International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(20s) 2026 pp. 1910-1933
https://doi.org/10.70917/ijcisim-2026-6114
Article

A Hybrid Transformer and Rule-Based Linguistic Framework for Phishing Detection and Manipulation Tactics Identification

Eman Ghannam AL-Ruwaili[1], Randa Allafi[2]

[1]Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia, emangd22@gmail.com
[2]Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia
*Corresponding author: emangd22@gmail.com

Abstract: Phishing attacks exploit linguistic manipulation, including urgency, authority signalling, impersonation, and persuasive framing, across email and SMS. This manuscript describes a two-stage framework coupling transformer-based binary classification with a rule-based Linguistic Manipulation Detection Tool (LMDT) that returns tactic spans and a configurable Linguistic Manipulation Score (LMS). Before splitting, the 2,000-message cross-channel corpus is grouped by near-duplicate and template similarity so that no shared-template family crosses the train/validation/test boundary; on the resulting test set (n=300), fine-tuned BERT reaches the highest classification performance (macro F1 97.67%), significantly ahead of a Random Forest baseline and of RoBERTa (Holm-adjusted exact McNemar, p=0.021 and p=0.039), while the deployed hybrid pipeline continues to gate the LMDT stage on RoBERTa (macro F1 94.99%, 95% CI [92.32, 97.33]), the backbone specified before the BERT comparison was run; replacing it with BERT is the first item of future work. The LMDT is evaluated against a 120-message stratified subsample of the frozen test set, independently labelled by two human annotators with disagreements resolved by adjudication (Cohen kappa reported per tactic, Appendix B): on this ground truth the original rule set outperforms a proposed refinement to the authority and impersonation detectors, which eliminates their false positives on legitimate messages but also eliminates their recall, so the original rules are retained in the deployed system. Per-message predictions, 95% bootstrap confidence intervals, and exact McNemar tests with Holm correction are reported for every classifier comparison. The LMS is reported only as a descriptive aggregation of the four detected tactics: with no independently rated severity outcome available, its weights remain unvalidated design choices rather than an evidenced risk score.
Keywords: phishing detection; RoBERTa; manipulation tactics; spaCy; tactic annotation; NLP; cybersecurity; hybrid framework; Linguistic Manipulation Score

1. Introduction

Phishing accounts for a large share of all reported cybersecurity breaches. Messages are crafted to trigger impulsive decisions; no software vulnerability need be present—a property that makes rule-based filters insufficient against novel campaigns [1]. Four recurring manipulation patterns structure most phishing content: urgency (time pressure compelling immediate action), authority (invoked institutional trust), impersonation (brand or person spoofing), and persuasion (reward framing or social proof) [2, 3]. Detection systems that produce only a phishing/legitimate verdict forgo a chance to teach users why a message is dangerous.

Transformer language models—BERT and RoBERTa in particular—have set new performance standards on short adversarial texts, capturing the semantic deceptive patterns that bag-of-words representations miss [4,5]. Most deployed transformer detectors remain opaque: the decision surfaces as a probability score with no linguistic grounding.

This paper makes three contributions. First, we assemble a balanced cross-channel evaluation set of 2,000 English messages from public sources, group near-duplicate and template-shared messages before splitting so that no template family crosses a partition boundary, and generate four tactic pseudo-labels with a deterministic Qwen3-4B procedure for the training and validation splits; two independent human annotators additionally label a 120-message stratified subsample (60 phishing, 60 legitimate) of the frozen 300-message test set, with inter-annotator agreement reported per tactic and disagreements resolved by adjudication, providing ground truth for the LMDT evaluation. Second, we fine-tune RoBERTa and BERT for binary phishing classification and compare both against two TF-IDF baselines under identical training conditions, with per-message predictions retained and exact McNemar tests reported for every pairwise comparison. Third, we build the LMDT, a deterministic, auditable rule engine that annotates messages classified as phishing with tactic-level evidence spans and a composite LMS, and we evaluate a proposed refinement to two of its four detectors against the same human-labelled test set used for the original rules.

2. Related Work

2.1 Traditional Classifiers

TF-IDF representations fed into SVMs or Random Forests established early automated phishing detection, reaching accuracies above 95% on constrained benchmarks [6, 7]. Dual-pipeline and ensemble approaches—combining URL features with content analysis—have demonstrated complementary detection capacity across message types [8, 9]. Performance degrades when phishers rephrase messages to evade known keyword patterns [10]. URL-based classifiers applying SVM with minimal lexical features report similar ceilings on constrained testbeds [11, 12]. Comparative studies consistently rank TF-IDF+SVM as a strong classical baseline [5], a finding partly confirmed here: SVM reaches 95.33% macro F1 on the leakage-resistant test set, essentially matching RoBERTa (94.99%) and trailing only fine-tuned BERT (97.67%) among the models compared.

2.2 Transformer-Based Detection

Fine-tuned BERT and RoBERTa models dominate recent phishing benchmarks. Meléndez et al. [5] reported RoBERTa at 99.43% accuracy on a large email corpus; Ibrahim et al. [4] replicated RoBERTa's advantage on the Nazario+Enron split. Cross-channel evaluation covering both email and SMS within one experiment is less common than single-domain studies [13, 14]. Several independent benchmarks confirm BERT variants outperform classical methods across evaluation settings [15, 16]. Intelligent phishing detectors evaluated on the CEAS_08 corpus confirm deep learning gains over bag-of-words approaches [17, 18]. Hybrid architectures that couple TF-IDF feature engineering with transformer embeddings have also demonstrated gains over either approach alone [19, 20]. Recent related studies also examine cross-channel NLP detection [30], multi-dataset evaluation [31], compact transformer comparisons [32, 33], BERT/LLM benchmarks [34, 35], XLM-RoBERTa feature fusion [36], and URL/HTML-assisted phishing detection [37].

2.3 Manipulation Tactics

Framing phishing content in terms of recurring manipulation-tactic categories has informed social engineering detection research and computational threat analysis [2, 21]. Computational approaches using spaCy pattern matching for security-relevant text appear in threat intelligence extraction [21] and

social engineering detection [2, 3]. Language model fine-tuning for domain-specific phishing variants—including ransomware lures—likewise leverages transformer rule integration [22]. Our LMDT extends this line with four phishing-specific detectors and a weighted composite score.

2.4 Explainable Phishing Detection

Post-hoc methods (LIME, SHAP, attention visualisation) have been applied to phishing and adjacent security classification tasks [23, 24], including deep learning for email threat intelligence [25, 26]. DistilBERT variants integrated with XAI span highlighting have shown improved user trust in automated filtering [27]. Acharya et al. [2]

combined lightweight transformers with token-level XAI highlights and found users preferred the explanations over bare labels. Our approach returns character-span evidence by design, without post-hoc approximation—a deterministic, auditable alternative suited to deployment environments where model-specific explanation methods may not be available.

3. Methodology

3.1 Corpus

The corpus draws from four public sources: phishing emails from the Nazario Phishing Corpus and the Kaggle Phishing Email Dataset; legitimate emails from the Enron corpus; and SMS messages from the SMS Smishing Collection. After deduplication, 2,000 messages remain: 1,000 phishing and 1,000 legitimate, with 1,600 emails (80%) and 400 SMS (20%), as detailed in Table 1. Before splitting, messages are grouped at three levels of strictness: exact text match, a template key that normalises URLs, email addresses, and phone numbers, and character n-gram (4-5) cosine similarity at a 0.85 threshold with transitive closure, so that near-duplicate or template-shared messages are assigned to the same partition. In this corpus, exact and template matching alone found no duplicates, but the near-duplicate criterion identified 44 multi-message groups covering 180 messages (6.8% of the corpus); the 70/15/15 split is stratified on channel and class over these groups rather than over individual messages, so no template family crosses the train/validation/test boundary. Short message channels introduce harder detection conditions due to limited lexical surface [28, 29].

