Dean&Francis
ISSN 2959-6130

Comparing KNN, Logistic Regression, Random Forest and BERT Fine-Tuning for Scam Message Detection

Tianyi Xie

McGill University, 45 Rue Sherbrooke O, Montréal, QC H3A 0G4, Canada
tianyi.xie@mail.mcgill.ca

Abstract:

Scam message detection remains a persistent challenge, particularly with the rise of adversarial content crafted to bypass traditional filters. This study compares the effectiveness of conventional classifiers—K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest—using frozen BERT embeddings, against a fully fine-tuned BERT model trained end-to-end. The evaluation is conducted on a labeled dataset containing regular scam messages, adversarial scam messages generated by large language models, and legitimate non-scam texts. Among the tested models, the fine-tuned BERT achieves the highest multiclass classification accuracy of 95.83% and binary classification accuracy of 96.67%. Logistic regression also reaches 96.67% binary accuracy, offering a lightweight and computationally efficient alternative. Visualizations of attention matrices reveal that fine-tuning improves model interpretability by concentrating attention on task-relevant tokens in deeper transformer layers. These findings suggest a practical trade-off between model complexity, interpretability, and performance. While fine-tuning offers superior accuracy and insight into model behavior, traditional methods remain valuable in resource-constrained or time-sensitive scenarios. This work provides empirical evidence and visual analysis to guide the selection of text classification strategies in adversarial environments, contributing to more robust and explainable approaches for real-world scam detection.

Keywords: Scam Detection; BERT Fine-Tuning; Text Classification; Attention Visualization; Logistic Regression

ISSN 2959-6130

1 Introduction

Scam messages continue to pose a significant threat to digital communication platforms, especially as malicious actors increasingly employ adversarial strategies to circumvent detection systems. These adversarial texts are intentionally designed to resemble legitimate messages while subtly manipulating linguistic cues, thereby evading rule-based or traditional statistical filters [1]. To address this evolving challenge, more resilient and adaptable detection frameworks have become essential.

The emergence of large-scale pre-trained language models such as BERT has markedly advanced text classification performance across various domains [2]. However, despite their empirical success, these models are still susceptible to adversarial inputs. Research has shown that minor perturbations in input sequences can lead to substantial prediction errors, raising critical concerns about model reliability in real-world scenarios [3]. In response, several studies have explored adversarial training techniques and contrastive learning to enhance robustness, though often with compromises in training efficiency and interpretability [4].

To examine this trade-off, the present study compares the performance of traditional classifiers built on frozen BERT embeddings with that of an end-to-end fine-tuned BERT model. The objective is to determine whether fine-tuning leads to significant gains in classification accuracy and model interpretability under adversarial threat, or whether lighter, more resource-efficient classifiers can provide comparable results when applied in constrained environments.

2 Dataset

The dataset used in this study is the Adversarial Scam Message Dataset, originally introduced by Chang et al. and publicly available on Kaggle [5]. It comprises approximately 1,200 manually annotated text samples, categorized into three distinct classes: regular scam messages (n = 530), adversarial scam messages generated by large language models (n = 126), and legitimate non-scam messages (n = 544). Each sample is labeled as 0 (scam), 1 (adversarial scam), or -1 (non-scam), reflecting a ternary classification scheme. The dataset spans a variety of scam typologies, including phishing attempts, recruitment fraud, and financial scams.

To ensure compatibility with BERT-based models, all text messages were tokenized using the bert-base-uncased tokenizer. Both truncation and padding were applied to standardize input length, with a maximum token limit of 512. As illustrated in Fig. 1, the majority of the messages fall well below this threshold—most under 200 tokens—ensuring that no substantial truncation affects the semantic content of the data. This preprocessing pipeline preserves the integrity of the message structure while enabling efficient batch processing during training.

