping of persuasion-oriented linguistic patterns to Cialdini's persuasion principles, thereby strengthening the theoretical relationship among natural language processing, behavioral persuasion theory, and cybersecurity.

Table 1. Conceptual Relationship Between Persuasion Techniques and Cialdini's Persuasion Principles

Persuasion Technique	Related Cialdini Principle
Appeal to authority	Authority
Appeal to fear	Scarcity / Fear of Loss
Doubt	Uncertainty Manipulation
Exaggeration or minimization	Scarcity / Cognitive Bias
Flag-waving	Unity / Social Identity
Loaded Language	Liking / Emotional Framing
Name Calling or Labeling	Social Pressure / Polarization
Repetition	Cognitive Reinforcement

Natural Language Processing (NLP) in Phishing Detection

Natural Language Processing (NLP) is an area of artificial intelligence that enables computers to process and understand human language. In cybersecurity, NLP has become essential for analyzing textual patterns in phishing emails and detecting malicious linguistic characteristics (Alkhodhairy & Saleem, 2025).

NLP techniques such as tokenization, stopword removal, stemming, and TF-IDF are commonly used to transform email content into numerical representations suitable for machine learning models. More advanced approaches employ semantic embeddings and Transformer-based models such as BERT, which improve phishing detection by capturing contextual meaning and linguistic relationships within emails (Alkhodhairy & Saleem, 2025).

Additionally, NLP has been applied to analyze users' perceptions and decision-making processes when identifying phishing emails. Studies have shown that users often rely on URL legitimacy, perceived trust, and prior experience when detecting phishing (Cazares et al., 2021).

Classification Algorithms for Phishing Detection

Machine learning classification algorithms are widely used to automatically distinguish phishing emails from legitimate messages. Common approaches include Support Vector Machine (SVM), Naive Bayes, Random Forest, Decision Trees, and Deep Learning models (Salloum et al., 2021).

Among these algorithms, SVM has shown strong performance in text classification tasks due to its ability to handle high-dimensional textual data. Alkhodhairy and Saleem (2025) proposed an NLP-based SVM framework that achieved high accuracy in detecting suspicious emails. Deep Learning and Transformer-based models have also demonstrated superior capability in capturing semantic and contextual phishing patterns.

Recent studies have additionally explored multi-label classification approaches to identify multiple persuasion techniques within a single phishing email. Consequently, integrating persuasion theory, NLP, and machine learning represents a promising approach to improving intelligent phishing detection and cybersecurity awareness systems.

Methodology

This study employed a quantitative, experimental research design to develop and evaluate a machine learning framework for phishing detection grounded in persuasion principles. The proposed approach integrates Natural Language Processing (NLP) techniques with supervised machine learning algorithms to identify phishing emails and classify the persuasion strategies embedded within the messages. The research process consisted of dataset preparation, text preprocessing, feature extraction, model training, and performance evaluation.

Data Preprocessing

To prepare the textual data for analysis, basic natural language preprocessing techniques were applied. All text was converted to lowercase, and irrelevant special characters and excessive whitespace were removed to standardize the dataset while preserving the semantic structure of the persuasion content. Since transformer-based architectures were employed, advanced preprocessing techniques such as stopword removal and lemmatization were intentionally avoided to maintain contextual and linguistic information.

Dataset and Synthetic Data Augmentation

The original PerSent dataset exhibited class imbalance and limited representation across several persuasion categories, potentially affecting the performance and generalization capability of transformer-based multilabel classification models. To mitigate this limitation, a synthetically balanced dataset was generated through a controlled semantic augmentation process based on the original labeled samples. Because several persuasion categories contained a limited number of training instances, controlled semantic augmentation was applied exclusively to the training set to reduce class imbalance while preserving the original persuasion labels.

For each underrepresented sample, semantically equivalent variants were generated through controlled lexical and syntactic transformations. The augmentation process included synonym substitution of non-essential words, sentence restructuring while preserving semantic meaning, and paraphrasing using a Large Language Model (LLM) with constrained prompts. All persuasion labels assigned to the original sample were retained. To ensure annotation consistency, several constraints were enforced throughout the augmentation process. Specifically, the central persuasive intent had to remain unchanged, no additional persuasion techniques could be introduced, the original multilabel annotations had to be preserved, duplicated samples were removed, and every generated sample was manually reviewed to verify semantic consistency. Finally, augmentation was performed only on the training partition to prevent information leakage into the validation and test sets.

