5-2026

Phishing Email Detection Using Subject and Body Text: A Comparative Study of Transformers, Bi-LSTM, and Traditional Classifiers

Phyo Htet Kyaw
Auckland University of Technology, phyohtet.kyaw@autuni.ac.nz

Jairo Gutierrez
Auckland University of Technology, jairo.gutierrez@rocketmail.com

Akbar Ghobakhlou
Auckland University of Technology, akbar.ghobakhlou@aut.ac.nz

21R. Phishing Email Detection Using Subject and Body Text: A Comparative Study of Transformers, Bi-LSTM, and Traditional Classifiers

Phyo Htet Kyaw
Auckland University of Technology
phyohtet.kyaw@autuni.ac.nz

Jairo Gutierrez
Auckland University of Technology
jairo.gutierrez@aut.ac.nz

Akbar Ghobakhlou
Auckland University of Technology
akbar.ghobakhlou@aut.ac.nz

Abstract

Phishing emails continue to bypass blacklist, signature, and rule-based defenses, motivating the development of detectors that can learn robust patterns from message content. This Research-in-Progress paper reports completed results for the subject+body text module of a planned multi-segment phishing email detector covering headers, URLs, and attachments. We compare Transformer-based models (BERT, DistilBERT, DeBERTa-v3-base, DeBERTa-v3-small) with recurrent neural models (LSTM, Bi-LSTM) and traditional classifiers (Logistic Regression, SVM, Random Forest, Decision Tree, Naïve Bayes) under consistent preprocessing and evaluation settings. DeBERTa-v3-small achieves the strongest overall performance on the text module (F1 = 0.9972, Matthews correlation coefficient (MCC) = 0.9950), followed closely by BERT (F1 = 0.9969, MCC = 0.9944). DistilBERT provides a strong efficiency–effectiveness trade-off (F1 = 0.9957, MCC = 0.9922), and a linear SVM remains competitive among traditional classifiers (F1 = 0.9963, MCC = 0.9933). At batch size 1, the model-only inference time is 4.84 ms/message for DistilBERT, compared with 8.98 ms/message for BERT and 12.27 ms/message for DeBERTa-v3-small. These results provide a strong baseline for the text module, and future work will implement the remaining modules and measure their added value through controlled ablation and fusion studies.

Keywords: Phishing detection, Phishing email, Email security, Deep learning.

1. Introduction

Social-engineering attacks manipulate users by impersonating trusted entities and crafting persuasive, deceptive communications. As a form of social engineering, email phishing tricks recipients into giving up their digital identities and confidential information, while sometimes distributing malware. In 2023, the Anti-Phishing Working Group (Anti-Phishing Working Group, 2024) reported nearly five million phishing incidents. The consequences are severe: between October 2013 and December 2022, business email compromise (BEC) accounted for an estimated USD 51 billion in losses, and phishing ranked among the most frequently reported cybercrimes (Federal Bureau of Investigation, 2022). Industry analyses further note that in 2022, emails delivered 86% of all malware, with ransomware comprising about 35% of those payloads (Check Point Research, 2023; Verizon, 2022). While blacklist, whitelist, and signature-based defenses are widely deployed, adversaries increasingly evade them by using well-crafted lures, embedded malicious content and links, and reputable mail infrastructure to slip past traditional filters. Attackers are also using machine learning (ML) to generate more convincing messages that undermine legacy detectors (Yamin et al., 2021).

To overcome these limitations, recent research has emphasized ML and deep learning (DL) for phishing detection, seeking both higher accuracy and better adaptability (Kocher & Kumar, 2021). Deep learning models such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and convolutional neural networks (CNNs) can learn discriminative patterns directly from email data and have shown strong performance for phishing detection (Lee et al., 2021). However, most systems analyze only a partial segment of the email, typically body text, headers, and URLs, leaving threats that hide in obfuscated links or in attachments insufficiently addressed. Our prior literature review (Kyaw et al., 2024) emphasized this gap and recommended models that consider all major segments of the email.

