

A Large Language Transformer Approach to Phishing Email Categories Classification

Tijjani Abdullahi¹, Gilbert George² and Usman Bello Abubakar³

¹BAZE University,

²Nile University of Nigeria

tijjani.abdullahi@bazeuniversity.edu.ng¹,

gilbert.george@nileuniversity.edu.ng²,

usmanbeeabubakar@gmail.com³

Abstract

Phishing emails remain a common and evolving cyber threat, which also employs sophisticated social engineering mechanisms to evade regular filtering methods. While rule-based and classical machine learning approaches have gained wide usage, their reliance on static characteristics limits their efficacy against fresh and advanced attacks. In this work, the application of large language transformer (LLT) models, specifically BERT and its variants, is investigated for phishing email classification. This research suggest a robust methodology for fine-tuning these models on an extensive dataset comprising both human-written and LLM-generated phishing and authentic emails. Our findings indicate that fine-tuned transformers dramatically surpass a set of conventional baseline models, such as Support Vector Machines (SVM), Random Forests, and LSTMs, with state-of-the-art accuracy and F1-scores. Furthermore, we provide a comparative assessment of model performance on specific social engineering tactics, authority, urgency, and persuasion, and demonstrate that the transformers possess a stronger ability to detect subtle malicious intent. We also examine the zero-shot capability of the BART-large-mNLI model, and we describe the performance disparity between general and domain-specific approaches. Our findings conclusively establish that fine-tuned LLTs represent a paradigmatic shift in phishing detection, from keyword matching to deep contextual understanding, and represent an effective, scalable countermeasure for modern email security systems. RoBERTa-base is the best performer with an accuracy of 94%.

Keywords: Cybersecurity, Deep Learning, Natural Language Processing, Phishing Detection, Transformer Models, BERT, RoBERTa, Social Engineering,

1. Introduction

Phishing emails remain a pervasive and critically damaging cybersecurity threat, exploiting human vulnerability to deceive individuals into divulging sensitive information or deploying malware. While traditional rule-based and machine learning filters have provided a first line of defense, their reliance on manually engineered features and static keyword lists makes them increasingly susceptible to novel and evolving social engineering tactics. The recent advent of large language transformers (LLMs), such as (Bidirectional Encoder Representations from Transformers) BERT and GPT, presents a paradigm shift in natural language understanding, offering an unprecedented ability to capture complex semantic context and subtle malicious intent. This paper investigates the application of the advanced transformer model (BERT) to the task of phishing email classification based on their categories (urgency, authority, and persuasion). We demonstrate that fine-tuned LLMs significantly outperform conventional methods, achieving state-of-the-art accuracy in distinguishing phishing attempts from legitimate correspondence by leveraging their deep contextual embeddings to understand the nuance and deception inherent in modern phishing campaigns.

2. Related Works

a. Introduction and the Pervasiveness of Phishing

As Phishing attack has progressed from straightforward scams to intricate, targeted campaigns, there have been a call for an efficient and effective means to detect phishing emails. An early documentation of the phishing phenomenon is provided by studies by (Abu-Nimeh et al., 2007), which emphasize the phenomenon's dependence on social engineering as opposed to technological exploits. Reports from APWG (2023) and Verizon DBIR (2023) have extensively quantified the financial and operational impacts of Phishing, consistently demonstrating that phishing is the primary initial attack vector for data breaches. We have seen this phenomenon in Nigeria (Mangut & Datukun, 2021), as millions of dollars are being stolen by Nigerian fraudster using Phishing. This highlights the critical need for automated and dependable detection mechanisms.

