Detecting Phishing Emails Using Deep NLP Models and Compression Methods

Omer Hausner, Liat Shabtay

August 22, 2024

Abstract

This paper explores the use of transformer-based models, specifically RoBERTa and ALBERT, for phishing email detection. We trained and evaluated the models' performances based on a kaggle dataset which includes above 80,000 emails. Through rigorous evaluation, RoBERTa demonstrated superior performance in accuracy, F1 score, and AUC compared to ALBERT. For instance, RoBERTa scored AUC of 0.99972, which outperformed ALBERT's 0.99949. To address the computational demands of these models, we applied compression techniques including quantization, pruning, and distillation, effectively reducing model size while maintaining high performance. Our findings highlight the potential of BERT-based models in cybersecurity and the importance of model optimization for deployment in resource-constrained environments. Future work will explore integrating these models into more complex systems for enhanced phishing detection.

1 Introduction

The proliferation of phishing emails has become a major concern in cybersecurity, with the number of attacks escalating rapidly in recent years. Phishing attacks, which involve deceptive emails designed to trick recipients into revealing sensitive information, have evolved in sophistication, making them increasingly difficult to detect using traditional methods. This growing threat has prompted researchers and cybersecurity professionals to explore more advanced techniques for phishing detection. Recent statistics highlight a significant increase in phishing activities. For instance, reports show that phishing attacks surged by 47.2% in 2022 compared to the previous year, driven by increasingly sophisticated techniques employed by cybercriminals. The adoption of remote work and digital communication has further amplified the risk, as these environments are particularly vulnerable to such attacks. In 2023, credential phishing, a common form of phishing where attackers aim to steal login credentials, saw a notable growth of 17%, with nearly 7 million detections reported [1], [2].

Recent advancements in Natural Language Processing (NLP), particularly the development of transformer-based models, have shown great promise in addressing this challenge. Models like BERT [3], RoBERTa [4], and ALBERT [5] have demonstrated exceptional capabilities in understanding the complex language patterns and contexts often employed in phishing emails. These models leverage deep learning to analyze and classify emails with a high degree of accuracy, capturing subtle cues that might indicate malicious intent.

However, the application of such powerful models in real-world settings is often hindered by their computational demands. Deploying these models on resource-constrained

devices, such as those found in enterprise environments or mobile platforms, requires efficient utilization of resources. This has led to an increased focus on model compression techniques, such as quantization, pruning, and distillation, which aim to reduce the size and computational requirements of NLP models without significantly compromising their performance.

By combining the strengths of transformer-based models with advanced compression techniques, this research aims to enhance the detection of phishing emails, providing a more scalable and efficient solution that can be deployed across various platforms. This approach not only improves detection accuracy but also ensures that the solution remains practical and accessible in diverse computing environments.

2 Related Work

The classification of emails between phishing and legitimate has been a significant research focus, with transformer-based models emerging as a dominant approach due to their superior performance in various natural language processing (NLP) tasks.

Transformer models such as BERT, RoBERTa, and ALBERT have shown remarkable effectiveness in text classification tasks, including phishing detection. For example, the study by Atawneh et al. [6] highlights the capabilities of transformer-based models in detecting phishing emails. They demonstrate that transformers, when fine-tuned on a phishing dataset, outperform traditional machine learning approaches by capturing intricate patterns in email content. The paper further compares deep learning architectures for phishing email detection, including CNNs, LSTMs, and transformers. Their findings indicate that transformer models, particularly those like BERT and its variants, offer significant advantages in terms of accuracy and robustness. Specifically, transformer models excel at understanding the context and semantics of the text, making them well-suited for detecting sophisticated phishing attempts.

The importance of feature extraction, feature selection and pre-processing of the data remains relevant, especially when combined with advanced models like transformers as described in [7], [8]. These papers suggest a hybrid approach that combines content-based features, traditional data pre-processing, text-based feature engineering, together with transformer embeddings, achieving better detection rates. This approach underscores the value of combining sophisticated pre-processing and feature engineering with powerful models to enhance phishing detection.