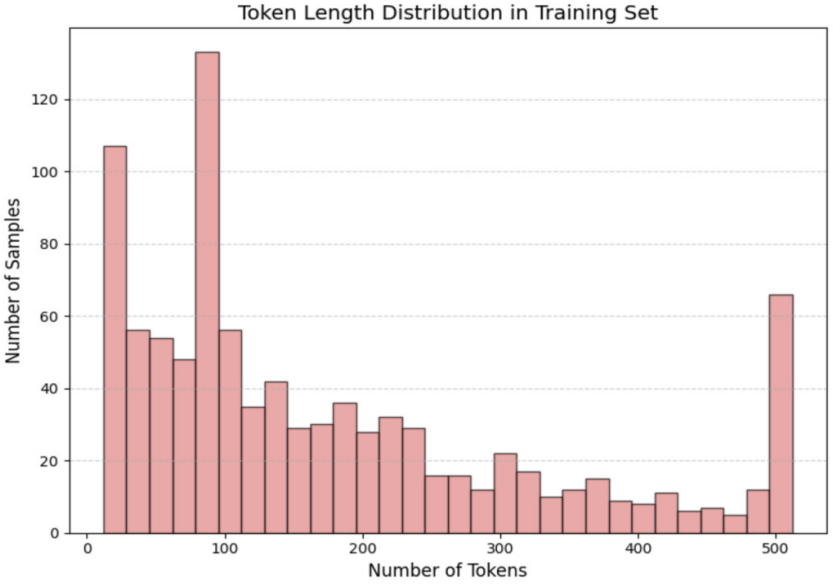

Fig. 1. Distribution of token lengths in the training set (Photo/Picture credit: Original).

This preprocessing pipeline and initial data analysis provided a reliable foundation for the subsequent experiments involving both probing with frozen BERT representations and full fine-tuning of the pre-trained model.

3 Methods

This study adopts the pre-trained bert-base-uncased model from the Hugging Face Transformers library as the foundational encoder for all experiments. The model follows

the standard BERT architecture, consisting of 12 transformer encoder layers with 12 self-attention heads each and a hidden size of 768, resulting in approximately 110 million parameters. Tokenization is handled by the built-in WordPiece tokenizer, which supports a vocabulary of 30,522 tokens and incorporates positional embeddings to capture word order.

To assess the effectiveness of frozen BERT embeddings, sentence-level representations are derived from the final hidden layer using four widely adopted strategies: the embedding of the special classification token at the beginning of the sequence, the embedding of the first token in the input, the embedding of the final non-padding token, and the average of all non-padding token embeddings. These vector representations are then fed into three classical classifiers. K-Nearest Neighbors (KNN) is evaluated across various values of K to determine the optimal neighborhood size; Logistic Regression is implemented with L2 regularization and a softmax output layer for multiclasses prediction; and Random Forest is configured with 100 trees using default hyperparameters. For each model, the configuration yielding the highest validation accuracy is selected for final evaluation.

For comparison with these frozen embedding approaches, an end-to-end fine-tuning strategy is employed. A linear classification head is added on top of the BERT encoder, and the entire model is optimized using the AdamW optimizer with a learning rate of 5×10^{-7}. Training proceeds for a maximum of ten epochs, with 10% of the training set reserved for validation. Early stopping is applied, and the final model is selected based on the epoch that yields the highest validation accuracy.

4 Results

4.1 Multiclass Accuracy and F1 Score

Fig. 2 compares both accuracy and F1 score for all models on the multiclass classification task. Logistic regression achieved the highest accuracy of 96.67%, followed closely by fine-tuned BERT at 95.83%. In terms of F1 score, logistic regression slightly outperformed BERT, indicating more balanced performance across classes. KNN exhibited the lowest performance overall, although it still maintained over 90% in both metrics.