Table 1: Class-by-channel corpus composition (n=2,000). All splits preserve the 80/20 email-to-SMS ratio and 50/50 class balance and are additionally stratified over 1,864 near-duplicate groups (Methodology): 1,820 singleton messages plus 44 multi-message groups covering 180 messages, rather than over individual messages.

Channel	**Phishing**	**Legitimate**	**Total**
Email	800	800	1,600
SMS	200	200	400
Total	1,000	1,000	2,000

Four binary tactic labels (urgency, authority, impersonation, persuasion) are assigned by deterministic zero-shot annotation with Qwen3-4B (Alibaba Cloud open-weight LLM; temperature 0, prompt identifier lmdt_qwen3_v1.1; repeated runs on the same input produced identical outputs in our implementation) for the training and validation splits. These pseudo-labels support dataset construction and the choice between rule versions on the validation split, where no human labels exist. The LMDT is scored against human labels: two annotators independently applied the same four tactic definitions (Appendix B) to a 120-message stratified subsample of the 300-message test split, legitimate messages included, and resolved their disagreements by adjudication. Every LMDT result reported on the test side is scored against these adjudicated labels. The 300-message test set is frozen and hash-verified in the notebook; the remaining 180 test messages carry no human labels.

Fig. 1 shows the two-stage system overview; Fig. 2 details the LMDT pipeline.

3.2 Stage 1 — RoBERTa Classifier

roberta-base (125M parameters) is fine-tuned with a [CLS]-pooled linear head. Training: AdamW, lr 2×10−5 with 10% linear warm-up, batch size 16, max sequence length

256 tokens, early stopping (patience 3, monitored on validation macro F1). Identical hyperparameters apply to BERT

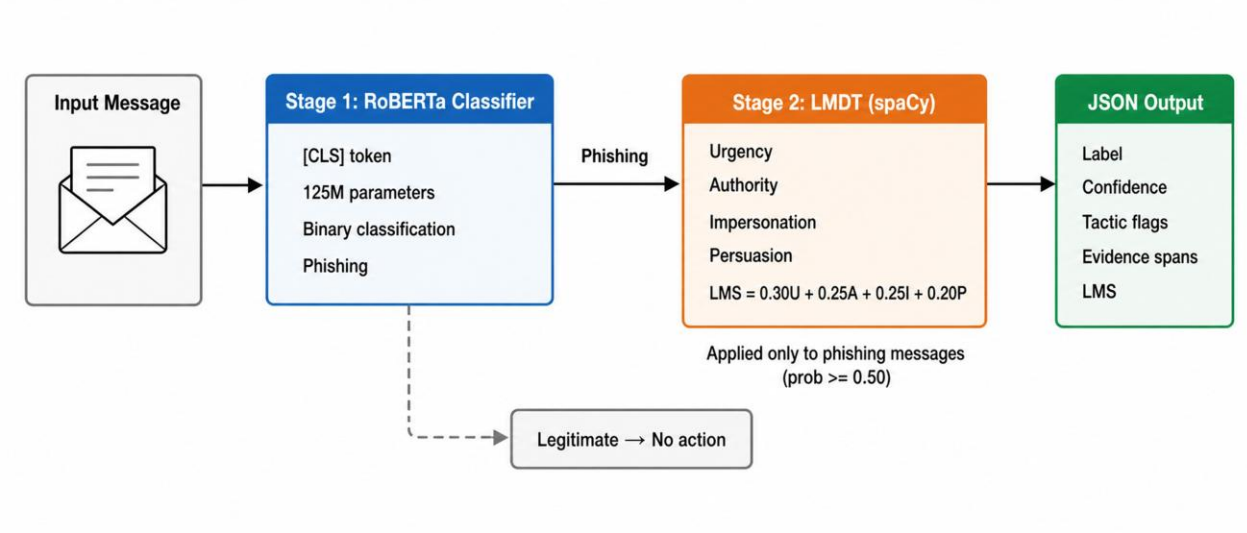


Figure 1: Two-stage hybrid phishing detection architecture. Stage 1 (RoBERTa) classifies all incoming messages; Stage 2 (LMDT) annotates confirmed phishing messages with tactic spans and an LMS.

(bert-base-uncased, 110M parameters). TF-IDF baselines use a linear SVM (C=1, linear kernel) and a 300-estimator Random Forest (max depth 50), both selected by grid search on the training split. All models train on 1,400 messages.

3.3 Stage 2 — LMDT

Messages classified as phishing (probability ≥ 0.50) enter the LMDT, which applies four independent rule sets implemented as regular expressions over the raw message text. spaCy is used for tokenisation only and contributes no parsing or entity recognition to any rule.

Urgency targets temporal quantifiers, modal imperatives, and negative-consequence constructions.

Authority (v1, deployed) matches a bare institution or role keyword (bank, government, IT department, compliance office, and comparable roles). A proposed refinement (v2) requires the term to share a sentence with, and sit within 120 characters of, a directive, compliance, verification, or consequence cue, or to appear inside an institutional self-identification; role mentions introduced by a personal possessive (e.g. "my manager") are excluded. v2 was evaluated against the human-annotated reference (Results; Appendix F) and is not adopted in the deployed system, which retains v1.

Impersonation (v1, deployed) matches a first-person organisational claim, an on-behalf-of claim, a named-brand mention with or without a role suffix, or a generic salutation (e.g. "dear customer") — the original (v1) rule set. A proposed refinement (v2) instead requires a brand paired with a service or department role, a first-person identity claim, or a forged sender address to be sufficient on its own, or a bare brand token, generic salutation, or automated-mail phrase to co-occur with an explicit account-action request; it was evaluated against the same ground truth (Appendix F) and is not adopted in the deployed system.

Persuasion covers reward framing, scarcity indicators, and social-proof constructions.

Each detector returns a binary flag and character-level evidence spans.

3.4 Linguistic Manipulation Score

$$LMS = 0.30\,U + 0.25\,A + 0.25\,I + 0.20\,P, \quad (1)$$

where U, A, I, P ∈ {0, 1}. The weights are heuristic design parameters, not estimated from data (Appendix E); urgency (U) receives the highest coefficient (0.30)

reflecting its reported status as the most pervasive temporal cue in phishing content [3, 21]. The LMS is configurable: operators may substitute empirically derived weights for deployment-specific tactic distributions. Because each component is binary, the score can take at most 16 distinct values (nine were observed on this study's test corpus), so it is a coarse indicator rather than a continuous risk measure (Discussion). LMS ∈ [0.00, 1.00].