In addition to leveraging powerful models like RoBERTa and ALBERT, researchers have explored various model compression techniques to deploy these models efficiently on resource-constrained devices. Xu, C., & McAuley, J. [9] discuss the importance of integrating NLP techniques with traditional methods to optimize both performance and resource usage. Their paper emphasizes the effectiveness of quantization, pruning and distillation as methods to reduce the computational footprint of transformer models without significantly sacrificing accuracy.

3 Exploratory Data Analysis

3.1 The Dataset

The dataset utilized in this research is sourced from Kaggle, titled the PHISHING EMAIL DATASET (link to dataset). This dataset was curated by researchers to investigate phishing

email strategies and combines emails from various origins to provide a robust resource for analysis. The final dataset comprises approximately 82,500 emails, of which 42,891 are phishing and 39,595 are legitimate. Key features include the sender and receiver email addresses, subject and body content, date, and a binary indicator of URL presence. The target label is binary, denoting whether an email is phishing or not.

3.2 Exploration of the Data

We began by exploring the dataset to understand its characteristics and underlying patterns. New features were engineered, such as the lengths of the email body and subject, the month and weekday of email transmission, and the domains of the sender and receiver. We investigated correlations between these features and the target label. While the subject length showed no correlation with the label, the body length did. The average body length for legitimate emails is 356.97 characters, compared to 195.40 characters for phishing emails. This suggests that longer body text may indicate a higher likelihood of legitimacy, making it a valuable feature (see Figure 1.

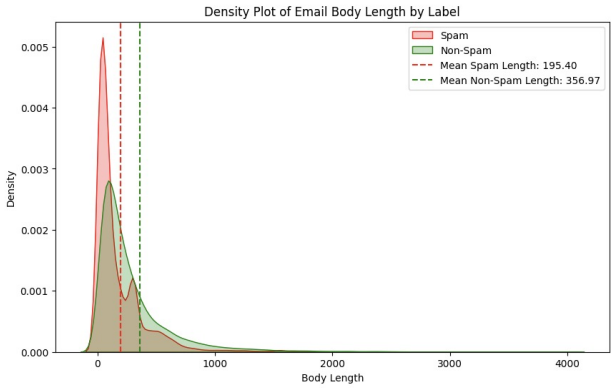

Figure 1: Density plot for body text length for each target value

Additionally, we analyzed the distribution of target labels across different sender and receiver domains, as shown in Figures 2, 3. Some domains, such as 'gmail.com' and 'twitter.com', are less likely to be associated with phishing emails, whereas others, like 'yahoo.co', 'hotmail.co', and 'msn.com', are more prone to phishing activity, often mimicking well-known brands.