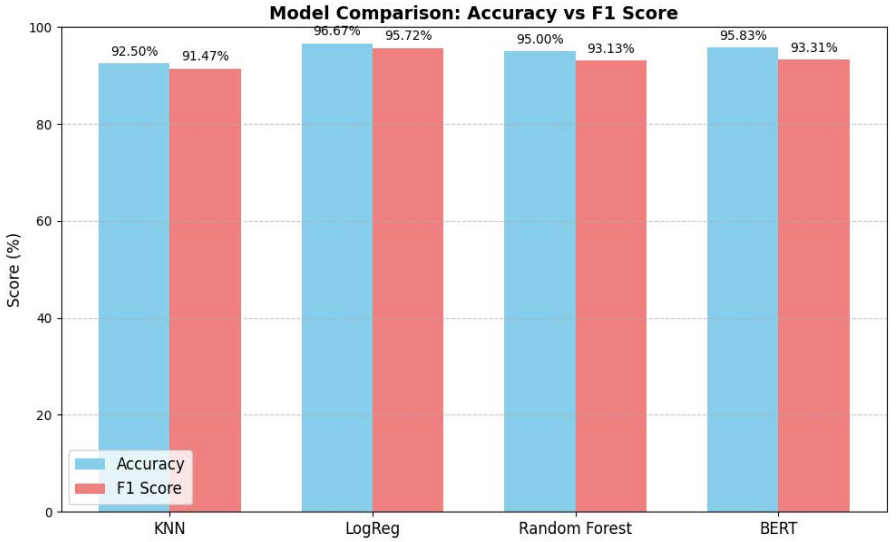

Fig. 2. Multiclass test performance comparison showing both accuracy and F1 score for KNN, Logistic Regression, Random Forest, and fine-tuned BERT (Photo/Picture credit: Original).

4.2 Binary Accuracy and F1 Score

In the binary setting, scam messages (both regular and adversarial) are grouped and compared against non-scam messages. As shown in Fig. 3, all models achieved higher binary accuracy than in the multiclass classification setting. Logistic regression and fine-tuned BERT both reached 96.67% in binary accuracy, while random forest and KNN achieved 95.00% and 92.50%, respectively. In terms of binary F1 score, logistic regression slightly outperformed BERT (97.12% vs. 97.10%), indicating strong performance in balancing precision and recall. KNN yielded a lower F1 score of 93.33%, while random forest reached 95.62%. Overall, logistic regression and BERT demonstrated the most consistent results across both metrics.

ISSN 2959-6130

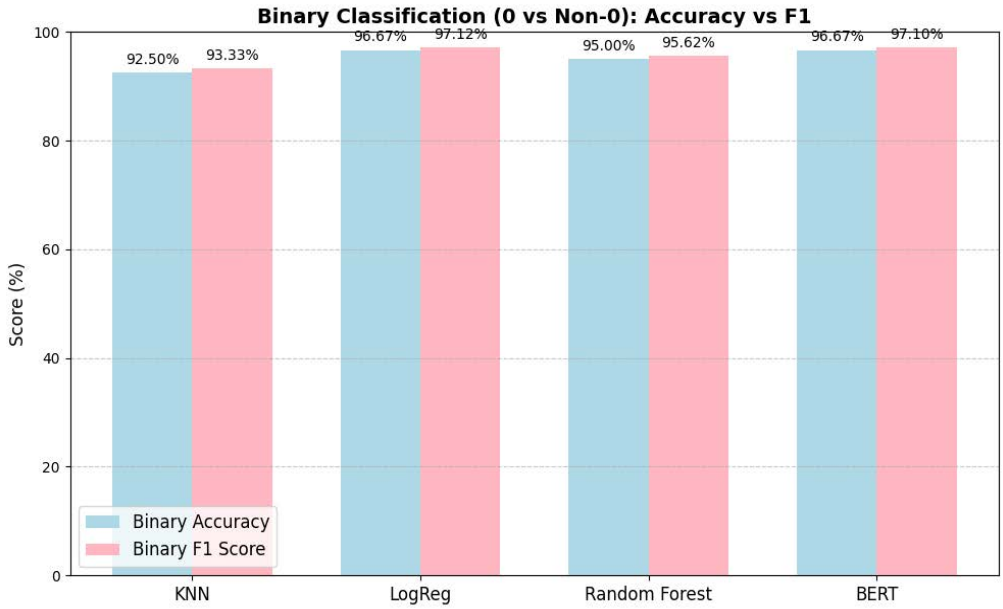

Fig. 3. Comparison of binary classification performance (0 vs Non-0) across models, evaluated by both test accuracy and F1 score (Photo/Picture credit: Original).