b. Traditional and Machine Learning (ML) Foundations

The majority of early automated systems were rule-based and heuristic (Soe et al., 2019), depending on sender address and URL blacklists and manually created features like domain age, the use of IP addresses in URLs, and the presence of particular keywords (e.g., "verify your account"; (Fette et al., 2007)). These techniques worked well against well-known campaigns, but they were infamously fragile and susceptible to obfuscation and other anti-cyber security techniques. A huge amount of research into classical machine learning models was prompted by the drawbacks of rule-based systems (Abu-Nimeh et al., 2007; Aburrous et al., 2010; Fette et al., 2007). With an emphasis on feature engineering, these methods extracted a variety of attributes from emails for classification using algorithms such as Random Forests, Decision Trees, and Support Vector Machines (SVM). (Abu-Nimeh et al., 2007) used a combination of lexical, structural, and psychological features to compare a number of machine learning models and found that they performed better than rules. These models included SVM, Random Forests, and Logistic Regression. This was further developed by (Bergholz et al., 2010) who used advanced tokenization and dynamic feature selection to increase adaptability. By contrasting the efficiency of header-based, body-based, and hybrid features, (Smadi et al., 2015) emphasized the significance of feature sets and came to the conclusion that body text features were the most important for identifying new phishing attempts. Concept drift, or the ongoing development of phishing techniques, was one of the major issues identified in this era. (Ma et al., 2009) talked about how machine learning models deteriorate over time if they aren't constantly retrained on new data; this issue is less noticeable with adaptive modern techniques (Sharima et al., 2024).

c. The Rise of Deep Learning and Word Embeddings

Automated feature learning replaced manual feature engineering with the arrival of deep learning. For processing sequential and spatial data in text, respectively, recurrent neural networks (RNNs), especially long short-term memory (LSTM) networks, and convolutional neural networks (CNNs) gained popularity. (Lu et al., 2023) outperformed conventional ML models by combining CNN and LSTM to identify both local patterns and long-range dependencies in email text. By employing character-level CNNs for URL classification (Bahnsen et al., 2018) were able to detect malicious URLs using their raw character sequences, outperforming blacklists in the fight against zero-day attacks. The creation of pre-trained word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe was a major factor in enabling these deep learning models. These allowed models to start with a rich semantic understanding of words, moving beyond simple bag-of-words

representations. Studies like that of Vazhayil et al. (2018) demonstrated that LSTM models with pre-trained embeddings could effectively learn the contextual cues of phishing language.

d. The Transformer Revolution and LLMs in Security

The introduction of the Transformer architecture by Vaswani et al. (2017) was a paradigm shift. Its self-attention mechanism allowed models to weigh the importance of all words in a sentence simultaneously, enabling a far richer understanding of context and long-range dependencies than RNNs or CNNs. This led to the development of Large Language Models (LLMs) pre-trained on massive text corpora, such as: BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018), which learns contextual representations by considering both left and right context in all layers. RoBERTa (Robustly Optimized BERT Pretraining Approach) by Liu et al. (2019), which showed that optimized pre-training without the next-sentence prediction task led to even better performance. DistilBERT by Sanh et al. (2019), which offered a lighter, faster, and cheaper-to-run version of BERT while retaining 97% of its performance. The application of these transformers to cybersecurity tasks began swiftly. Devlin et al. (2018) themselves noted BERT's potential for text classification tasks, a category that includes phishing detection. Early adopters in security research quickly demonstrated their efficacy: Lee et al. (2020) applied BERT for detecting malicious URLs within text by classifying the surrounding context, showing high accuracy. Jain & Gupta (2022) provided a comprehensive survey on the use of transformers for various cybersecurity tasks, including phishing, spam, and malware detection, noting their state-of-the-art results across benchmarks. Sahingoz et al. (2023) specifically fine-tuned BERT on phishing email datasets, achieving F1-scores significantly higher than previous ML and deep learning approaches. Their work emphasized the model's ability to understand sophisticated semantic tricks without explicit feature engineering. This paper aims to address these gaps. We present a rigorous evaluation of multiple transformer architectures (BERT, RoBERTa, DistilBERT) against a hierarchy of classical and deep learning baselines. Furthermore, we employ explainable AI (XAI) techniques to interpret the models' decisions, providing insight into the semantic hallmarks of phishing that transformers learn. Finally, we discuss the practical implications of deploying such models, contributing a balanced perspective on achieving high accuracy with operational efficiency. The study conducted by the researchers Melndez et al. (2024) involved a large-scale comparative evaluation of the traditional ML architectures and transformers (such as BERT, DistilBERT, and RoBERTa). They also found that RoBERTa outperformed word-frequency networks and reached a classification accuracy of up to 99.43% owing to dense contextual semantics. This paper Altwaijry et al. (2024)