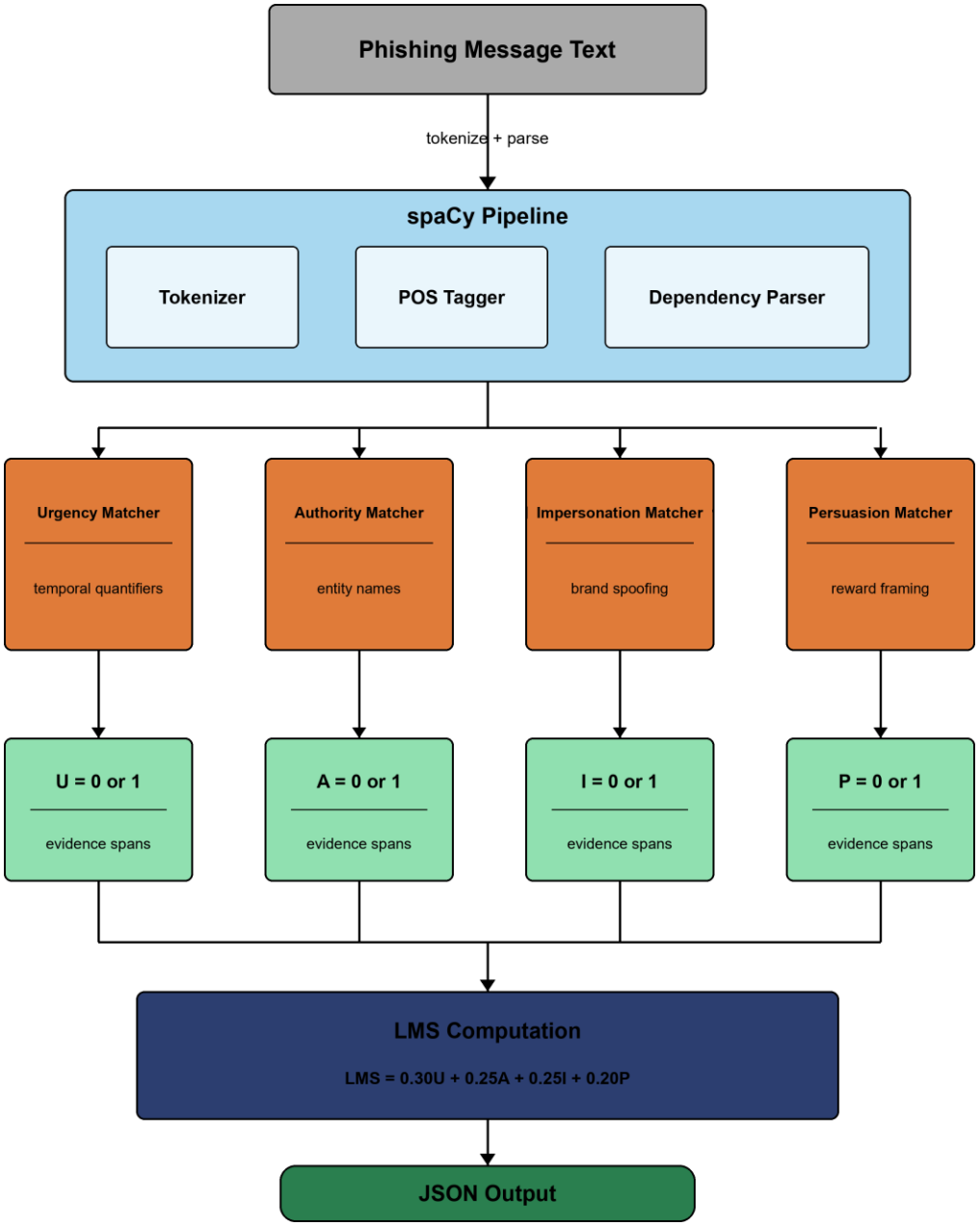


Figure 2: LMDT internal pipeline. spaCy provides tokenisation only; the four tactic detectors are regular-expression rule sets applied to the raw message text (Methodology), and their flags and evidence spans are aggregated into the LMS.

4. Experimental Setup

The corpus is partitioned 70/15/15: train 1,400 / val 300 / test 300 (1,400+300+300=2,000), stratified on channel and class over near-duplicate groups (Methodology), preserving the 80/20 email-to-SMS ratio and 50/50 class balance across all splits. The test set was withheld from all hyperparameter decisions and from LMDT rule-version selection, which is made on the validation split only, scored against the Qwen3-4B pseudo-labels that are the only labels available for that split (Table F3).

Primary metrics are macro-averaged precision, recall, F1, and accuracy. Secondary metrics are AUC-ROC, MCC, Cohen's kappa, and average precision. LMDT tactic metrics are reported on a 120-message stratified subsample (60 phishing, 60 legitimate) of the 300-message frozen test set, independently labelled by two human annotators; inter-annotator agreement (Cohen's kappa per tactic) and the adjudication protocol used to resolve disagreements are reported in Appendix B. Per-message predictions are retained for every classifier; 95% confidence intervals use a 2,000-resample percentile bootstrap (Wilson intervals for rate metrics), and every pairwise comparison uses an exact two-sided McNemar test with Holm-Bonferroni correction across the family of comparisons. All timing benchmarks were measured in the reported Colab GPU environment (NVIDIA L4).

5. Results

5.1 Phishing Classification

Table 2 reports test-set performance (macro F1 comparison in Fig. 3). Stage 2 adds tactic annotation but does not alter the binary classification decision from Stage 1, so the hybrid's classification metrics are identical to RoBERTa's alone and are omitted from Fig. 3 to avoid a redundant bar.

On the leakage-resistant, near-duplicate-grouped test split, BERT reaches the highest scores (F1 97.67%, AUC 99.66%, MCC 0.954), ahead of TF-IDF+SVM (F1 95.33%), RoBERTa (F1 94.99%, 95% CI [92.32, 97.33]), and TF-IDF+RF (F1 94.00%). Exact McNemar tests with Holm-Bonferroni correction across all six pairwise comparisons find BERT significantly ahead of TF-IDF+RF ($p=0.021$) and of RoBERTa ($p=0.039$); no other pair separates significantly at the 0.05 level, including BERT versus TF-IDF+SVM ($p=0.474$). Discordant counts (b = BERT correct only, c = other model correct only) are 12 and 1 against TF-IDF+RF, 8 and 0 against RoBERTa, and 11 and 4 against TF-IDF+SVM. The deployed hybrid pipeline gates the LMDT stage on RoBERTa's decision, consistent with the architecture proposed for this work; BERT's advantage on this split is discussed below as a candidate for a future revision of the hybrid's Stage 1 component. The RoBERTa confusion matrix records TN=150, FP=0, FN=15, TP=135 across the 300 test messages (BERT: TN=150, FP=0, FN=7, TP=143). Per-message predictions for every classifier are retained, making every reported test and interval reproducible from the stored predictions alone.

Table 2: Test-set performance on the leakage-resistant, near-duplicate-grouped split (n=300). Bold = best. F1 confidence intervals: 95% percentile bootstrap, 2,000 resamples.

Model	Acc.	F1	AUC	MCC	κ	F1 95% CI
TF-IDF+RF	94.00	94.00	98.32	0.881	0.880	[91.28, 96.33]
TF-IDF+SVM	95.33	95.33	99.39	0.907	0.907	[92.95, 97.66]
BERT	**97.67**	**97.67**	**99.66**	**0.954**	**0.953**	[95.97, 99.33]
RoBERTa	95.00	94.99	99.27	0.905	0.900	[92.32, 97.33]

Table 3: RoBERTa per-class precision, recall, and F1 on the leakage-resistant test set (n=300); Cohen's kappa (0.900) is a single classifier-level statistic, reported once.