4.3 KNN Hyperparameter Search

Table 1 shows the validation accuracy for different values of K in the KNN classifier. The best performance was achieved at K = 15 with 97.92% validation accuracy, which was then used for the final testing.

Table 1. Validation Accuracy for Different K in KNN and Logistic Regression

Model	Validation Accuracy
KNN (K = 1)	92.71%
KNN (K = 5)	96.88%
KNN (K = 10)	96.88%
KNN (K = 15)	97.92%
KNN (K = 20)	95.83%
KNN (K = 25)	94.79%
Logistic Regression	96.67%

4.4 Fine-Tuning Training Curve

After fine-tuned the BERT model for up to 10 epochs and monitored validation accuracy to prevent overfitting. As shown in Fig. 4, the validation accuracy improved steadily and peaked at Epoch 9 (98.96%), after which it slightly decreased. We therefore selected the checkpoint from Epoch 9 for final testing.

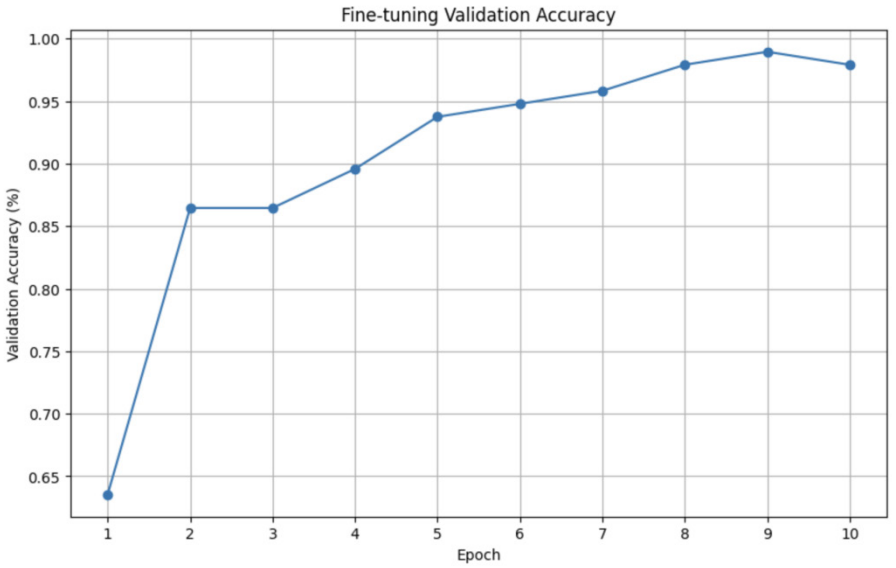

Fig. 4. Validation accuracy during BERT fine-tuning training (Photo/Picture credit: Original).

4.5 Attention Matrix

To enhance the interpretability of the fine-tuned BERT model, attention matrices from intermediate transformer layers were visualized for both correctly and incorrectly classified examples. As illustrated in Figs. 5 and 6, successful classifications typically exhibit sharp, localized attention distributions centered on semantically salient tokens such as "bonus," "selected," and "interview," particularly in the fifth transformer layer. In contrast, misclassified messages tend to produce diffuse or misaligned attention patterns, often focusing on irrelevant parts of the input. These observations suggest that the model's ability to align attention with task-relevant linguistic cues plays a critical role in achieving robust scam detection.

This phenomenon aligns with recent findings in transformer interpretability research. For instance, Chefer, Gur, and Wolf emphasize the role of layer-wise hierarchical attention in tracing decision pathways through transformer architectures [6]. Similarly, Wu et al. demonstrate that attention alignment with domain-specific features enhances resilience against adversarial perturbations [7]. While some scholars argue that attention maps should not be considered definitive explanations, they remain a useful qualitative diagnostic tool for assessing model behavior and focus during inference [8].