The augmentation strategy preserved the semantic meaning, persuasion intent, and original multilabel annotations while introducing linguistically diverse instances to improve class distribution. This approach aligns with prior research in natural language processing and cybersecurity, where synthetic data augmentation techniques are commonly employed to address data scarcity and class imbalance in supervised learning tasks (Feng et al., 2021). Furthermore, recent studies have shown that synthetic data generation and augmentation strategies can improve transformer-based classification performance, robustness, and generalization capability in imbalanced datasets (Won et al., 2026). Although synthetic

samples may not fully capture the complexity and variability of real-world phishing communications, the augmentation process improved classification performance across persuasion categories (Chen et al., 2023).

Label Taxonomy and Overlapping Persuasion Techniques

The PerSent dataset represents persuasion as a multilabel classification problem rather than a multiclass task. Each textual instance may simultaneously contain multiple persuasion techniques because persuasive messages frequently combine different rhetorical strategies to increase their effectiveness. The dataset includes eight persuasion categories: Appeal to Authority, Appeal to Fear, Doubt, Exaggeration or Minimization, Flag-waving, Loaded Language, Name Calling or Labeling, and Repetition. Each category is represented as an independent binary label, where a value of 1 indicates the presence of the corresponding persuasion technique and 0 indicates its absence. Since these persuasion techniques are not mutually exclusive, the proposed model predicts each label independently using sigmoid activation. Consequently, a single message may receive multiple labels whenever several persuasion techniques coexist within the same text.

During controlled semantic augmentation, the complete multilabel annotation associated with each original sample was preserved. Synthetic variants inherited exactly the same label vector as their corresponding original instance, ensuring semantic consistency and preventing the introduction of new persuasion categories during augmentation.

Transformer-Based Classification Model

Textual information was transformed into contextual embeddings using the DistilRoBERTa transformer architecture, a lightweight variant of RoBERTa optimized for computational efficiency. DistilRoBERTa was selected due to its strong performance in natural language processing tasks and its ability to capture contextual semantic relationships in persuasion-oriented text. The model was fine-tuned for multilabel classification of persuasion techniques commonly associated with phishing and social engineering attacks.

The dataset was randomly divided into training, validation, and testing subsets using an approximate 64/16/20 ratio. Transfer learning was employed using a pre-trained DistilRoBERTa model, which was subsequently fine-tuned using the persuasion dataset. Binary cross-entropy loss with sigmoid activation was used for the multilabel classification task. Hyperparameter tuning included optimization of learning rate, batch size, number of epochs, and dropout regularization to improve model generalization and reduce overfitting.

Evaluation Metrics

Model performance was evaluated using metrics commonly applied in multilabel text classification studies, including precision, recall, F1-score, and Hamming Loss. Precision measures the proportion of correctly predicted persuasion labels among all predicted labels, indicating the model's ability to minimize false positives. Recall evaluates the proportion of correctly identified persuasion labels among all actual labels, reflecting the model's capability to detect relevant persuasion techniques and minimize false negatives. The F1-score represents the harmonic mean of precision and recall, providing a balanced assessment between both metrics (Sokolova & Lapalme, 2009). Because the study involved multilabel persuasion detection, both micro- and macro-average F1 Scores were calculated. Micro-average F1-score evaluates overall classification performance by aggregating all label predictions, while macro-average F1-score calculates the average performance across all persuasion categories independently, giving equal importance to each class regardless of its frequency (Manning, 2009). Additionally, Hamming Loss measures the proportion of incorrectly predicted labels relative to the total number of labels, where lower values indicate better multilabel classification performance (Tsoumakas & Katakis, 2007). Together, these metrics enabled a

comprehensive assessment of the model's capability to identify multiple persuasion techniques simultaneously within phishing-related textual content.

Experimental Environment
All experiments were conducted using Python within the Google Colab environment. The implementation used libraries such as Pandas, NumPy, PyTorch, Scikit-learn, and Hugging Face Transformers. Graphics Processing Unit (GPU) acceleration available in Google Colab was utilized to improve computational efficiency during transformer model training.

Ethical Considerations
This research exclusively used publicly available datasets and synthetically augmented textual samples derived from the original corpus. No sensitive personal information or human participant interaction was involved, and the study was conducted solely for academic and cybersecurity research purposes.

Results

The proposed DistilRoBERTa-based multilabel classification model achieved strong performance in detecting persuasion techniques in phishing communications, as shown in Table 2. The model achieved micro-F1 and macro-F1 scores of 0.84, indicating balanced predictive performance across persuasion categories. Furthermore, the low Hamming Loss value of 0.026 suggests a relatively small proportion of incorrect multilabel predictions.