Class	Precision	Recall	F1	Cohen kappa

Phishing	100.00	90.00	94.74	0.900
Legitimate	90.91	100.00	95.24	—

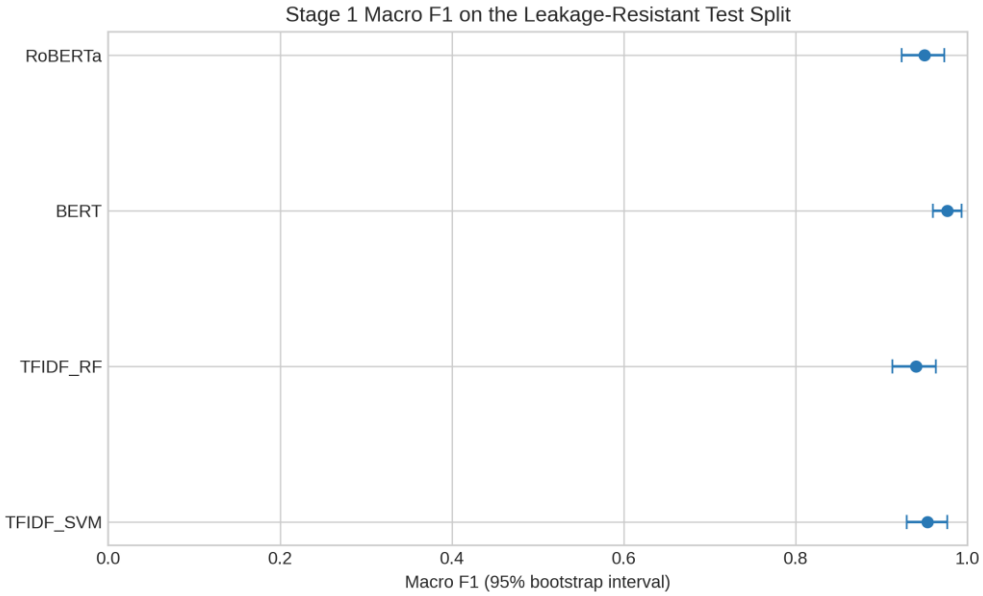


Figure 3: Macro F1 with 95% bootstrap confidence intervals for the four classifiers, on the leakage-resistant test set. The hybrid's classification decision is identical to RoBERTa's (94.99%) and is not plotted separately; BERT achieves the highest macro F1 (97.67%).

5.2 Cross-Channel Performance

RoBERTa reaches identical macro F1 (94.99%) on the email and SMS subsets of the leakage-resistant test split, with zero false positives in both channels; AUC is somewhat higher for email (99.63%) than for SMS (97.44%), which is consistent with the smaller SMS subset (n=60) producing a noisier estimate rather than a systematic channel effect.

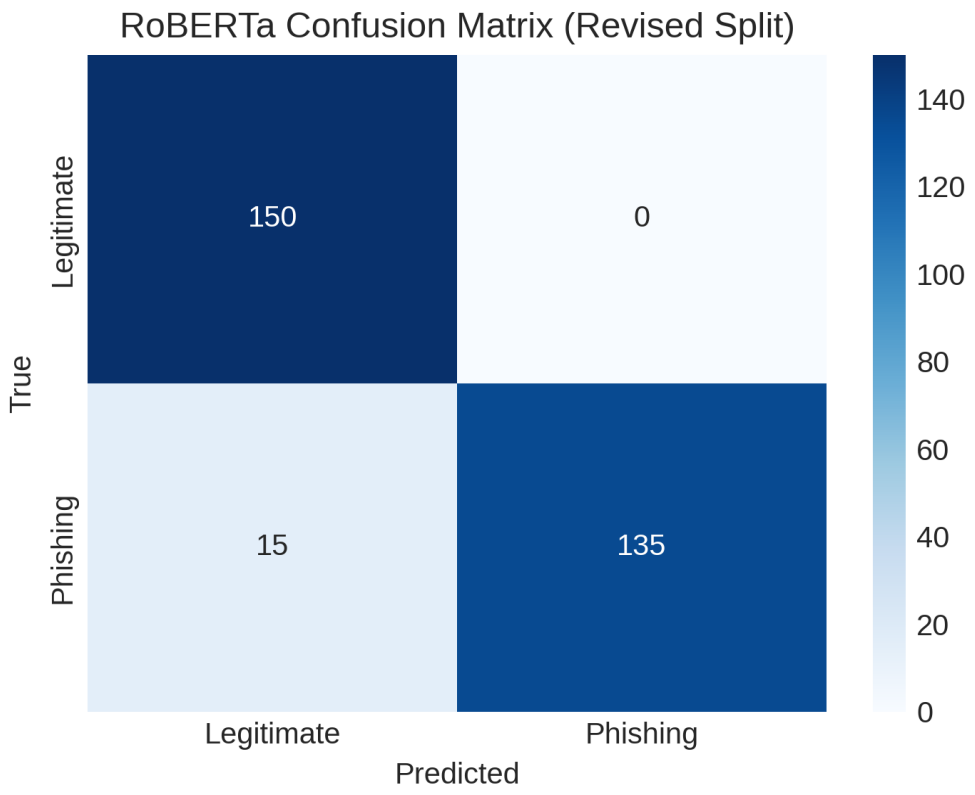


Figure 4: RoBERTa confusion matrix on the leakage-resistant test set (n=300): TN=150, FP=0, FN=15, TP=135.

Table 4: RoBERTa cross-channel results on the leakage-resistant test split.

Channel	**n**	**F1**	**Spec.**	**Sens.**	**AUC**
Email	240	94.99	100.00	90.00	99.63
SMS	60	94.99	100.00	90.00	97.44
All	300	94.99	100.00	90.00	99.27

5.3 Training Dynamics

RoBERTa's validation macro F1 peaks at epoch 3 (0.970) and training halts at step 528 (epoch 6 of the 10-epoch budget) once macro F1 fails to improve for three consecutive epochs; BERT's validation macro F1 first reaches its plateau (≈0.967) at epoch 2 and training halts at step 616 (epoch 7). Rapid phishing-domain convergence within the first few epochs has been observed across fine-tuning studies [38, 39].

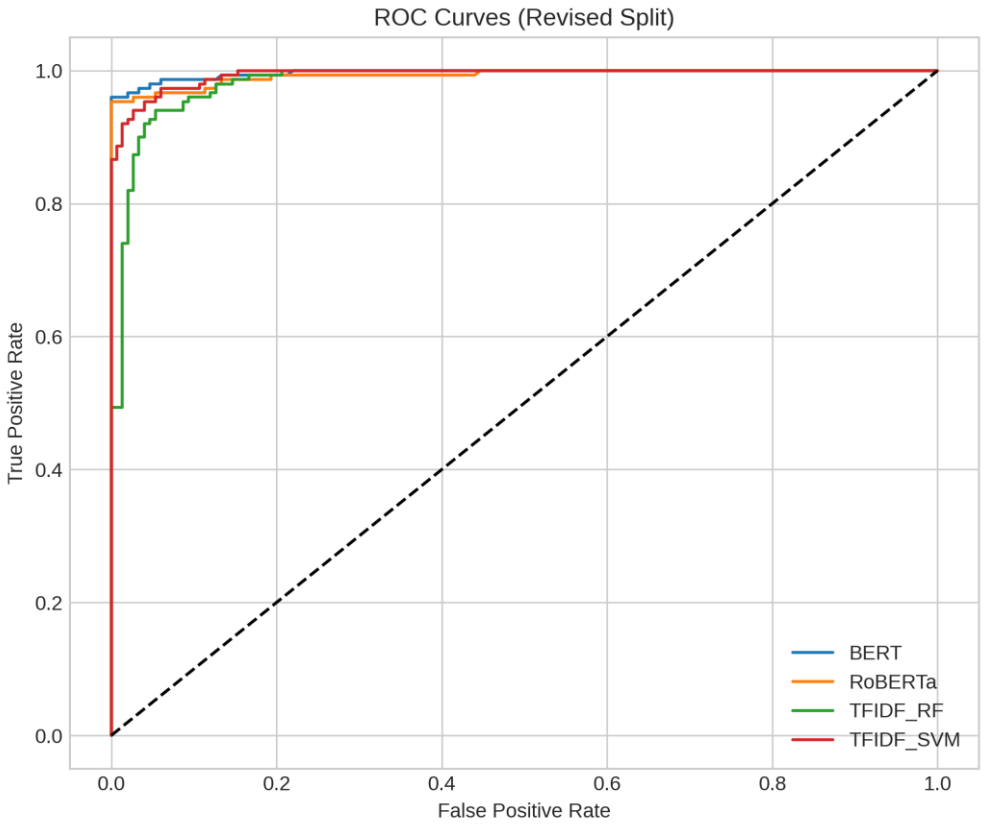


Figure 5: ROC curves on the leakage-resistant test set; RoBERTa AUC 0.9927, BERT AUC 0.9966.