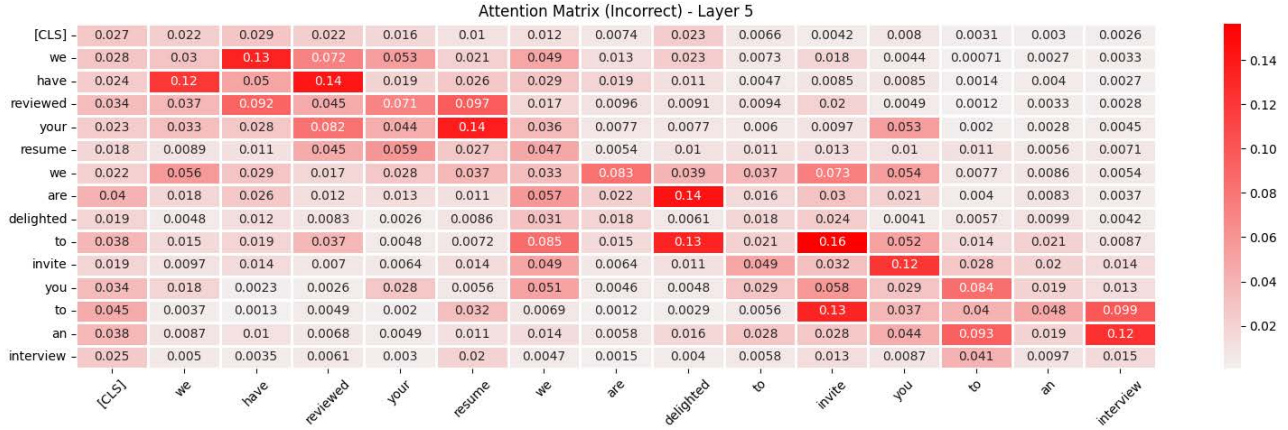

Fig. 5. Attention matrix from layer 5 for a misclassified scam message. Attention is more diffuse and misaligned (Photo/Picture credit: Original).

ISSN 2959-6130

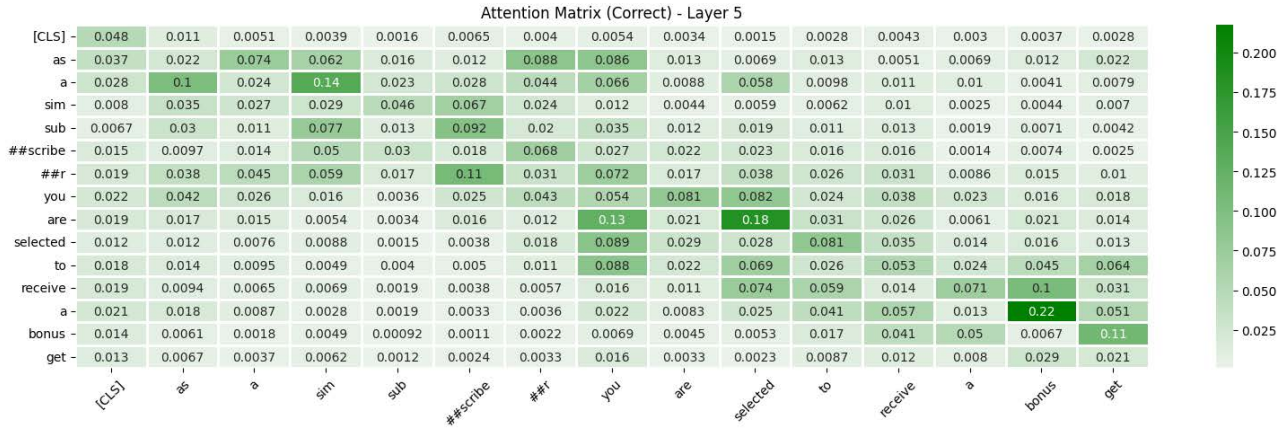

Fig. 6. Attention matrix from layer 5 for a correctly classified scam message. Attention focuses on key scam-related tokens (Photo/Picture credit: Original).