Table 2. Overall Performance Metrics of the Proposed DistilRoBERTa Multilabel Classification Model

Metric	Value
Micro F1-score	0.84
Macro F1-score	0.84
Hamming Loss	0.026
Micro Precision	0.93
Micro Recall	0.77

The highest classification performance was observed for Repetition (F1 = 0.98), Flag-Waving (F1 = 0.95), and Name Calling or Labeling (F1 = 0.93), as shown in Table 3. These persuasion techniques exhibit more explicit linguistic structures and repetitive contextual patterns, which are easier for transformer models to identify.

In contrast, Loaded Language (F1 = 0.62) and Appeal to Fear (F1 = 0.69) obtained comparatively lower scores. These categories often involve semantic ambiguity, implicit emotional framing, and contextual subjectivity, making them more challenging for automated classification models.

Table 3. Per-label Precision, Recall and F1-score for each persuasion category

Persuasion Technique	Related Cialdini Principle	Precision	Recall	F1-score
Appeal to authority	Authority	1.00	0.81	0.89
Appeal to fear	Scarcity / Fear of Loss	0.81	0.61	0.69
Doubt	Uncertainty Manipulation	0.94	0.84	0.89
Exaggeration or minimization	Scarcity / Cognitive Bias	0.94	0.62	0.74
Flag-waving	Unity / Social Identity	1.00	0.90	0.95
Loaded Language	Liking / Emotional Framing	0.71	0.56	0.62
Name Calling or Labeling	Social Pressure / Polarization	1.00	0.88	0.93
Repetition	Cognitive Reinforcement	1.00	0.95	0.98

Discussions

The results of this study demonstrate that transformer-based multilabel classification models can effectively identify persuasion techniques commonly used in phishing and social engineering communications. The proposed DistilRoBERTa model achieved micro-F1 and macro-F1 scores of 0.84, indicating strong overall predictive performance and relatively balanced classification across persuasion categories. Furthermore, the low Hamming Loss value (0.026) suggests that the model generated a limited number of incorrect label assignments, reinforcing the effectiveness of the proposed approach for multilabel persuasion detection.

The findings also reveal important differences in classification performance among persuasion techniques. Categories such as Repetition, Flag-Waving, and Name Calling or Labeling achieved the highest F1-scores, suggesting that these persuasion mechanisms contain more explicit linguistic and contextual patterns that are easier for transformer models to recognize. In contrast, techniques such as Loaded Language and Appeal to Fear performed comparatively poorly. These categories often involve implicit emotional framing, contextual ambiguity, and subjective interpretation, making automated detection more challenging. This finding is consistent with previous NLP research indicating that emotionally nuanced and context-dependent persuasion strategies are more difficult to classify accurately.

From a theoretical perspective, the results support the conceptual relationship between linguistic persuasion techniques and the psychological persuasion principles proposed by Robert Cialdini. Persuasion categories associated with explicit influence mechanisms, such as authority, social identity, and repetition, demonstrated stronger predictive performance. This suggests that transformer-based NLP models can capture linguistic patterns associated with well-established psychological influence strategies commonly used in phishing attacks.

The use of synthetic semantic augmentation also contributed positively to the study. The relatively balanced macro-F1 score indicates that the augmentation strategy helped mitigate the class imbalance in PerSent. By preserving semantic meaning and persuasion intent while increasing representation across underrepresented categories, the synthetic dataset improved model robustness and generalization capability. Nevertheless,

synthetic samples may not fully capture the diversity and complexity of real-world phishing communications, representing an important limitation of the study.

This research contributes to the intersection of cybersecurity, natural language processing, and behavioral persuasion theory by proposing an interdisciplinary approach for detecting persuasion mechanisms in phishing-related text. Unlike traditional phishing detection models that focus primarily on binary classification, the proposed approach enables simultaneous identification of multiple persuasion techniques, providing deeper insights into the psychological manipulation strategies used in phishing campaigns.
Despite the promising results, several limitations should be acknowledged. First, the study relied on a publicly available dataset and synthetically generated samples, which may not fully reflect the evolving phishing tactics observed in real-world environments. Second, the model was primarily evaluated on textual content and did not incorporate additional contextual indicators, such as metadata, URLs, or sender behavior. Future research could explore multimodal approaches to phishing detection that integrate textual, behavioral, and network-based features. Additionally, future studies may evaluate larger transformer architectures, ensemble methods, and explainable AI techniques to improve interpretability and classification performance across subtle persuasion categories.