This Research-in-Progress paper reports completed results for phishing email detection using subject and body text. We compare Transformer-based models (BERT, DistilBERT, DeBERTa-v3-base, DeBERTa-v3-small), recurrent neural models (LSTM, Bi-LSTM), and traditional classifiers (Logistic Regression, SVM, Random Forest, Decision Tree, Naïve Bayes) under consistent preprocessing and evaluation settings (Devlin et al., 2019; Graves & Schmidhuber, 2005; He et al., 2021; Hochreiter & Schmidhuber, 1997; Sanh et al., 2019). The contributions of this paper are: (1) a completed subject+body text module with comparative results across multiple models under consistent settings; (2) reporting imbalance-aware measures (F1, MCC) together with model-only inference time at batch sizes 1 and 32.

While this paper focuses on the text module, our broader research goal is a modular, multimodal detector that will later incorporate headers, URLs, and attachments. The remaining modules and fusion strategy are left for future work, and will be evaluated using controlled ablation to quantify the added value of each module.

2. Related Work

Most recent phishing-email detectors focus on a single email segment. Our prior systematic literature review found that most studies use body-text features, while fewer incorporate headers, URLs, or attachments, which also makes cross-study comparisons difficult because datasets and evaluation settings vary (Kyaw et al., 2024). A smaller line of work moves toward multi-segment analysis. Muralidharan and Nissim (Muralidharan & Nissim, 2023) classify headers, body, and attachments with separate models and combine them using stacking, reporting strong performance on VirusTotal data. Lee et al. (2021) combine structure, text, and URL modules with a meta-classifier and report high performance on enterprise data, but they do not analyze attachments.

These studies highlight the value of multi-segment designs, but attachments and advanced obfuscation (e.g., shortened links and QR-encoded URLs) remain under-represented in many published systems. In this Research-in-Progress, we focus on the subject+body text module and provide a strong baseline by comparing multiple model families under consistent settings. The remaining modules and fusion strategy are planned as future work and will be evaluated through controlled ablation to measure the added value of each segment.

3. Methodology

All experiments were implemented in Python and executed on Google Colab using an NVIDIA A100 GPU runtime. The workflow trains multiple model families on a fixed training split and evaluates them on a fixed held-out test split under consistent settings. We use Hugging Face Transformers (PyTorch) for Transformer fine-tuning, TensorFlow/Keras for recurrent neural models, and scikit-learn for traditional classifiers. To improve reproducibility, the same random seed (42) was used for Python, NumPy, and PyTorch.

3.1 Data and Pre-processing

This study uses the Enron (Cohen, 2009) and Nazario (Nazario, 2006) datasets, which are well-established benchmarks in phishing detection research. The experimental corpus consists of 10,000 randomly selected legitimate emails from Enron and 8,122 phishing emails from Nazario. For each message, the subject and body are parsed, concatenated, and stored as a single text field with a binary label (0 = legitimate, 1 = phishing). The dataset is split into 80% training and 20% testing, and the same predefined split is used for all models. A limitation of this evaluation is that legitimate and phishing emails are drawn from different and relatively old public corpora. As a result, some class separation may reflect corpus-specific characteristics rather than broadly applicable phishing indicators. The results are therefore treated as a controlled baseline for the text module.