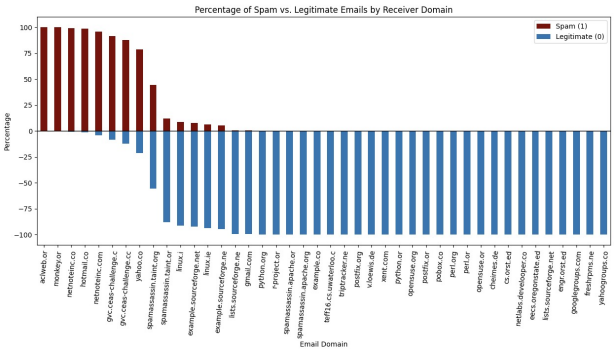

Figure 2: Target value ratio vs Receiver Email Domain

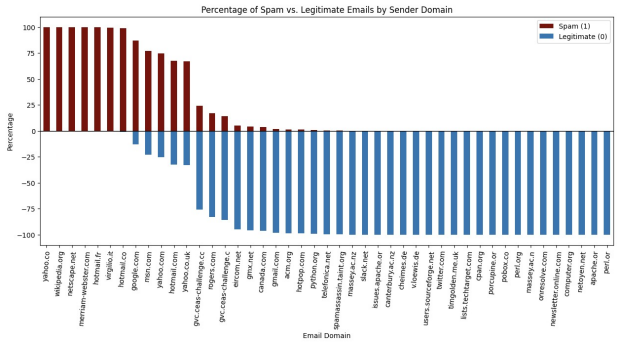

Figure 3: Target value ratio vs Sender Email Domain

Our examination of temporal features (month and weekday of email transmission) revealed no significant influence on the target label. Similarly, the presence of a URL in the email body did not show a strong correlation with phishing activity.

Finally, we explored the most common words used in phishing versus legitimate emails. As illustrated in Figures 4, 5, the word clouds for legitimate and phishing emails reveal distinct patterns in language use. Legitimate emails feature terms like "Submission", "note", and "ID", indicating structured communication. In contrast, phishing emails prominently display words like "video," "html", "index", and "cnn", which suggest a focus on deceptive

tactics, using clickbait elements and impersonation of well-known brands to lure recipients into malicious actions. This highlights the importance of analyzing language patterns to effectively differentiate between legitimate and phishing emails for better detection and security.

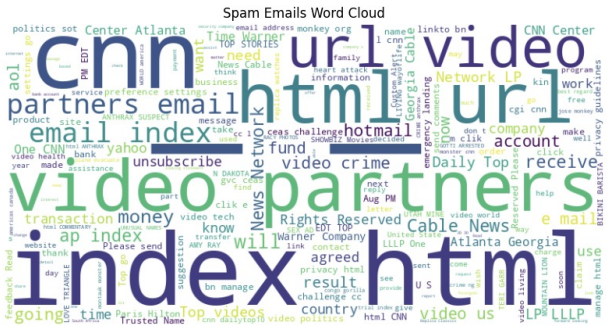

Figure 4: Most common words for spam emails

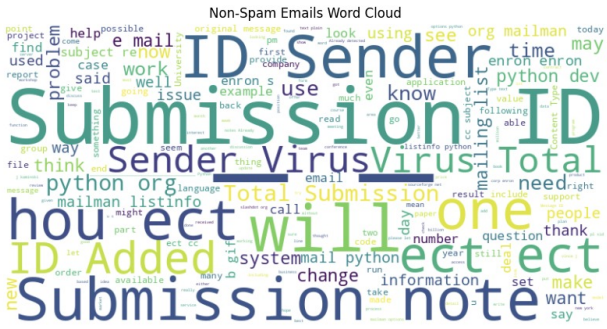

Figure 5: Most common words for legitimate emails

4 Proposed Models

To detect phishing emails using deep learning, we conducted a pre-processing stage focused on text-based information from each email, specifically the subject line and body content. We began by utilizing the NLTK library to clean the text, removing special characters and 'stop words' deemed irrelevant to the task. Next, we concatenated the cleaned subject and body texts of each email, introducing distinct tokens to distinguish between the two. The token [SUBJECT] was placed before the subject text, and [BODY] before the body content. This combined text was then tokenized using the built-in tokenizers provided by models from the Hugging Face platform.

4.1 Model Selection

For our experiments, we selected two BERT-based models: RoBERTa [4] and ALBERT [5]. RoBERTa builds on BERT [3] by adjusting key hyperparameters to enhance the training process. In contrast, ALBERT employs parameter reduction techniques to reduce memory usage and accelerate training. Both models were implemented using open-source code from the Hugging Face AI community, with their selection driven by their distinct approaches—RoBERTa enhances BERT's capabilities through deeper training (involving more parameters), while ALBERT focuses on efficiency and reduced model size. Therefore comparing the two models might raise interesting insights regarding the different approaches in the specific task of detecting phishing emails.