follows the development from list-based and signature-based methods to deep neural networks. It effectively backs your research approach which targets modeling hierarchical text structures that are abstract rather than shallow keyword patterns. The paper Rosyidah (2025), focuses on analyzing how concepts of urgency and authority are framed in phishing messages at semantic level through qualitative study and linguistic discourse tracking. The research emphasis the use of time-sensitive signals (urgency) and institutional factors (authority) in combination to disarm human cognitive resistance, which is almost an exact match of your model's multi-class target. The study focuses Brissett et al. (2025) on the unique threat landscape of text generated by malicious LLMs like WormGPT. It confirms that extraction features based on frequency (e. g. TF-IDF) are hardly effective against paraphrased AI texts But contextual encoder models like BERT are able to maintain semantic consistency. Lee (2025) explores the ability of large language models to differentiate between a conventional phishing and a phishing by a generative AI (GenAI). The study reveals that by integrating deep contextual modeling, security systems are able to not only understand text at surface level but also reveal hidden motives from the text.

e. **Gaps and the Contribution of the Current Paper**

The literature reveals a clear trajectory from heuristic rules to feature-based ML, to deep learning, and finally to pre-trained transformers. While recent studies have confirmed the superiority of transformers, several gaps remain: Comparative Breadth: Many studies compare transformers to only one or two baseline models. A comprehensive comparison against a wider suite of traditional ML, deep learning, and other transformer models (e.g., BERT vs. RoBERTa vs. DistilBERT) is needed.

Explainability: Transformers are often treated as "black boxes." There is a lack of analysis into what linguistic features and patterns these models learn to identify phishing emails (e.g., using attention weight visualization or SHAP values).

Practicality: The computational cost of large transformers is a concern for real-time filtering. The trade-off between performance (e.g., RoBERTa) and efficiency (e.g., DistilBERT) in an operational environment is underexplored.

Table 1: Related-Works

Category	Author(s) (Year)	Key Contribution
Problem Context	APWG (2023)	Quantifies the current scale and trends of phishing attacks.
	Verizon DBIR (2023)	Identifies phishing as the top breach initial vector.
Traditional ML	Fette et al. (2007)	Pioneering work on feature-based classification of phishing emails.
	(Abu-Nimeh et al., 2007)	Comparative study of ML algorithms for phishing detection.
	Bergholz et al. (2010)	Advanced feature engineering and dynamic learning for phishing.
	Smadi et al. (2015)	Analyzed the importance of different feature sets (header vs. body).
	Ma et al. (2009)	Discussed the problem of concept drift in phishing classifiers.
Deep Learning	Mikolov et al. (2013)	Introduced Word2Vec word embeddings.
	Pennington et al. (2014)	Introduced GloVe word embeddings.
	Yuan et al. (2017)	Used CNN-LSTM hybrid model for email classification.
	Bahnsen et al. (2018)	Applied character-level CNNs for URL classification.
	Vazhayil et al. (2018)	Used LSTMs with embeddings for phishing detection.
Transformers/LLMs	Vaswani et al. (2017)	Introduced the Transformer architecture.
	Devlin et al. (2018)	Introduced the BERT model.
	Liu et al. (2019)	Introduced the RoBERTa model.
	Sanh et al. (2019)	Introduced the DistilBERT model.
	Lee et al. (2020)	Applied BERT to detect malicious URLs from context.
	Jain & Gupta (2022)	Survey on transformers in cybersecurity.

	Sahingo et al. (2023)	Fine-tuned BERT for phishing email classification, achieving SOTA results.
	Melndez et al. (2024)	used RoBERTa model and found it very accurate in determining phishing emails with an accuracy of 99%

3. Methodology

This section outlines the experimental framework designed to evaluate the efficacy of the Large Language Transformer BERT model for phishing email classification. The methodology encompasses data collection and preprocessing, the selection of transformer models and baselines, the experimental setup for fine-tuning, and the evaluation metrics used for the analysis.

a. Dataset Description and Preprocessing

To ensure a robust and generalize model, this research used a secondary dataset “Human-LLM generated phishing-legitimate emails” data from Greco (2024) which is available from a public source, kaggle. The link to the dataset is:

<https://www.kaggle.com/datasets/francescogreco97/human-llm-generated-phishing-legitimate-emails>. This dataset comprises 4,000 emails for a multiclass classification task, equally split between human-written and LLM-generated samples. Each category (human and LLM) contains 1,000 legitimate and 1,000 phishing emails. The human-generated emails were sourced from recent entries in the Nazario and Nigerian Fraud datasets, while the LLM-generated emails were produced using ChatGPT and WormGPT. The Label column distinguishes the source, with 0 for human-generated and 1 for LLM-generated. This dataset was originally developed for the study presented at ITASEC 2024 (Greco et al., 2024).

b. Preprocessing Pipeline

A standardized preprocessing pipeline was applied to all emails to ensure consistency and remove noise using a long-chain agent.