Dataset diversity was assessed using pairwise cosine similarity (Equation 1), computed on the combined subject+body text. Before feature extraction, HTML tags were removed, and URLs and email addresses were replaced with placeholder tokens. L2-normalized term frequency–inverse document frequency (TF–IDF) features were extracted using word N-grams (1–2), a minimum document frequency min_df = 3, and a maximum of 120,000 features. Next, 50,000 random pairs were sampled for each pair type (legitimate–legitimate, phishing–phishing, and legitimate–phishing), and similarity values were summarized as percentages in fixed bins as shown in Table 1, consistent with prior reporting practice (Aassal et al., 2020). Most pairs fall within 0–10% similarity, indicating high lexical diversity. We also visualize the dataset using t-Distributed Stochastic Neighbor Embedding (t-SNE) on TF–IDF representations after dimensionality reduction with Singular Value Decomposition (TF–IDF → SVD → t-SNE)(Maaten & Hinton, 2008). Figure 1 shows multiple phishing clusters and partial overlap with legitimate emails, suggesting both lexical variation and a realistic classification challenge.

$$\cos\_\mathrm{sim}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{\boldsymbol{x_i} \cdot \boldsymbol{x_j}}{\|\boldsymbol{x_i}\|_2 \, \|\boldsymbol{x_j}\|_2} \tag{1}$$

Pair Type	[0-10]	(10-20]	(20-30]	(30-40]	(40-50]	>50
Legit – Legit (0↔0)	98.29	1.49	0.14	0.04	0.01	0.02
Phish – Phish (1↔1)	92.26	5.53	1.15	0.56	0.21	0.29
Legit – Phish (0↔1)	99.99	0.01	0	0	0	0

Table 1: Distribution of pairwise email similarities within the dataset

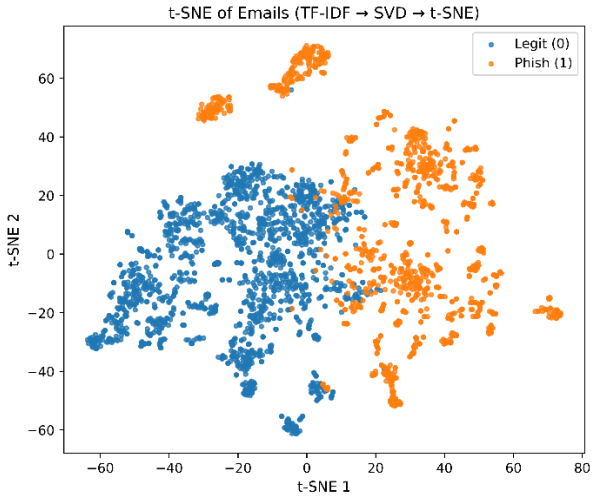

Figure 1: 2D t-SNE projection of the dataset

Pre-processing is model-specific and follows representative pipelines for each model family. Transformer models use the corresponding pretrained tokenizer with truncation and padding to a maximum sequence length of 512 tokens. LSTM and Bi-LSTM inputs are lowercased, stripped of URLs and non-alphanumeric characters, tokenized, stopwords removed, and lemmatized. A Keras tokenizer with a vocabulary size of 10,000 is fitted on the training set, and sequences are padded or truncated to length 512. Traditional classifiers use TF–IDF word uni-grams and bi-grams with max_features = 50,000 and min_df = 2.

3.2 Models and Training

We compare four Transformer models (BERT, DistilBERT, DeBERTa-v3-base, DeBERTa-v3-small), two recurrent neural models (LSTM and Bi-LSTM), and five traditional classifiers (LR, SVM, RF, DT, and NB). Transformer models are fine-tuned for 5 epochs with maximum sequence length 512, learning rate 0.00005, AdamW optimization, weight decay 0.01, and mixed-precision training on GPU. BERT and DistilBERT use per-device batch size 16 with gradient accumulation 2. DeBERTa models use gradient checkpointing, disabled cache, and fallback memory-safe settings (batch size 8 with gradient accumulation 4, then batch size 4 with gradient accumulation 8, with max length reduced to 256 only if required). LSTM and Bi-LSTM use a vocabulary size of 10,000, sequence length 512, embedding dimension 128, 64 hidden units, recurrent dropout 0.2, batch size 32, Adam optimizer, binary cross-entropy loss, and 5 training epochs. Traditional classifiers use TF-IDF features with uni-grams and bi-grams, max_features = 50,000, and min_df = 2; Logistic Regression uses max_iter = 2000, the linear SVM uses sigmoid calibration with 3-fold cross-validation on the training data only, and Random Forest uses 300 trees. Binary predictions are generated using the default decision threshold for each model, without additional class weighting or threshold tuning.