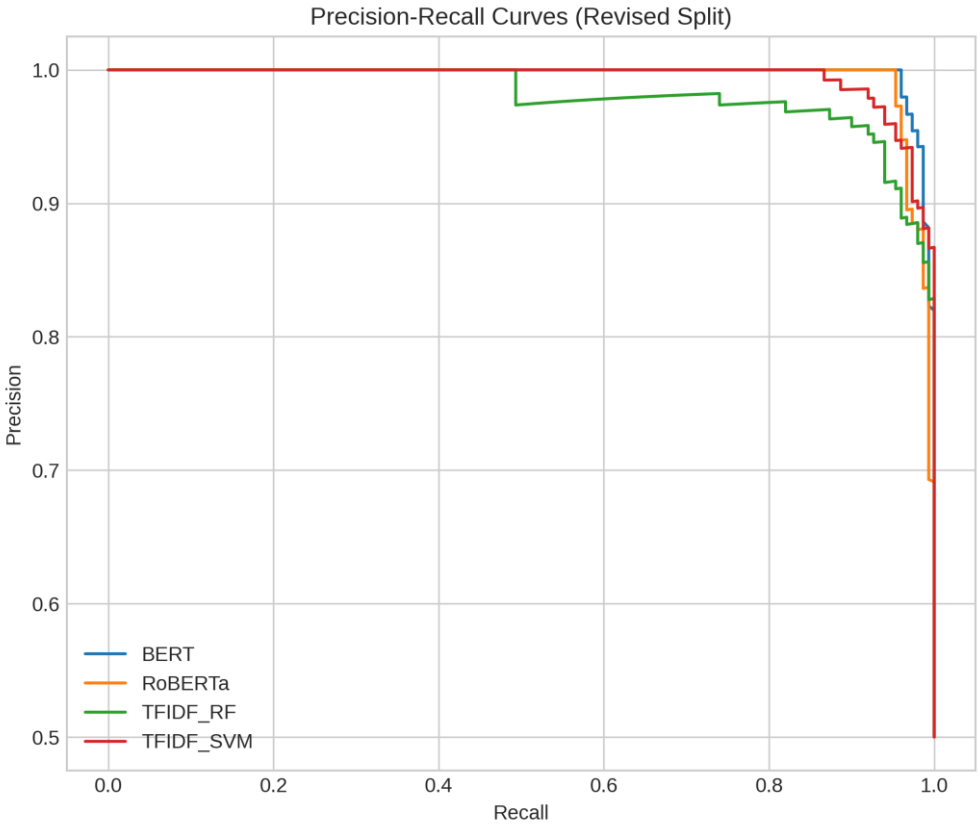


Figure 6: Precision-recall curves on the leakage-resistant test set.

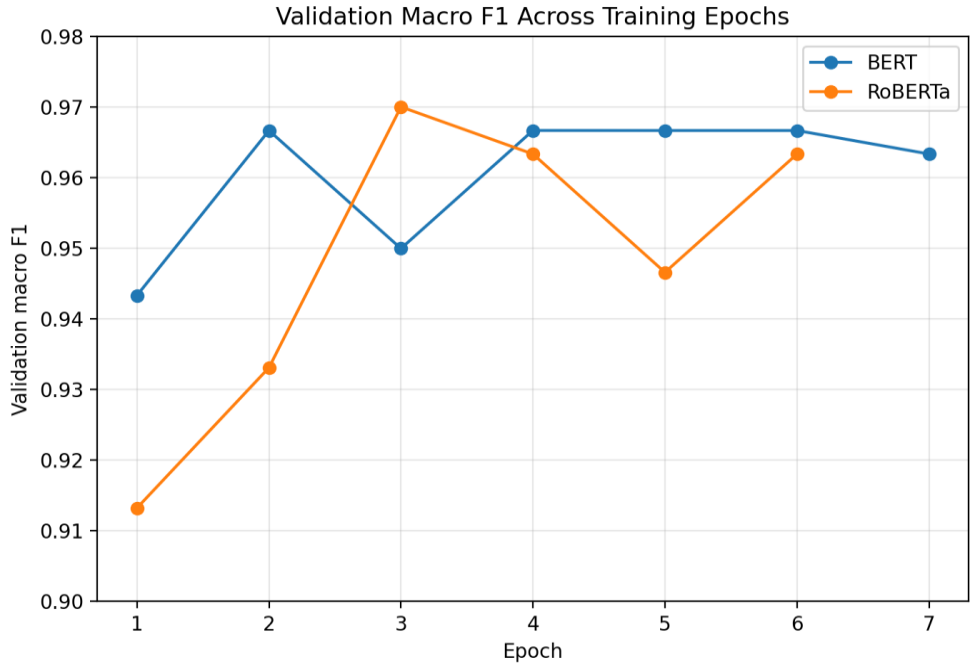


Figure 7: Validation macro-F1 across training epochs, this run. RoBERTa peaks at epoch 3 (0.970) and training halts at epoch 6; BERT plateaus by epoch 2 (≈0.967) and training halts at epoch 7.

5.4 LMDT Tactic Detection (Oracle-Gated Intrinsic Evaluation)

Table 5 reports LMDT performance against the adjudicated labels of two independent human annotators on a 120-message stratified subsample of the frozen test set (60 phishing, 60 legitimate; Appendix B), for the deployed rule set (v1). The proposed Authority and Impersonation refinement (v2, Appendix F) is evaluated separately below and is not adopted in the deployed system.

LMS agreement with the human-derived reference aggregation: the LMS computed from LMDT-detected tactics shows a Spearman correlation of 0.337 (95% CI [0.174, 0.503], p<0.001; 2,000-resample bootstrap) with the same aggregation computed from the adjudicated human labels at the paper's own weights, on the 120-message annotated subsample (Appendix B). The confidence interval excludes zero, indicating a moderate positive association between the two scores. The same computation against a Qwen3-4B pseudo-label aggregation gives a Spearman correlation of 0.326. The ranking the LMS induces over messages is comparatively stable across the same 2,000 weightings (Kendall tau fifth percentile 0.829, Figure 9), but this stability is a property of the weights, not evidence that the score tracks a real severity construct (Appendix E).

Table 5: LMDT (v1, deployed) evaluation against two-annotator adjudicated human ground truth (n=120 stratified subsample: 60 phishing, 60 legitimate, of the 300-message frozen test set; Cohen's kappa and the adjudication protocol are reported in Appendix B). n+ = human-positive count per tactic; Prec+, Rec+, F1+ = positive-class precision, recall, F1; BalAcc = balanced accuracy = (Sensitivity + Specificity)/2. All counts sum to n=120 per row. v1 was selected over the proposed refinement v2 on the validation split (Table F3); the same comparison on this annotated subsample is reported immediately below.