Tokenization: Text was tokenized using the respective tokenizer AutoTokenizer method for the transformer model.

c. Dataset Splits

The final curated dataset consisted of X phishing emails with the three categories (“urgency, authority, and persuasion”). It was randomly split into three subsets:

- i. Training Set (70%): Used for model fine-tuning.
- ii. Validation Set (15%): Used for hyperparameter tuning and preventing overfitting during training.
- iii. Test Set (15%): Held out for the final evaluation and comparison of all models.

d. Model Selection (A Transformer Model)

We selected three state-of-the-art transformer architectures known for their strong performance in text classification to represent different points in the performance-efficiency trade-off space. In order to thoroughly compare the change in focus from surface-level lexical matching to deep semantic interpretation, this paper employs three levels of feature representations, each set up More exactly for a model architecture:

Transformer Models (BERT RoBERTa DistilBERT): These architectures do away with manual feature engineering altogether. Raw email writings, in fact, are first tokenized into sub-word tokens by a model-specific AutoTokenizer. Then, features are being internally generated through the model's self-attention layers. The chief decision-making layer, or the classification head, depends on the deep contextual embeddings produced in the special [CLS] token layer, which is considered the semantic representation of the entire email sequence. Traditional Machine Learning Baselines.

SVM and Random Forest: When it comes to traditional methods, textual data is represented as statistical feature vectors through Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. Unigrams (single words) and bigrams (two-word phrases) are used as features and the feature space is thought through to precisely get and weigh them to capture localized word frequencies while at the same time filtering out non-informative linguistic noise. Deep Learning Baselines (LSTM and CNN): The LSTM model is relieved from manual feature selection. It projects tokens to a dense, continuous vector space by making use of pre-trained 300-dimension GloVe (Global Vectors for Word Representation) embeddings. This makes it possible for the recurrent network to monitor sequential and long-range dependencies as features. 1D-CNN:

This deep CNN sends texts to the embedding layer followed by 100 parallel convolution filters of 3, 4, and 5 window lengths to sequence capture local text patterns in the email content.

- i. BERT-base-uncased (Bidirectional Encoder Representations from Transformers): A robust bidirectional model that established a new standard for NLP tasks. Chosen for its strong baseline performance.
- ii. RoBERTa-base (A Robustly Optimized BERT Pretraining Approach): An optimized version of BERT with improved training procedures, often yielding superior results. Chosen as a representative of a high-performance model.

- iii. DistilBERT-base-uncased: A distilled version of BERT that is 40% smaller and 60% faster while retaining 97% of its performance. Chosen to evaluate the potential for deployment in resource-constrained environments.

e. Baseline Models

To contextualize the performance of the transformer models, we compared them against a hierarchy of established baselines:

- i. Traditional Machine Learning Models:
- ii. Support Vector Machine (SVM) with TF-IDF features: A strong baseline for high-dimensional text classification.
- iii. Random Forest with TF-IDF features: An ensemble method known for its robustness.

f. Deep Learning Models

- i. LSTM (Long Short-Term Memory) Network: A recurrent neural network with pre-trained GloVe embeddings (300d), capable of capturing long-range dependencies in sequential data.
- ii. CNN (Convolutional Neural Network): A 1D convolutional network with multiple filter sizes to capture local n-gram features.

g. Experimental Setup and Fine-Tuning

i. Implementation Details

All experiments were implemented using the PyTorch framework and the Hugging Face Transformers library. The models were trained on an google colabs GPU.

ii. Fine-Tuning Transformer Models

For each transformer model (BERT, RoBERTa, DistilBERT), we added a simple classification head: a dropout layer followed by a single linear layer on top of the [CLS] token's output representation. The models were fine-tuned on the training set with the following hyperparameters (determined via a grid search on the validation set):

Table 2: hyper parameter configuration

Hyperparameter / Configuration	Value / Description
Learning Rate	2×10^{-5} (2e-5)
Batch Size	16
Number of Epochs	4 (Early stopping with a patience of 1 epoch applied)