3.3 Evaluation and Inference Timing

All models are evaluated on the same test set. Because phishing detection commonly involves class imbalance, we assess model performance primarily using F1-score and Matthews correlation coefficient (MCC). The latter provides a more balanced evaluation by accounting for asymmetry between legitimate and phishing samples. Confusion matrices are also produced for each model. Inference timing is measured on a shared subset of test messages using warm-up iterations before timing. We report model-only inference time at batch sizes 1 and 32. For Transformers, model-only time measures the forward pass. For LSTM/Bi-LSTM it measures the prediction step. For traditional classifiers it measures classifier scoring on already vectorized features.

4. Results

Results for the completed subject+body text module are reported on the held-out test split. Figure 2 and Figure 3 rank models by MCC and F1-score, while Figure 4 summarizes all metrics. DeBERTa-v3-small achieves the best results (F1 = 0.9972, MCC = 0.9950), followed by BERT (F1 = 0.9969, MCC = 0.9944). DistilBERT remains strong (F1 = 0.9957, MCC = 0.9922), and a linear SVM is competitive among traditional classifiers (F1 = 0.9963, MCC = 0.9933). However, these near-ceiling results should be interpreted cautiously because the current benchmark may contain corpus-level cues that make legitimate and phishing emails easier to separate than in real-world settings. Figures 2–4 summarize the rankings and metric overview. Figure 5 reports model-only inference time (ms/message) for bs = 1 and bs = 32. At bs = 1, DistilBERT is the fastest Transformer (4.84 ms/message), ahead of BERT (8.98 ms/message) and DeBERTa-v3-small (12.27 ms/message), while recurrent models are

substantially slower (LSTM: 315.17 ms/message; Bi-LSTM: 420.03 ms/message). With batching (bs = 32), per-message inference time decreases (e.g., DistilBERT: 2.79 ms/message; DeBERTa-v3-small: 4.53 ms/message), supporting DistilBERT as a strong accuracy–latency trade-off and DeBERTa-v3-small when peak detection performance is preferred.