Tactic	n+	TP	FP	FN	TN	Prec+	Rec+	F1+	BalAcc	MCC
Urgency	13	9	15	4	92	0.375	0.692	0.486	0.776	0.429
Authority	3	1	6	2	111	0.143	0.333	0.200	0.641	0.188
Impersonation	7	0	5	7	108	0.000	0.000	0.000	0.478	-0.052
Persuasion	22	15	29	7	69	0.341	0.682	0.455	0.693	0.310
Macro avg.	---	---	---	---	---	0.215	0.427	0.285	---	---

On this annotated subsample, the comparison between v1 and v2 diverges by tactic once McNemar's test (overall correctness) is read alongside the bootstrap interval on positive-class F1 (tactic detection specifically). For impersonation, v2 is more often correct overall (accuracy 94.2% vs 90.0%, McNemar b=0, c=5, raw p=0.063, Holm-adjusted p=0.125, not significant) because it never fires on this tactic at all (0 false positives, vs 5 for v1), even though neither version recovers a meaningful positive-class F1 (both 0.0%; v1 misses 7 of 7 human-flagged instances, v2 the same 7 by construction). For authority, the same pattern holds: v2 reaches higher overall accuracy (97.5% vs 93.3%, McNemar b=1, c=6, raw p=0.125, Holm-adjusted p=0.125, not significant) by eliminating v1's false positives, but at the cost of all recall (F1 drops from 20.0% to 0.0%; only 3 human-positive instances exist for this tactic in the sample, so this difference is not statistically resolvable here and should be read descriptively). Read together, v2 trades away v1's already-modest recall on both tactics for a quieter false-positive profile; it is not a clear improvement on either, v1 retains materially more recall on both and is the version selected on the validation split (Methodology) and deployed in the hybrid, with v2 documented here as a negative result. The per-tactic comparison with confidence intervals and discordant counts is in Table F1 (Appendix F).

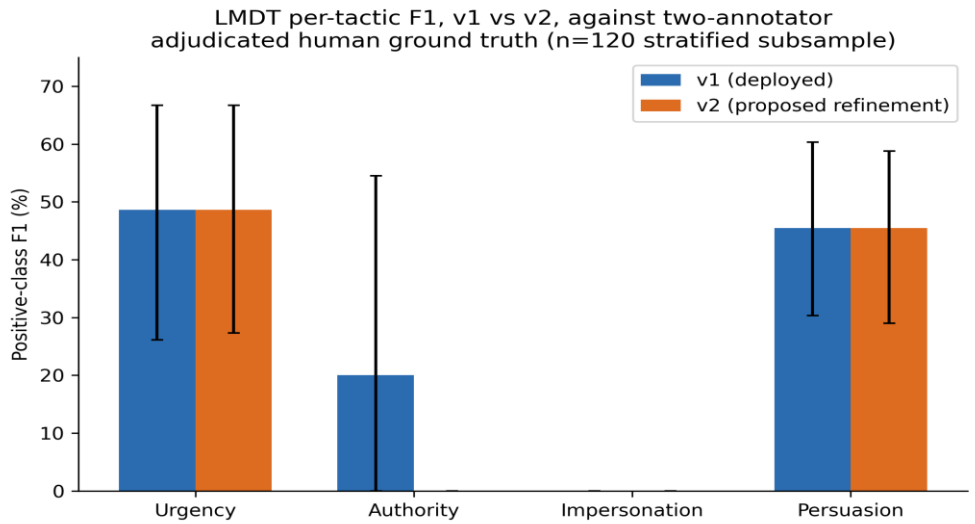


Figure 8: LMDT per-tactic F1, original (v1) versus revised (v2) rules, against two-annotator adjudicated human ground truth on the 120-message stratified subsample (60 phishing, 60 legitimate).

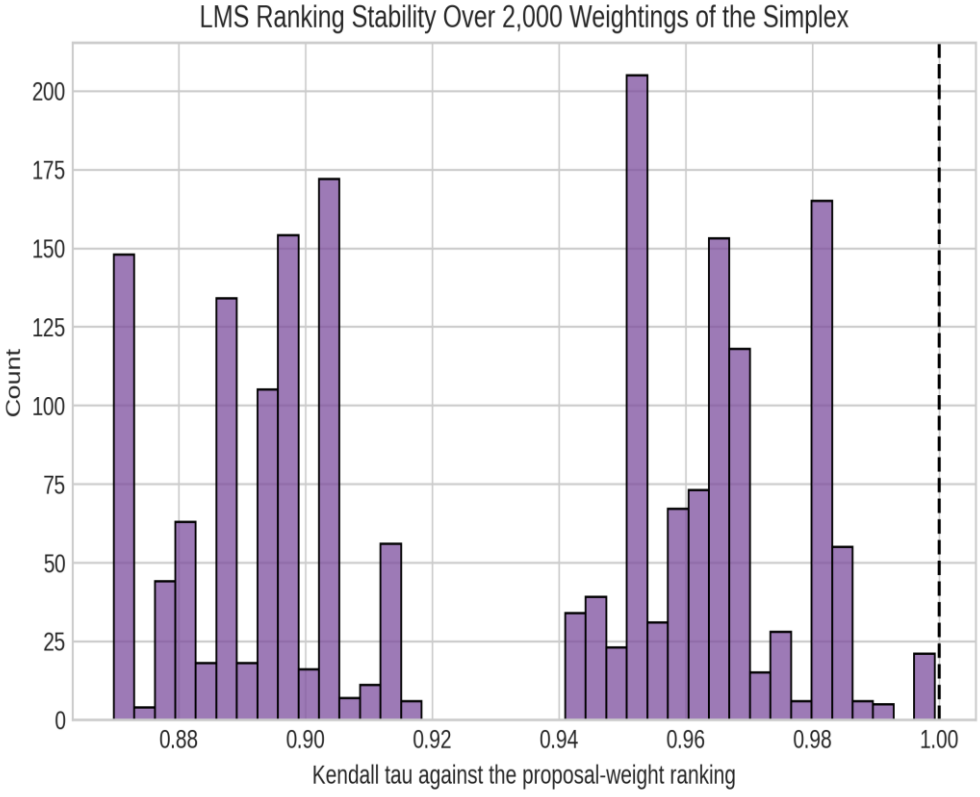


Figure 9: Stability of the LMS message ranking across 2,000 sampled non-negative weightings (Kendall tau against the paper's own weights; 5th percentile 0.829). A value of 1.0 indicates no change in ranking.

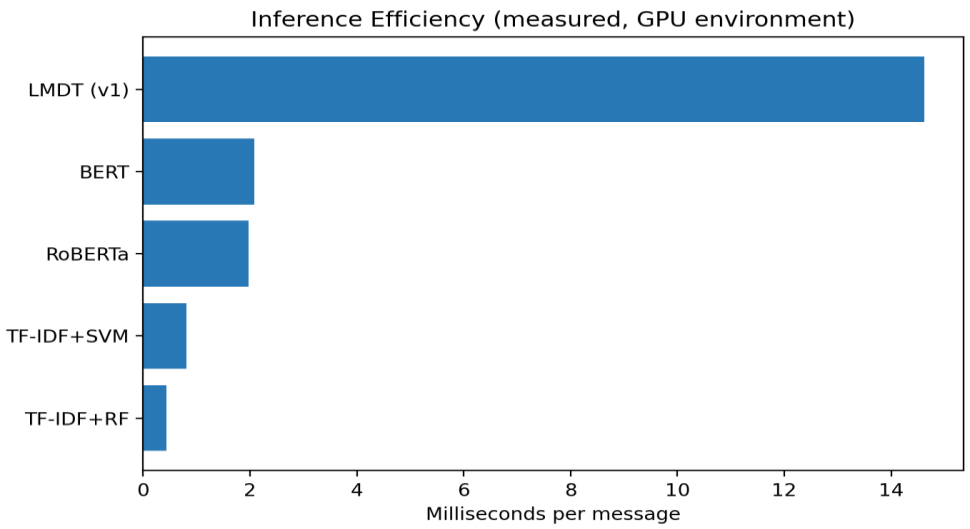


Figure 10: Inference speed (ms/message) in the reported Colab GPU environment; LMDT cost applies only to messages Stage 1 classifies as phishing. See Table 6 for exact values from this run.

