

Experimental Findings Report

Explainable Hybrid DeBERTa–XGBoost Architecture for Email Spam and Phishing Detection

1. Aim of the Experiment

The experiment aimed to investigate and evaluate whether the contextual representations learned by DeBERTa-v3-base could be effectively decoupled from its native classification stage and combined with XGBoost as a downstream classifier for email spam and phishing detection, while enabling useful and interpretable analysis of the resulting classification decisions through SHAP and LIME.

The investigation was therefore structured around three questions:

1. **Classification:** Can XGBoost effectively classify DeBERTa's learned representations, and does this produce measurable differences compared with standalone DeBERTa?
2. **Computational cost:** What computational benefit, if any, is obtained by separating Transformer-based representation generation from downstream classifier optimisation?
3. **Explainability:** Can SHAP and LIME provide useful insight into the decisions made by the hybrid model?

2. Research Gaps Addressed

The experiment was motivated by two interconnected research gaps in Transformer-based email spam and phishing detection. First, existing end-to-end Transformer architectures couple computationally intensive contextual representation learning with downstream classification, leaving a need to investigate whether these functions can be decoupled so that learned representations can be reused with a separate classifier and

downstream optimisation can be performed more efficiently. Second, the high-dimensional latent representations underlying Transformer decisions remain difficult to interpret, creating a need for complementary approaches that can investigate both the latent features influencing classification and the textual evidence associated with individual predictions. The study therefore investigated a DeBERTa-v3-base–XGBoost architecture, using SHAP for representation-level analysis and LIME for local text-level explanation.

2.1 Classification and Computational Cost

Transformer models such as DeBERTa are capable of learning rich contextual representations from email content. However, they are commonly deployed as end-to-end classifiers in which the same architecture performs both representation learning and final classification.

This creates an opportunity to investigate whether these two functions can be separated:

Email → DeBERTa-v3-base → 768-dimensional representation → XGBoost → Classification

The experiment therefore examined not only whether XGBoost could exploit the semantic information contained in DeBERTa's representations, but also whether separating the two stages could make downstream classifier training, optimisation and experimentation more computationally efficient without repeatedly processing the emails through the Transformer.

Importantly, the experiment did **not** assume that XGBoost would necessarily outperform DeBERTa's native classification head. The objective was to empirically determine whether a separate classifier could effectively exploit the representation already learned by the Transformer.

2.2 Explainability

Although Transformer models can achieve strong classification performance, their high-dimensional learned representations are difficult to interpret directly. In an email-security setting, identifying an email as malicious or legitimate is not always sufficient;

understanding the evidence contributing to a decision can be important for analysing false positives, false negatives and model behaviour.

The experiment therefore adopted a two-level explainability strategy:

- **SHAP** to investigate influential dimensions within the DeBERTa representation used by XGBoost.
- **LIME** to investigate influential textual features in individual email prediction

3. Experimental Finding 1 — Classification Efficiency

The first experimental question was whether DeBERTa's learned contextual representations contained sufficient discriminative information for a separate XGBoost classifier to distinguish legitimate and malicious emails.

The results demonstrate that they did. The 768-dimensional DeBERTa representations provided a highly informative feature space for downstream classification

Metric	Standalone DeBERTa	DeBERTa XGBoost →
Accuracy	99.40%	99.50%
Precision	97.73%	96.70%
Recall	95.56%	97.78%
F1-score	96.63%	97.24%
Balanced Accuracy	97.67%	98.72%
ROC-AUC	99.92%	99.98%
PR-AUC	99.38%	99.78%
MCC	96.31%	96.96%

The results indicate a **modest but measurable improvement** with the hybrid architecture rather than a substantial change in overall classification performance. The most notable improvement was observed in malicious-class recall, which

increased from **95.56% to 97.78%**, indicating that the hybrid model identified a greater proportion of malicious emails. This improvement was accompanied by a decrease in precision from **97.73% to 96.70%**, reflecting a small increase in false-positive classifications. Nevertheless, the hybrid model achieved higher **F1-score (96.63% → 97.24%)**, **balanced accuracy (97.67% → 98.72%)**, **ROC-AUC (99.92% → 99.98%)**, **PR-AUC (99.38% → 99.78%)**, and **MCC (96.31% → 96.96%)**, indicating an overall improvement across the principal classification metrics.

3.1 Security-Oriented Error Analysis

The confusion-matrix comparison provides an additional perspective:

Outcome	Standalone DeBERTa	DeBERTa XGBoost →
True Positives	86	88
False Negatives	4	2
False Positives	2	3
True Negatives	908	907

The hybrid model correctly identified **two additional malicious emails**, reducing false negatives from four to two. This improvement came at the cost of one additional false positive.

From a classification perspective, this indicates that XGBoost was able to **refine the final decision boundary using the semantic representation generated by DeBERTa**. At the same time, the relatively small magnitude of the improvement indicates that DeBERTa's representation was already highly discriminative and

that XGBoost primarily provided additional decision refinement rather than fundamentally transforming the classification capability.

The controlled XGBoost experiments further support this interpretation. Moderate changes to XGBoost complexity produced largely identical thresholded validation predictions, suggesting that the quality of the underlying DeBERTa representation was a major determinant of performance.

Summary of Finding 1

DeBERTa's contextual representations could be effectively combined with XGBoost as a separate classification layer, producing a modest improvement in malicious-email detection, particularly in recall. The hybrid framework reduced false negatives from four to two on the held-out test set, while increasing false positives from two to three.

3.2. Experimental Finding 2 — Computational Cost

The computational investigation produced a more nuanced result.