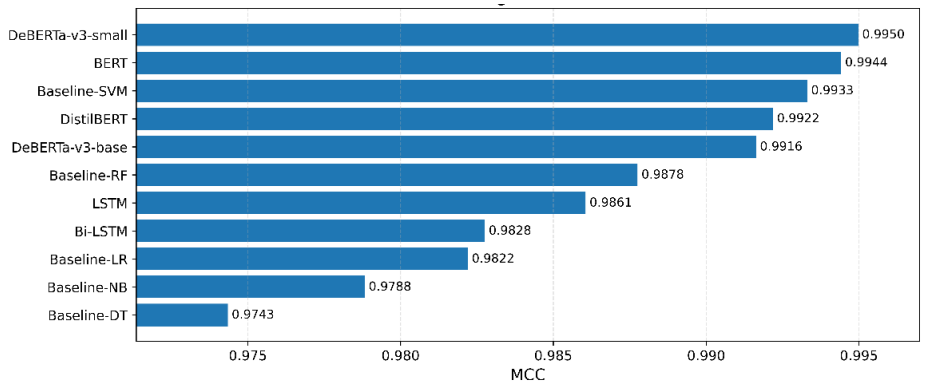

Figure 2: MCC comparison across all evaluated models

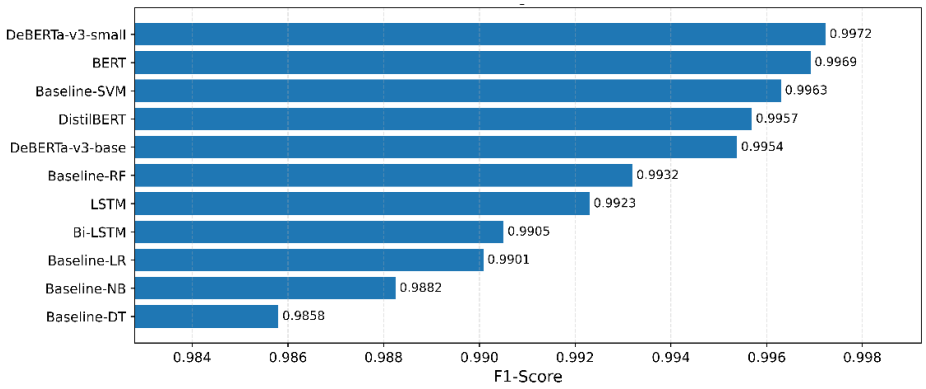

Figure 3: F1-score comparison across all evaluated models.

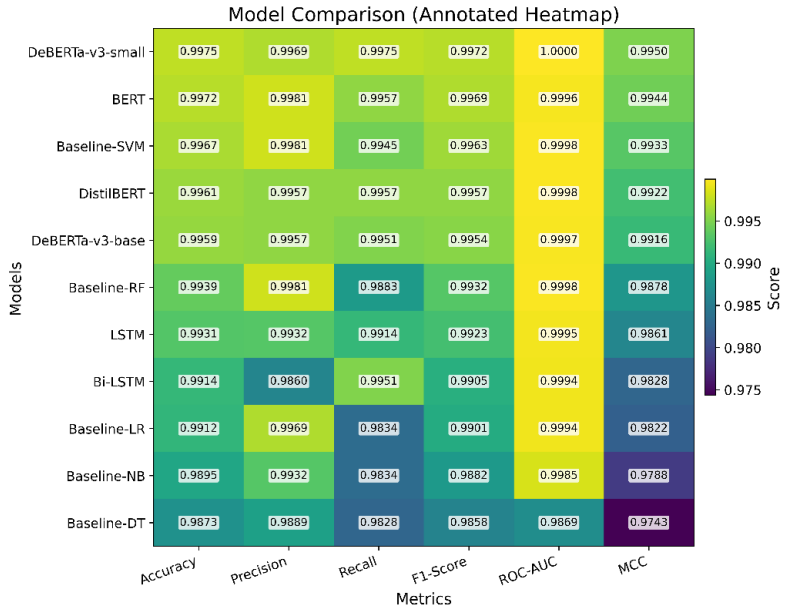

Figure 4: Consolidated metric overview for all models using an annotated heatmap.

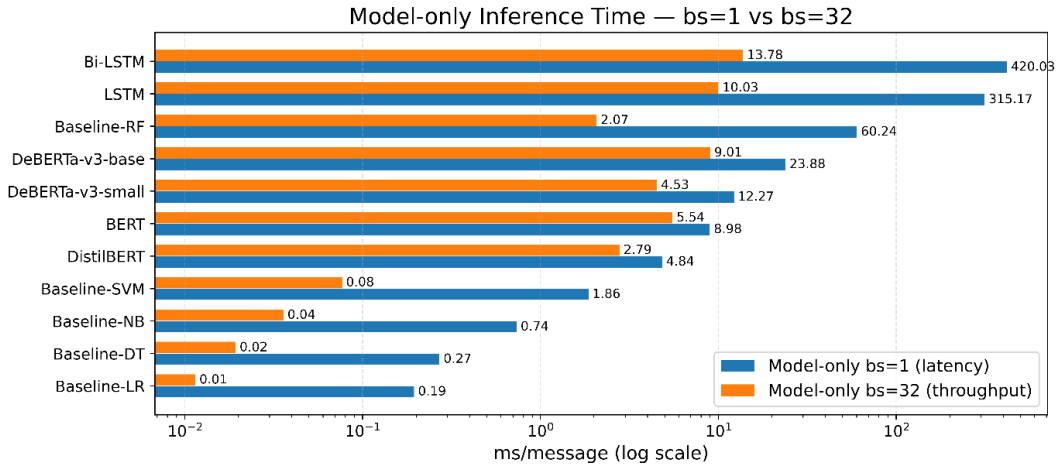

Figure 5: Model-only inference time (ms/message) at bs=1 and bs=32 (A100; log scale).