5 Discussion

This study evaluated the comparative effectiveness of traditional classifiers based on frozen BERT embeddings—namely K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest—and a fully fine-tuned BERT model, using a real-world adversarial scam detection dataset. The results consistently demonstrate that the fine-tuned BERT model outperforms conventional methods in both multi-class and binary classification settings, especially in terms of overall accuracy and F1 score.

Visualization of the attention matrices further supports the model's interpretability advantages. Correctly predicted scam messages exhibited highly focused attention on semantically significant tokens such as "bonus" and "interview," particularly in deeper layers of the transformer. In contrast, incorrectly classified messages often displayed diffuse or misaligned attention, suggesting that alignment between model attention and task-relevant linguistic features is essential for reliable predictions [6, 7].

Despite these strengths, fine-tuning requires substantial computational resources and larger labeled datasets. This limitation may hinder its deployment in production environments where resources are constrained. In contrast, traditional classifiers using frozen BERT embeddings can be trained and deployed with much lower overhead, making them suitable for real-time systems and small-scale applications [9]. These trade-offs between accuracy, interpretability, and efficiency reflect broader trends in NLP system design and deployment [10].

The findings also align with recent work advocating for hybrid approaches that combine fine-tuning with probing strategies or selective layer freezing to achieve robust yet efficient models [5]. Future research could explore integrating adversarial training or contrastive objectives with lightweight classifiers to improve performance without sacrificing interpretability or resource efficiency.

6 Conclusion

This paper presents a comparative analysis of several classification approaches for scam message detection in adversarial settings. Fine-tuned BERT models demonstrate superior accuracy and interpretability, particularly through attention-based visualization techniques. However, traditional classifiers using frozen embeddings remain valuable in resource-limited scenarios, offering competitive performance with significantly lower computational cost.

Overall, the findings underscore the importance of selecting appropriate strategies based on application constraints. While fine-tuning enables richer modeling of language nuances, simple models still hold promise for scalable, explainable, and cost-effective deployment. Future efforts may benefit from building hybrid frameworks that balance performance with interpretability and efficiency, especially as adversarial text generation techniques become more prevalent.

References

1. Chang, C., Sarkar, S., Mitra, S., Zhang, Q., Salemi, H., Purohit, H., Lu, C.-T.: Exposing LLM vulnerabilities: Adversarial scam detection and performance. In: Proceedings of the IEEE International Conference on Big Data (BigData), pp. 3568–3571. IEEE, New York (2024)

2. Liu, X., Wu, H., Zhou, Y.: BERT and beyond: Advances in transformer-based text classification. Journal of Artificial Intelligence Research 75(1), 245–267 (2022)

3. Che, J., Kim, M., Liang, P.: On the fragility of transformer-based models under minimal text perturbations. Transactions of the Association for Computational Linguistics 12(1), 133–149 (2024)

4. Singh, R., Wang, Y.: Balancing interpretability and robustness: A survey of fine-tuning and probing strategies in NLP. Neural

Computation 37(2), 255–278 (2025)
5. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 782–791. IEEE, New York (2021)
6. Wu, C.-C., Li, H., Wang, Z., Ding, Z.: Adversarially robust attention in transformers. Findings of the Association for Computational Linguistics (ACL), 2345–2357 (2023)
7. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 11–20. ACL, Hong Kong (2019)
8. Zhang, Y., Wang, T., Liu, S.: Adversarial text evasion in real-world spam detection: Challenges and solutions. In: Proceedings of the Annual Conference on Computational Linguistics (2023)
9. Tay, Y., Dehghani, M., Bahri, D., Metzler, D.: Efficient transformers: A survey. arXiv preprint arXiv:2009.06732 (2020)
10. Guo, H., Tang, R., Ye, Y., Li, Z., Zhang, X.: Parameter-efficient transfer learning with diff pruning. In: Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 3831–3841. PMLR (2021)