Threats to Validity

Several threats to validity should be considered when interpreting the results of this study. First, the persuasion labels used in this research are derived from the PerSent taxonomy and represent linguistic manifestations of persuasion rather than direct measures of psychological influence. Although the selected categories were conceptually related to Cialdini's persuasion principles, this correspondence is interpretive rather than strictly equivalent. Furthermore, persuasion techniques are not mutually exclusive and may co-occur within the same message, introducing a degree of annotation subjectivity when identifying overlapping categories such as Loaded Language and Appeal to Fear. Consequently, the labels should be interpreted as linguistic indicators of persuasion rather than as direct representations of human cognitive processes.

A second threat concerns the controlled semantic augmentation strategy. Since synthetic samples were generated using predefined lexical and syntactic transformations, the augmented data may reflect patterns introduced by the augmentation process rather than the full diversity of naturally occurring phishing messages. To mitigate this risk, augmentation was applied exclusively to the training partition, duplicate instances were removed, and generated samples were reviewed to verify semantic consistency while preserving the original multilabel annotations. Nevertheless, residual linguistic artifacts may remain and could influence the model's learning process.

Another limitation relates to the characteristics of the evaluation dataset. Although PerSent provides a suitable benchmark for persuasion analysis, it was not developed exclusively from real-world phishing communications. Consequently, domain shift may occur when the proposed framework is applied to operational phishing emails, SMS messages, instant messaging platforms, or multilingual attack campaigns. Differences in writing style, organizational context, message structure, and attacker behavior may reduce the generalizability of the reported results. Therefore, future research should evaluate the proposed framework using additional phishing datasets collected from operational environments.

The experimental evaluation also presents methodological limitations. All experiments were conducted using a fixed random seed to ensure reproducibility under a consistent experimental configuration. Although this approach provides a reproducible evaluation setting, it does not capture the variability associated with different random initializations and therefore does not permit statistical significance testing

of the reported performance. Consequently, the reported results should be interpreted as reproducible outcomes obtained under a fixed experimental setting rather than as estimates of performance variability. Future research should evaluate the proposed framework using multiple random seeds and appropriate statistical significance tests to further assess its robustness.

The scope of the experimental evaluation represents an additional limitation. The objective of this work was to investigate the use of controlled semantic augmentation within a fixed DistilRoBERTa architecture rather than to perform a comprehensive benchmark across multiple machine learning and transformer-based models. Accordingly, the proposed framework was evaluated using a single transformer architecture under a consistent experimental configuration. Although comparisons with additional baseline models could provide broader benchmarking evidence and further strengthen the empirical evaluation, such analyses were beyond the scope of the present study and are considered important directions for future research.

Finally, the proposed framework may be affected by adversarial adaptation. As phishing campaigns continue to evolve, attackers may intentionally modify their linguistic strategies to reduce the effectiveness of persuasion-aware detection models. For example, explicit urgency or authority cues may be replaced with more subtle persuasive language, benign contextual information, or text generated by large language models to evade detection. Such adaptations may reduce the effectiveness of classifiers trained on relatively stable persuasion patterns. Consequently, the proposed framework should be regarded as one component of a broader phishing detection strategy rather than as a standalone defense mechanism. Future research should evaluate the robustness of the proposed framework against adversarially adapted phishing messages and investigate the integration of persuasion-based analysis with complementary technical indicators, including URLs, sender reputation, metadata, and behavioral features.

References

Al Tawil, A., Almazaydeh, L., Qawasmeh, D., Qawasmeh, B., Alshinwan, M., & Elleithy, K. (2024). Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT. *Computers, Materials and Continua*, *81*(2), 3395–3412. https://doi.org/10.32604/cmc.2024.057279

Alkhodhairy, G., & Saleem, K. (2025). Machine learning algorithm for detecting suspicious email messages using Natural Language Processing NLP. *Alexandria Engineering Journal*, *128*(April), 153–165. https://doi.org/10.1016/j.aej.2025.04.067

Auton, J. C., & Sturman, D. (2025). Persuasion under pressure: the influence of persuasion principles and time constraints on phishing email susceptibility. *Information and Computer Security*, *33*(5), 845–859. https://doi.org/10.1108/ICS-07-2024-0163

Bustio-Martínez, L., Herrera-Semenets, V., García-Mendoza, J. L., Álvarez-Carmona, M. Á., González-Ordiano, J. Á., Zúñiga-Morales, L., Quiróz-Ibarra, J. E., Santander-Molina, P. A., & van den Berg, J. (2024). Uncovering phishing attacks using principles of persuasion analysis. *Journal of Network and Computer Applications*, *230*(July), 103964. https://doi.org/10.1016/j.jnca.2024.103964