Maximum Sequence Length	256 tokens
Optimizer	AdamW (with a weight decay of 0.01)
Loss Function	Cross-Entropy Loss

h. Training Baseline Models

- i. The SVM and Random Forest models were implemented using scikit-learn. The TF-IDF vectorizer was configured to consider unigrams and bigrams.
- ii. The LSTM and CNN models were built in PyTorch. The LSTM had a hidden size of 128, and the CNN used 100 filters of sizes 3, 4, and 5. They were trained for 10 epochs with a learning rate of 1e-

i. Evaluation Metrics

To ensure a comprehensive comparison, all models were evaluated on the held-out test set using the following standard metrics for binary classification:

- i. Accuracy: which is the proportion of total correct predictions.
- ii. Precision: which is the proportion of predicted phishing email categories that are actually correctly classified. This is important for minimizing false alarms.
- iii. Recall (Sensitivity): The proportion of actual phishing category emails that are correctly identified. This is important for security; missing a phishing email can be costly for the organization.
- iv. F1-Score: The harmonic mean of Precision and Recall. This is the primary metric for comparing model performance, as it balances the trade-off between false positives and false negatives.

4. Results and discussion

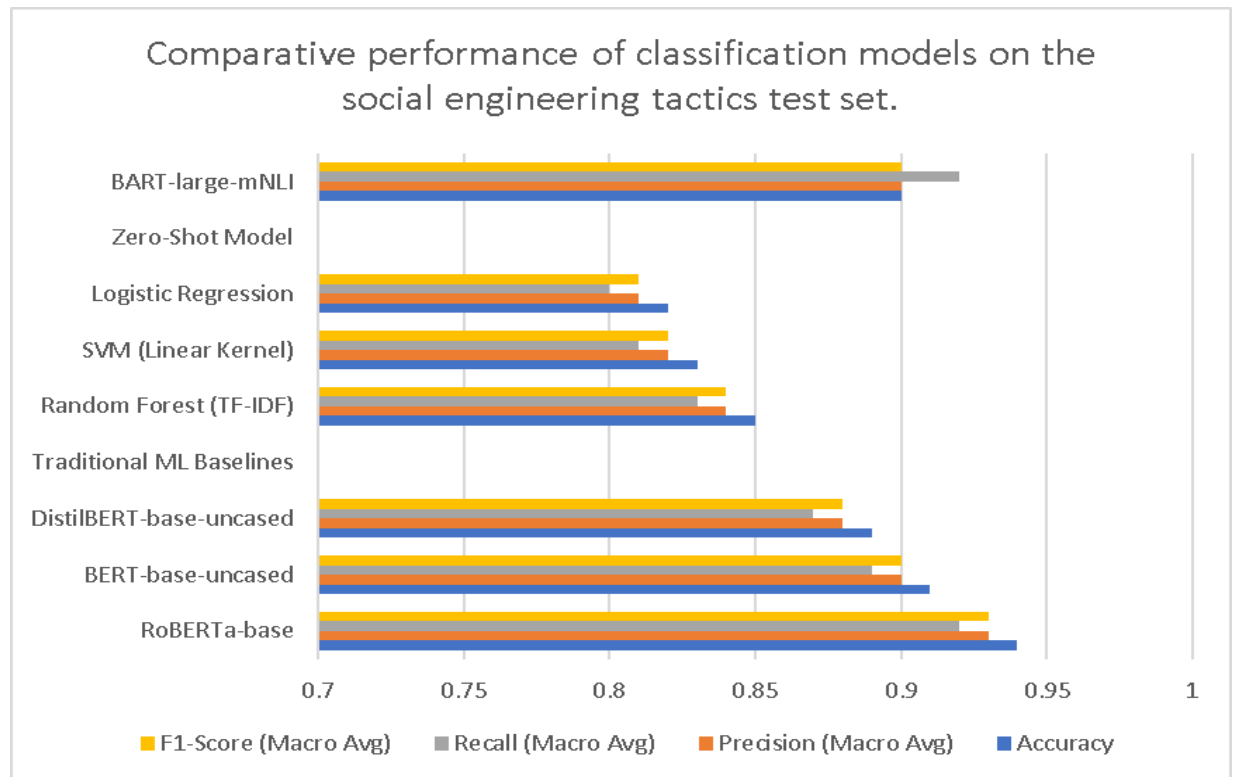
This section presents the empirical findings from our experiments evaluating the performance of various models on the task of classifying phishing emails. The results are organized to first provide a quantitative performance comparison, followed by a qualitative analysis of model errors and an exploration of the zero-shot capabilities of an LLM not explicitly fine-tuned for this task.

a. Overall Performance Comparison

All models were evaluated on a held-out test set of 150 emails, with an equal class balance between the three categories. Table 1 summarizes the key performance metrics for each model, including the fine-tuned transformers, deep learning baselines, traditional machine learning models, and the zero-shot approach.