The dataset was split into training and validation sets with a 70/30 ratio. Each model was trained for 5 epochs with a batch size of 50 and a weight decay of 0.01, while all other parameters were left at their default settings. Training was conducted on an NVIDIA H100 80GB HBM3 GPU. To monitor the training process, we integrated our project with the Wandb platform, allowing us to track real-time progress across multiple model parameters. For instance, Figure 6 illustrates the improvement in loss during training for both the RoBERTa and ALBERT models.

4.2 Model Compression Techniques

Following the evaluation of these models, we applied three compression techniques to the model that demonstrated superior performance in terms of AUC (RoBERTa). The first technique was Quantization, where we reduced the model's parameters from 32-bit floating point to 16-bit floating point precision. The second approach involved Magnitude-Based Pruning, where 30% of the weights in each layer were pruned. Lastly, we employed a distillation-based technique, training a smaller student model, DistilBERT [10], to replicate the output of the larger model while being more computationally efficient.

5 Results

Models	Loss	F1 Score	Accuracy	Precision	Recall	AUC	Inference Time (sec)	Model Size (Mb)
RoBERTa	0.03723	0.99331	0.99305	0.9937	0.99292	0.99972	0.0063	475.52
ALBERT	0.04339	0.9918	0.99147	0.99115	0.99246	0.99949	0.0032	44.5
Quantized RoBERTa	0.03722	0.99331	0.99305	0.9937	0.99292	0.99972	0.0028	237.77
Pruned RoBERTa	0.03583	0.99259	0.99228	0.99047	0.99471	0.99963	0.0028	475.52
DistilBERT	0.03755	0.99203	0.99175	0.99026	0.99382	0.99956	0.0028	313

Table 1: Comparison of Model Performance Metrics

5.1 Comparison between ALBERT and RoBERTa

As shown in Table 1, the results indicate that while both ALBERT and RoBERTa demonstrated strong performance in detecting phishing emails, RoBERTa consistently outperformed ALBERT across several key metrics. Specifically, RoBERTa achieved an F1 score of 0.99331, slightly higher than ALBERT's 0.9918. This trend is reflected in the accuracy as well, where RoBERTa's 0.99305 marginally surpassed ALBERT's 0.99147. In terms of the Area Under the Curve (AUC), RoBERTa scored 0.99972, edging out ALBERT's 0.99949. While ALBERT demonstrated faster inference times and significantly reduced model size, these benefits came at the cost of slightly lower predictive performance compared to RoBERTa.

The loss curves during training, as depicted in Figure 6, reveal that both models experienced a rapid decline in loss initially, followed by a gradual stabilization. However, RoBERTa's loss consistently remained higher than ALBERT's throughout training, suggesting that RoBERTa required more steps to converge. Despite this, RoBERTa's slightly better overall performance highlights its ability to generalize better at the cost of requiring more training iterations.

5.2 Comparison of Compressed RoBERTa Models

Given RoBERTa's superior performance, three model compression techniques were applied to further optimize it: quantization, pruning, and distillation. Among these, quantization effectively reduced the model size by almost half (from 475.52 Mb to 237.77 Mb) while maintaining the same high performance across all metrics. The pruned RoBERTa model, on the other hand, achieved a slightly lower F1 score (0.99259) and accuracy (0.99228) compared to the full RoBERTa, yet still maintained a commendable AUC of 0.99963. Distillation with DistilBERT also demonstrated effective compression, reducing the model size to 313 Mb while still delivering competitive performance across all metrics. Based on these results, quantization appears to be the most efficient compression method, achieving

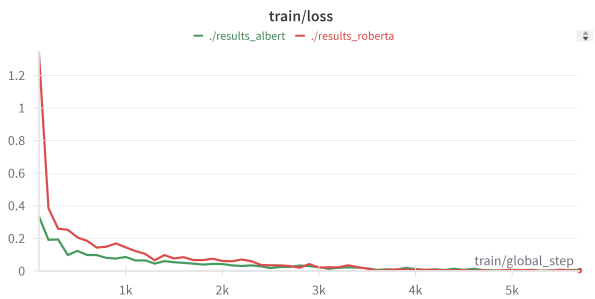

Figure 6: Training Loss Progression for RoBERTa and ALBERT Models - Monitored via Wandb Platform