Cazares, M. F., Andrade, R., Navas, G., Fuertes, W., & Herrera, J. (2021). Characterizing phishing attacks using natural language processing. *Proceedings of the 2021 5th World Conference on Smart Trends in Systems Security and Sustainability, WorldS4 2021*, 224–229. https://doi.org/10.1109/WorldS451998.2021.9514039

Chen, J., Tam, D., Raffel, C., Bansal, M., & Yang, D. (2023). An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Transactions of the Association for Computational Linguistics*, *11*, 191–211. https://doi.org/10.1162/tacl_a_00542

Cialdini, R. B. (2009). *Influence: Science and Practice (5th Edition)*. 272.

Dandotiya, M., Goyal, N. K., Khunteta, A., & Tiwari, B. (2026). Real time identification of phishing attacks through machine learning enhanced browser extensions. *Scientific Reports*, *16*(1), 1–25. https://doi.org/10.1038/s41598-026-35655-7

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 968–988. https://doi.org/10.18653/v1/2021.findings-acl.84

Ferreira, A., & Teles, S. (2019). Persuasion: How phishing emails can influence users and bypass security measures. *International Journal of Human Computer Studies*, *125*(November 2018), 19–31. https://doi.org/10.1016/j.ijhcs.2018.12.004

Goldenits, G., Koenig, P., Raubitzek, S., & Ekelhart, A. (2025). *Small Language Models for Phishing Website Detection: Cost, Performance, and Privacy Trade-Offs*. 1–34. https://doi.org/10.3390/jcp6020048

IBM. (2024). *Informe Cost of a Data Breach 2024 Resumen ejecutivo*. https://www.ibm.com/downloads/cas/BPG4KJRX

IBM. (2025). Cost of a data breach 2025. *IBM Corportation*, 1–59. https://www.ibm.com/downloads/documents/us-en/131cf87b20b31c91

Ibrahim, M., & Elhafiz, R. (2026). Phishing Email Detection Using BERT and RoBERTa. *Computation*, *14*(2). https://doi.org/10.3390/computation14020046

Jerjes, A. Z. A. M., Hussein, I. H., & Ali-Attar, I. T. (2026). Smart Detection and Prevention of Whaling Attacks in Cyber Security. *Journal of Internet Services and Information Security*, *16*(1), 346–364. https://doi.org/10.58346/JISIS.2026.I1.020

Manning, C. (2009). Introduction to Modern Information Retrieval (2nd edition). *Library Review*, *53*(9), 462–463. https://doi.org/10.1108/00242530410565256

Pan, T. B., Yang, Q., Jordyn Cole, A., Wilson, R., & L Woodard, D. (2026). Psychology of phishing emails: Quantifying persuasion principles and simulating detection with large language models. *Expert Systems with Applications*, *313*(August 2025), 131507. https://doi.org/10.1016/j.eswa.2026.131507

Rao, R. S., Kondaiah, C., Pais, A. R., & Lee, B. (2025). A hybrid super learner ensemble for phishing detection on mobile devices. *Scientific Reports*, *15*(1), 1–17. https://doi.org/10.1038/s41598-025-02009-8

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2021). Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey. *Procedia CIRP*, *189*(2019), 19–28. https://doi.org/10.1016/j.procs.2021.05.077

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Stylianou, I., Bountakas, P., Zarras, A., & Xenakis, C. (2025). Suspicious minds: Psychological techniques correlated with online phishing attacks. *Computers in Human Behavior Reports*, *19*(July 2024), 100694. https://doi.org/10.1016/j.chbr.2025.100694

Tian, C. (Annie), Jensen, M. L., & Durcikova, A. (2023). Phishing susceptibility across industries: The differential impact of influence techniques. *Computers and Security*, *135*(March), 103487. https://doi.org/10.1016/j.cose.2023.103487

Tsoumakas, G., & Katakis, I. (2007). *Multi-Label Classification : 3*(3), 1–13.

Won, D. H., Shin, K. S., & Youm, S. (2026). Synthetic Data Augmentation for Imbalanced Tabular Data: A Comparative Study of Generation Methods. *Electronics (Switzerland)*, *15*(4), 1–27. https://doi.org/10.3390/electronics15040883

Zhang, H., Shi, Y., Liu, M., Chen, L., Wu, S., & Xue, Z. (2025). A combined feature selection approach for malicious email detection based on a comprehensive email dataset. *Cybersecurity*, *8*(1). https://doi.org/10.1186/s42400-024-00309-6