5. Conclusion

This Research-in-Progress paper establishes the subject+body text module as a robust foundation for our planned multi-segment detector, facilitating subsequent system extensions. DeBERTa-v3-small achieves the best overall effectiveness, while DistilBERT offers the most favorable efficiency–effectiveness trade-off. Model-only inference measurements at batch size 1 further support DistilBERT for latency-sensitive deployment (4.84 ms/message) compared with BERT (8.98 ms/message) and DeBERTa-v3-small (12.27 ms/message). Despite these strong results, text-only analysis may miss attacks that primarily appear in headers, obfuscated URLs (including shortened or QR-encoded links), or malicious attachments. Future work will therefore implement the remaining modules, evaluate their added value through ablation and fusion studies, and enrich the dataset to better represent evolving phishing tactics. Although the Enron and Nazario datasets are well-established benchmarks in phishing detection literature, they may not fully capture the latest phishing tactics or modern legitimate email patterns, and some of the observed separation may reflect corpus-specific patterns rather than broadly applicable phishing signals. Consequently, the near-perfect performance (F1 ≈ 0.99) should be treated as a controlled baseline, requiring additional validation on more diverse and up-to-date datasets to confirm the system's robustness and real-world applicability.

References

Aassal, A. E., Baki, S., Das, A., & Verma, R. M. (2020). An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs. *IEEE Access*, *8*, 22170-22192. https://doi.org/10.1109/ACCESS.2020.2969780

Anti-Phishing Working Group. (2024). *Phishing Activity Trends Report, 4th Quarter 2023*. https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf

Check Point Research. (2023). *Cyber Security Report*. https://resources.checkpoint.com/report/2023-check-point-cyber-security-report

Cohen, W. W. (2009). *Enron Email Dataset* (https://www.cs.cmu.edu/~enron/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers),

Federal Bureau of Investigation. (2022). *FBI Internet Crime Report 2022*. https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.,

He, P., Liu, X., Gao, J., & Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Kocher, G., & Kumar, G. (2021). Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges. *Soft Computing*, *25*(15), 9731-9763.

Kyaw, P. H., Gutierrez, J., & Ghobakhlou, A. (2024). A Systematic Review of Deep Learning Techniques for Phishing Email Detection. *Electronics*, *13*(19), 3823. https://www.mdpi.com/2079-9292/13/19/3823

Lee, J., Tang, F., Ye, P., Abbasi, F., Hay, P., & Divakaran, D. M. (2021, 6-10 Sept. 2021). D-Fence: A Flexible, Efficient, and Comprehensive Phishing Email Detection System. 2021 IEEE European Symposium on Security and Privacy (EuroS&P),

Maaten, L. v. d., & Hinton, G. E. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579-2605.

Muralidharan, T., & Nissim, N. (2023). Improving malicious email detection through novel designated deep-learning architectures utilizing entire email. *Neural Networks*, *157*, 257-279. https://doi.org/https://doi.org/10.1016/j.neunet.2022.09.002

Nazario, J. (2006). *Nazario Phishing Corpus* (https://monkey.org/~jose/phishing/

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*.

Verizon. (2022). *Data Breach Investigations Report*. https://www.phishingbox.com/downloads/Verizon-Data-Breach-Investigations-Report-DBIR-2022.pdf

Yamin, M. M., Ullah, M., Ullah, H., & Katt, B. (2021). Weaponized AI for cyber attacks. *Journal of Information Security and Applications*, *57*, 102722. https://doi.org/https://doi.org/10.1016/j.jisa.2020.102722