Table 6: Training and inference efficiency on the leakage-resistant split, measured in the reported Colab GPU environment (NVIDIA L4).

Component	**Train (s)**	**Infer (ms/msg)**
TF-IDF+RF	1.51	0.436
TF-IDF+SVM	7.94	0.815
BERT	99.43	2.086
RoBERTa	92.40	1.970
LMDT (v1, deployed)	---	14.625
Hybrid (RoBERTa+LMDT v1)	92.40	16.595

5.5 Threshold Sensitivity

We adopt θ=0.50 as an a priori default threshold. Fig. 11 plots precision, recall, and macro F1 across the full threshold range as a post-hoc reference. In this run, precision stays at 1.000 across every threshold from 0.30 to 0.70; recall (and therefore macro F1) is unaffected until θ=0.65, where it drops from 0.900 to 0.887, a one-point change that raising θ further does not reverse.

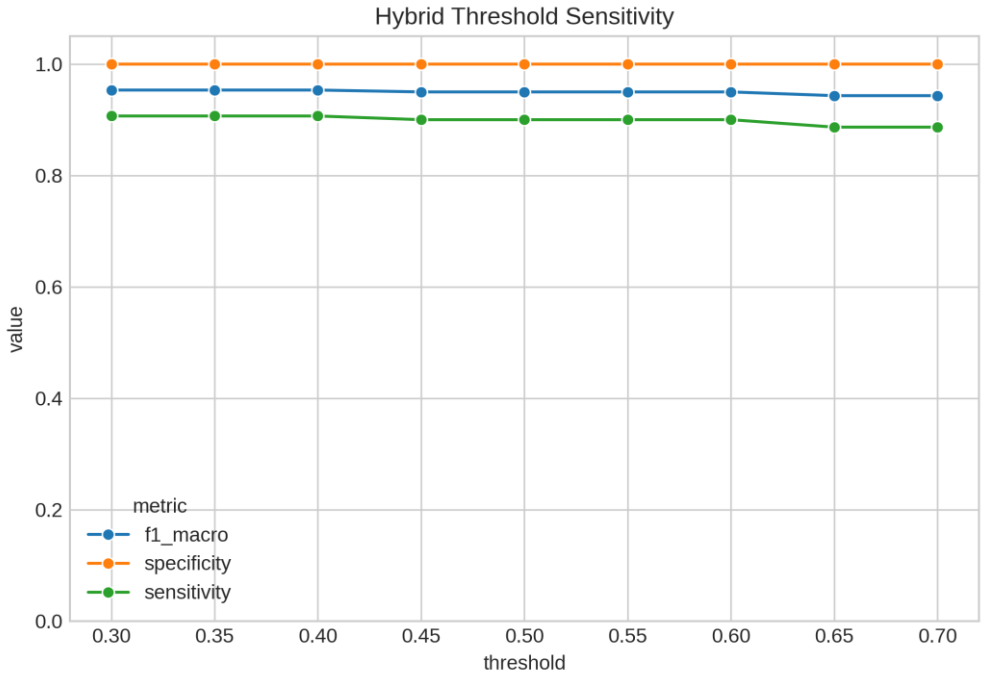


Figure 11: RoBERTa precision, recall, and macro F1 versus classification threshold on the leakage-resistant test set; θ=0.50 is the a priori default used throughout this study.

5.6 Inference Efficiency

Timing measurements were obtained on the leakage-resistant split in the reported Colab GPU environment (Table 6) and should be read as implementation benchmarks rather than deployment guarantees, since they depend on hardware, batch size, and software versions; CPU-only inference would be markedly slower for both transformers. LMDT timing applies only to messages that reach Stage 2 (Stage 1 phishing decisions).

6. Discussion

6.1 Classification

On the leakage-resistant split, BERT's macro F1 of 97.67% and RoBERTa's 94.99% are broadly comparable to email-focused benchmarks in the literature (Table 7), though direct comparison is limited by differing corpora and channel composition. Hash-based exact-text matching and the historical template key found no duplicate groups in this corpus draw; a similarity-based audit of the historical (template-grouped) split nonetheless found one test message at ≥99.9% cosine similarity to a training message, and 27 more test messages (9.0% of the test set) with a near-duplicate (≥0.85 cosine similarity) in the training or validation split. The revised near-duplicate grouping reduces both figures to zero. The measured change in RoBERTa macro F1 between the two protocols was small (+0.99 points) in this run; because the two protocols evaluate different test sets under a single training seed, this should be read as the combined effect of the protocol change and test-set variation, not as an isolated leakage estimate.

Table 7: Contextual accuracy tier for this work's leakage-resistant results against one external email-only benchmark; corpus size and channel composition differ, making direct comparison inadvisable.

Study	**Setting**	**F1 / Acc (%)**
Meléndez et al. [5]	Large email, RoBERTa	99.43
This work	Email+SMS bal., BERT	97.67

This work	Email+SMS bal., RoBERTa (hybrid backbone)	94.99

RoBERTa shows no macro F1 gap between the email and SMS subsets on the leakage-resistant split (94.99% for both, Table 4); a channel gap was not specifically re-evaluated under the historical protocol in this study, so this should not be read as evidence that near-duplicate grouping removed a channel effect reported elsewhere [13, 14]. Short-text phishing challenges have been noted across adjacent spam detection tasks as well [40, 41].

6.2 LMDT and LMS

Tactic detection metrics in Table 5 are measured against the adjudicated labels of two independent human annotators (Appendix B), rather than against the Qwen3-4B pseudo-labels used only in Part A dataset construction. Overall recall remains modest for all four tactics (mean per-tactic positive-class F1 28.5% on the 120-message annotated subsample), and the impersonation detector recovers zero true positives against human judgment. First-pass Cohen's kappa ranged from fair for urgency (kappa=0.35) to slight for authority, impersonation, and persuasion (kappa<0.03), and the 72 of 120 messages with at least one disagreement were resolved by adjudication before this evaluation (Appendix B). LMDT provides post-classification linguistic annotation; it is not a faithful explanation of RoBERTa's internal decision process and does not constitute an XAI method in the post-hoc approximation sense.

The impersonation detector recovers no true positives against the human-annotated reference (0 of 7 positive instances in the 120-message annotated subsample), the weakest result among the four tactics.

Urgency detection balances precision (37.5%) and recall (69.2%) against the human-annotated reference, recovering most of the messages the annotators marked urgent at the cost of a substantial false-positive rate.

To illustrate the LMDT's output, consider a representative phishing message pattern: "URGENT: Your account has been suspended by [Bank]. Verify your identity at the secure link below within 24 hours or access will be permanently revoked. Authorised by: Customer Compliance Team." The LMDT annotates: urgency (URGENT, within 24 hours, permanently revoked), authority (Compliance Team), impersonation ([Bank]), and persuasion (loss-framing). LMS = 0.30(1)+0.25(1)+0.25(1)+0.20(1) = 1.00. A legitimate notification with no such markers scores LMS = 0.00.

The LMS correlation with the human-derived reference aggregation (Spearman 0.337, 95% CI [0.174, 0.503]) is a real, moderate positive association rather than strong agreement; the interval excludes zero but the point estimate still leaves most of the variance in the human-derived aggregation unexplained by the LMS. LMS remains a deterministic heuristic ordering of detected tactic combinations, capped at 16 attainable values; its relationship with expert-rated manipulation severity or operational phishing risk is unvalidated, and no such external outcome was available in this study (Appendix E).