The hybrid architecture **did not make the DeBERTa component computationally cheaper**. DeBERTa remained the dominant computational component because every email still had to pass through the Transformer to generate its contextual representation.

Generation of the train, validation and test embeddings on the local CPU required approximately: **264.55 minutes**

However, once the 768-dimensional representations had been generated, the computational characteristics of the downstream classification stage were substantially different.

The important computational benefit therefore arises **after representation generation**. Once the DeBERTa embeddings have been generated and stored, different XGBoost configurations can be trained and compared without regenerating the Transformer representations.

This creates a modular workflow in which the expensive representation-learning stage can be separated from downstream classifier experimentation.

However, the experiment does **not** provide evidence that the complete hybrid system has a lower end-to-end training or inference cost than standalone DeBERTa. In particular, the Transformer remains necessary for generating representations for new emails during inference.

Summary of Finding 2

Separating DeBERTa representation learning from XGBoost classification made downstream classifier optimisation inexpensive and rapidly repeatable once the Transformer representations had been generated. However, the experiment did not demonstrate a reduction in the overall computational cost of the complete Transformer-based detection system

Consequently, the computational-cost research gap was experimentally investigated but only partially addressed as a practical cost-reduction solution.

3.3 Experimental Finding 3 — Explainability

The third component investigated whether the hybrid architecture could provide useful explanations of its classification behaviour.

The combination of SHAP and LIME provided two complementary perspectives:

SHAP → What components of the learned representation influenced the XGBoost decision?

The combination of SHAP and LIME provided two complementary perspectives:

SHAP → What components of the learned representation influenced the XGBoost decision?

LIME → What textual features were influential in a particular email prediction?

3.3.1 SHAP: Latent Representation-Level Analysis

SHAP was applied to the XGBoost classifier operating on the 768-dimensional DeBERTa representations.

The analysis showed that the 768 dimensions were not equally influential. The dimension with the largest mean absolute SHAP magnitude was:

DeBERTa_dim_440

Other highly influential dimensions included:

dim_129

dim_587

dim_230

dim_451

dim_555

dim_157

dim_510

dim_659

dim_303

This demonstrates that the downstream XGBoost classifier relied more heavily on certain components of the learned representation than others.

However, these dimensions cannot be directly equated with human-readable concepts. For example, it would not be scientifically justified to state that DeBERTa_dim_440 represents "phishing", "malicious links", or another specific semantic concept.

The interpretation :

Certain components of the learned DeBERTa representation exerted substantially greater influence on the downstream classification decision than others.

3.3.2 LIME: Text-Level Analysis

LIME provided the complementary text-level perspective by identifying influential words or tokens in individual emails.

Four representative cases were examined:

- a) True Positive
- b) True Negative
- c) False Positive
- d) False Negative

This allowed the analysis to investigate both successful classifications and classification errors.

The false-positive example was particularly informative because the email was labelled legitimate but contained strong promotional and spam-like characteristics. The model therefore encountered textual evidence that resembled malicious email.

Conversely, the false-negative example was labelled malicious but classified as legitimate. Its analysis indicated competing textual and latent signals, with legitimate-oriented evidence ultimately dominating the decision.

This provides a more informative security analysis than simply recording that a prediction was incorrect. It allows the investigation to ask:

What characteristics contributed to this particular classification decision?

4.0 Integrated Experimental Finding

The three strands of the investigation can be brought together into a single experimental conclusion:

The experiment demonstrated that DeBERTa-v3-base contextual representations can be effectively decoupled from the native classification stage and combined with XGBoost as a downstream classifier for email spam and phishing detection. The resulting hybrid architecture produced a modest improvement in held-out classification performance, particularly in malicious-email recall, reducing false negatives from four to two while increasing false positives from two to three.

The computational investigation showed that the separation did not reduce the computational cost of DeBERTa representation generation, which remained the dominant computational stage. However, once the representations had been generated, XGBoost enabled downstream classifier training and optimisation to be performed rapidly without repeatedly processing the emails through the complex Transformer embeddings representation process. This significantly improves operation cost.

Finally, SHAP and LIME provided complementary explainability mechanisms: SHAP exposed influential dimensions within the learned DeBERTa representation, while LIME provided case-specific textual evidence. The LIME robustness analysis demonstrated that although model predictions remained stable, individual textual feature rankings were less stable and should therefore be treated as local and illustrative rather than definitive explanations.

5. Overall Contribution of the Experiment

The experimental contribution can consequently be expressed through three connected findings.

5.1 Hybrid Classification

The study demonstrated the **feasibility of separating Transformer representation learning from downstream classification**, with XGBoost successfully exploiting the resulting semantic representations.

5.2 Security-Oriented Performance

The hybrid produced a **modest improvement in malicious-email detection**, particularly recall, with false negatives decreasing from four to two on the held-out test set.

5.3 Multi-Level Explainability

The study demonstrated a complementary XAI framework:

- A. **SHAP** — latent representation-level analysis;
- B. **LIME** — local text-level analysis;
- C. **LIME robustness testing** — assessment of the stability of textual explanations.

6.0. Summary and Conclusion

6.1 Summary: The experiment supports the conclusion that: DeBERTa provides a strong semantic representation, XGBoost can effectively exploit that representation and modestly refine the classification decision, while SHAP and LIME provide complementary mechanisms for investigating the resulting model behaviour.

6.2. Conclusion

The experiment provides evidence that **Transformer representation learning and downstream classification can be effectively decoupled** in an email spam and phishing detection architecture. DeBERTa-v3-base generated rich contextual representations that were sufficiently discriminative for XGBoost to perform the downstream classification task, producing a modest improvement over standalone DeBERTa on the held-out test set.