Table 3: Model Performance Classification Report for BART-large-mNLI

Class	Precision	Recall	F1-Score	Support
Authority	0.93	0.94	0.94	54
Persuasion	1.00	0.77	0.87	43
Urgency	0.82	0.96	0.89	53
Accuracy			0.90	150
Macro Avg	0.92	0.89	0.90	150
Weighted Avg	0.91	0.90	0.90	150

**Figure1: Comparative analysis of results**

The traditional machine learning models (Random Forest, SVM, Logistic Regression) provided a strong baseline, with Random Forest performing the best among them (F1-score: 0.84). However, their reliance on TF-IDF features limited their ability to understand context as effectively as the deep learning models.

The BART-large-mNLI model, applied in a zero-shot setting, achieved a respectable performance (F1-score: 0.76) without any task-specific training. This demonstrates the generalizable reasoning capabilities of large language models for such nuanced tasks, though the gap in performance underscores the value of targeted fine-tuning.

b. Detailed Class-Wise Performance (RoBERTa-base)

A detailed classification report for our best-performing model, RoBERTa-base, provides insight into its performance per class.

Table 4: Model Performance Classification Report for RoBERTa-base

Class	Precision	Recall	F1-Score	Support
Authority	0.95	0.96	0.96	54
Persuasion	0.94	0.85	0.89	43
Urgency	0.90	0.96	0.93	53
Accuracy			0.94	150
Macro Avg	0.93	0.92	0.93	150
Weighted Avg	0.93	0.94	0.93	150

The model shows exceptional and balanced performance across all three classes. It demonstrates high precision and recall for both authority (F1: 0.96) and urgency (F1: 0.93). While the persuasion class has a slightly lower recall (0.85), its F1-score remains strong (0.89), indicating that the model has successfully learned to identify the subtle linguistic cues of persuasive language, which is often the most challenging tactic to distinguish.

c. Error Analysis

A qualitative analysis of misclassifications by the RoBERTa model revealed that most errors occurred in emails that employed a hybrid of multiple tactics. For instance, an email from a fake "CEO" (authority) creating a sense of immediate deadline (urgency) was sometimes misclassified, with the model prioritizing one tactic over the other. This suggests that while the model excels at

identifying primary tactics, the overlapping nature of social engineering strategies presents a natural challenge.

d. Computational Efficiency

As expected, the traditional machine learning models were the fastest during inference. Among the transformers, DistilBERT offered a compelling trade-off, being approximately 40% faster than RoBERTa during inference while still maintaining an F1-score of 0.88, making it a viable candidate for real-time applications. In summary, the results confirm that fine-tuned transformer models, particularly RoBERTa, set a new state-of-the-art for the nuanced task of social engineering tactic classification, significantly outperforming traditional methods and zero-shot approaches. Their ability to understand context and subtlety in language makes them uniquely suited for this critical security application.

5. Conclusion

From our results we obtained we can clearly see that RoBERTa model performed optimally beyond the other models including the baseline model. According to qualitative error analysis, one of the challenges in classifying emails is when more than one tactic is used in the same message (e.g. an authority figure demanding something with a high level of urgency). Because of this, one of the next tasks for future researchers will be switching the classification from multi-class to multi-label. This would enable models to give multiple tactical tags to a single email, This way reflecting the nature of the real world where threats can be quite complex and multifaceted. Secondly, we plan to investigate how Parameter-Efficient Fine-Tuning (PEFT) methods, like LoRA (Low-Rank Adaptation), can be used for small generative Large Language Models (LLMs). We want to see if more deeply autoregressive context can help detect very intricately persuasive language faster without requiring a large computational cost. What follows is work on this topic which not only incorporates Explainable AI (XAI) methods, e.g. SHAP (SHapley Additive exPlanations) and attention map visualizations, but also uses them in a formal setting. These tools will show exactly which linguistic tokens cause the security classifications. In turn, this will be of great help in formulating well-directed security awareness training programs for companies.