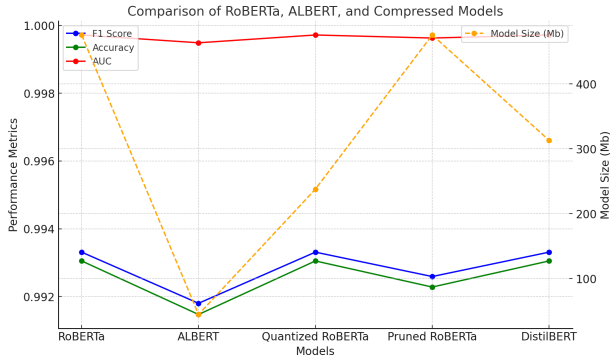

Figure 7: Graph Results for the Models' Performance

significant size reduction without sacrificing accuracy, although DistilBERT also offers a strong balance between model size and performance.

6 Discussion

In this study, we explored the use of transformer-based models, specifically RoBERTa and ALBERT, for the task of phishing email detection. We conducted a comprehensive analysis, comparing the performance of these models and applying various model compression techniques to optimize their deployment in resource-constrained environments. Our results indicate that RoBERTa outperformed ALBERT across several key metrics, demonstrating superior accuracy, F1 score, and AUC. To further enhance the practicality of using RoBERTa in real-world applications, we applied quantization, pruning, and distillation, which reduced model sizes while maintaining competitive performance levels.

The impressive results of transformer models in this study align with the broader trend observed in the field of Natural Language Processing (NLP), where BERT-based models have consistently demonstrated high effectiveness across various language tasks. Our findings reinforce the capability of these models to accurately detect phishing emails, which is a critical application given the increasing sophistication of phishing attacks. The ability of these models to capture subtle linguistic cues that are often indicative of phishing attempts underscores their potential to significantly enhance cybersecurity measures.

Model compression techniques play a crucial role in making these powerful models more accessible and usable in different operational contexts, especially where computational resources are limited. Techniques such as quantization and distillation not only reduce the model size but also preserve a high level of accuracy, making it feasible to deploy these models in environments where traditional transformer models would be too resource-intensive. The success of these techniques in our study highlights the importance of ongoing research into optimizing NLP models for practical use without compromising on performance.

Looking forward, the integration of these phishing detection models as part of a larger ensemble or feature set within more complex systems could further enhance their utility. By incorporating the outputs of these models as additional features in broader cybersecurity frameworks, we may uncover new synergies that could lead to even greater predictive performance. Such integration represents a promising direction for future work, especially as the nature of phishing attacks continues to evolve.

References

[1] D. Desai *et al.*, "2023 phishing report reveals 47.2

[2] E. Dzuba and J. Cash, "Introducing cloudflare's 2023 phishing threats report," https://blog.cloudflare.com/2023-phishing-report/, 2024, accessed: 20 August 2024.

[3] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[6] S. Atawneh and H. Aljehani, "Phishing email detection model using deep learning," *Electronics*, vol. 12, no. 20, p. 4261, 2023.

[7] S. Magdy, Y. Abouelseoud, and M. Mikhail, "Efficient spam and phishing emails filtering based on deep learning," *Computer Networks*, vol. 206, p. 108826, 2022.

[8] D. O. Otieno, A. S. Namin, and K. S. Jones, "The application of the bert transformer model for phishing email classification," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2023, pp. 1303–1310.

[9] C. Xu and J. McAuley, "A survey on model compression for natural language processing," *arXiv preprint arXiv:2202.07105*, 2022.

[10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.