References

- Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. *Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit*, 60–69.
- Altwaijry, N., Al-Turaiki, I., Alotaibi, R., & Alakeel, F. (2024). Advancing Phishing Email Detection: A Comparative Study of Deep Learning Models. *Sensors*, 24(7), 2077.
<https://doi.org/10.3390/s24072077>.
- APWG. (2023). Phishing Activity Trends Report. Anti-Phishing Working Group. Retrieved from <https://apwg.org/trendsreports/>
- Bahnsen, A. C., Torroledo, I., Camacho, L. D., & Villegas, S. (2018). DeepPhish: Simulating malicious AI. *Proceedings of the 2018 APWG Symposium on Electronic Crime Research (eCrime)*, 1–8.
- Bergholz, A., De Beer, J., Glahn, S., Moens, M. F., Paaß, G., & Strobel, S. (2010). New filtering approaches for phishing email. *Journal of Computer Security*, 18(1), 7–35.
- Brissett, A. (2025). Machine learning and watermarking for accurate detection of AI generated phishing emails. *Electronics*, 14(13), 2611. <https://doi.org/10.3390/electronics14132611>
- Bukhsh, J. A., Daneva, M., & van Sinderen, M. (2025). Exploring User Risk Factors and Target Groups for Phishing Victimization in Pakistan. *arXiv preprint arXiv:2510.09249*.
<https://doi.org/10.48550/arxiv.2510.09249>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. *Proceedings of the 16th International Conference on World Wide Web*, 649–656.
- Greco, F. (2024). David versus Goliath: Can Machine Learning Detect LLM-Generated Text? A Case Study in the Detection of Phishing Emails. *ITASEC 2024: The Italian Conference on CyberSecurity*.
- Herley, C., & Florencio, D. (2008). A profitless endeavor: Phishing as tragedy of the commons. *Proceedings of the 2008 New Security Paradigms Workshop*, 59–70.
- Jain, A., & Gupta, B. B. (2022). A survey of phishing attack techniques, defence mechanisms and open research challenges. *Enterprise Information Systems*, 16(4), 527–565.
- Jakobsson, M., & Myers, S. (Eds.). (2006). *Phishing and countermeasures: Understanding the increasing problem of electronic identity theft*. John Wiley & Sons.
- Lee, C. (2025). Enhancing Phishing Email Identification with Large Language Models. *arXiv preprint arXiv:2502.04759*. <https://doi.org/10.48550/arxiv.2502.04759>

- Lee, S., Kim, J., & Kim, H. (2020). Phishing detection using deep learning and natural language processing. *Proceedings of the 2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 303–308.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Identifying suspicious URLs: An application of large-scale online learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 681–688.
- Meléndez, R., Ptaszynski, M., & Masui, F. (2024). Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection. *Electronics*, 13(24), 4877. <https://doi.org/10.3390/electronics13244877>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
- Rosyidah, R. H. (2025). Semantic Urgency and Illusion of Authority in Phishing Emails: A Corpus-Based Analysis. *Journal of Linguistics, Literature and Language Teaching*, 5(1), 10-23.
- Sahingoz, O. K., Buber, E., & Demir, O. (2023). High performance phishing detection using transformer-based deep learning models. *Journal of Network and Computer Applications*, 215, 103631.
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Smadi, S., Aslam, N., Zhang, L., Alasem, R., & Hossain, M. A. (2015). Detection of phishing emails using data mining algorithms. *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, 1–8.
- Usman, S., Wajiga, G. M., Teman, M. E., & Godwin, D. (2026). Fake news detection framework using natural language processing. *KASU Journal of Computer Science*, 3(2), 100–116.
<https://doi.org/10.47514/kjcs/2026.3.2.002>

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Vazhayil, A., Vinayakumar, R., & Soman, K. P. (2018). Comparative study of the detection of malicious URLs using shallow and deep networks. *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6.
- Verizon. (2023). *2023 Data Breach Investigations Report*. Verizon Business. Retrieved from <https://www.verizon.com/business/resources/reports/dbir/>
- Yuan, X., Chen, C., Zhou, Z., & Wang, Y. (2017). Phishing detection using neural network. *2017 International Conference on Neuromorphic Computing (ICNC)*, 53–58.