6.3 Deployment

The hybrid system's ≈16.6 ms latency per phishing message (RoBERTa plus the deployed LMDT v1, GPU environment) is compatible with real-time email gateway insertion; CPU-only deployment would be slower. The structured JSON output—label, confidence, tactic flags, evidence spans, LMS—can drive header enrichment, client-side warning banners, or SIEM rule generation. Email in-telligence pipelines that integrate NLP-based scoring with gateway metadata have been proposed as a complementary layer to filter-rule engines [42].

6.4 Limitations

The 2,000-message corpus is modest; performance may differ on larger or geographically diverse data. Tactic ground truth derives from a 120-message stratified subsample of the frozen test set, independently labelled by two annotators and resolved by adjudication (Appendix B), rather than the full 300-message test set; first-pass inter-annotator agreement was fair for urgency and slight for the other three tactics, reflecting ambiguity in the tactic definitions themselves, particularly for persuasion. Disagreements were adjudicated by the same two annotators rather

than a third, independent adjudicator, which, given the low first-pass agreement (kappa <=0.03 on three of the four tactics), limits the reliability of the resulting reference labels, particularly for authority (n+=3) and impersonation (n+=7), where positive instances are too few to resolve the v1-vs-v2 comparison statistically; a third-party tiebreak on the disputed labels is noted as a priority for future work. The corpus covers English only. Transformer classifiers were fine-tuned under a single random seed each, so their reported confidence intervals reflect test-set sampling variance only, not initialisation variance; the TF-IDF baselines were refit under three seeds, and only the Random Forest showed a non-zero (though narrow) spread (F1 std 0.005) — the linear SVM's seed affects only its internal probability calibration, so its spread was exactly zero and is not evidence of stability. No independently rated severity, priority, or behavioural outcome was collected for the LMS, so its weights remain unvalidated design choices rather than evidenced risk parameters. No user study was conducted; whether tactic explanations reduce click rates in practice remains empirically open. BERT's classification advantage over RoBERTa on this split (Results) is not reflected in the deployed hybrid, which was specified around RoBERTa before the BERT comparison was run and has not been rebuilt or re-evaluated with BERT as the Stage 1 backbone; this rebuild is the first item of future work.

7. Conclusion

We presented a two-stage hybrid phishing detection framework pairing a fine-tuned transformer classifier with a deterministic, regex-based LMDT. On a 300-message, near-duplicate-grouped held-out test set, fine-tuned BERT achieved the highest macro F1 (97.67%) and AUC (99.66%) among four classifiers compared under identical training conditions, significantly ahead of TF-IDF+RF and of RoBERTa (Holm-adjusted exact McNemar, $p<0.05$); RoBERTa (macro F1 94.99%) remains the classifier gating the deployed hybrid because the architecture was specified before the BERT comparison was run; replacing the backbone with BERT is the first item of future work. The LMDT added character-level tactic evidence spans and a weighted LMS at ≈16.6 ms per phishing message in the reported GPU environment; the LMDT rule engine itself adds ≈14.6 ms of that total. Tactic metrics reflect agreement with the adjudicated labels of two independent human annotators on a 120-message stratified subsample of the frozen test set, rather than an LLM pseudo-label reference; against this ground truth, the original rule set retains more recall than a proposed refinement to the authority and impersonation detectors and is kept in the deployed system. Cross-channel results for RoBERTa: F1 94.99% for both email and SMS.

Future directions include multilingual extension, extending the two-annotator adjudicated protocol (Appendix B) to the full 300-message test set to tighten the tactic-level confidence intervals, multi-seed transformer training to bound initialisation variance, collection of an independently rated severity outcome to validate or refine the LMS weights, and a controlled usability study measuring whether tactic explanations reduce click rates in simulated phishing scenarios.

8. Data and Code Availability

All phishing email and SMS messages used in this study are drawn from publicly available corpora (Nazario Phishing Corpus, Kaggle Phishing Email Dataset, Enron corpus, and SMS Smishing Collection). The LMDT rule sets, the Qwen3-4B annotation prompt (Appendix A), the per-message predictions and confidence intervals underlying Tables 2-4, and the human-annotated LMDT reference labels described in Appendix B are available from the corresponding author upon reasonable request.

9. Acknowledgments

The authors acknowledge the open-source communities behind transformers (Hugging Face), spaCy, and scikit-learn for the libraries on which this work depends. The phishing email and SMS corpora were obtained from publicly available datasets (Nazario Phishing Corpus, Kaggle Phishing Email Dataset, Enron corpus, and SMS Smishing Collection). No proprietary data were used.

A Qwen3-4B Tactic Annotation Prompt

The following prompt was applied to each message under deterministic inference (temperature = 0; prompt identifier lmdt_qwen3_v1.1):

You are an expert phishing analyst. Given the following message, identify which of these four linguistic manipulation tactics are present. Return ONLY a JSON object with boolean values. Tactic definitions: urgency — time pressure, deadlines, or threatened negative consequences; authority — reference to official entities, compliance requirements, or institutional sender roles; impersonation

— brand-name spoofing or first-person brand ownership claims; persuasion — reward framing, scarcity cues, or social proof. Message: [message text].

Expected output format:

{"urgency": bool, "authority": bool, "impersonation": bool, "persuasion": bool}

In our implementation, repeated runs on the same input

produced identical outputs; hardware or library differences may affect reproducibility.

B Human Annotation Codebook and Validation Protocol

Two independent annotators applied this codebook, blinded to model, LMDT, and Qwen outputs and to one another, to a stratified subsample of 120 messages (60 phishing, 60 legitimate) drawn from the 300-message frozen test set (Table 5). First-pass Cohen kappa per tactic (95% bootstrap CI, 2,000 resamples): urgency 0.35 [0.14, 0.55] (fair agreement), authority 0.01 [-0.11, 0.21], impersonation 0.00 (interval degenerate: the second annotator marked no positives), persuasion 0.03 [-0.10, 0.15] (slight agreement for the latter three, reflecting ambiguity in the tactic definitions). The 72 of 120 messages with at least one tactic-level disagreement were resolved through adjudication by the two annotators (Adjudication and reporting, below), producing the reference labels used in Table 5 and Figure 8. Extending this protocol to the remaining 180 messages of the frozen test set remains future work (Limitations). The Qwen3-4B labels used elsewhere in this study remain automated pseudo-labels and are not relabelled as human annotations.

Sampling. A 120-message sample was drawn from the 300-message frozen test set, stratified by class and channel with 30 messages in each of the four class-by-channel cells (60 phishing, 60 legitimate; 60 email, 60 SMS), and shuffled into a common order. The sample was not enriched for candidate-positive examples, which is why positives are few for authority (3) and impersonation (7). Each annotator received an identical copy holding the message text, the channel, and four empty binary columns, without the class label, the LMDT output, or the Qwen3-4B labels, and worked without sight of the other annotator's labels.

Unit and decision rule. Each annotator marked each of the four tactics as present (1) or absent (0) for every message, using the definitions in the codebook table below. An "unclear" option and span marking were not used, so borderline judgments are folded into the binary labels. Agreement was computed on the independent first-pass labels: observed agreement, Cohen's kappa per tactic, and a 95% bootstrap interval for each kappa (2,000 resamples).

Tactic	Present when	Do not label present solely because
Urgency	A deadline, immediate-action demand, or threatened time-linked consequence pressures the recipient.	A neutral date, routine reminder, or ordinary scheduling language.
Authority	An institution, official role, or compliance power is invoked to compel, verify, restrict, or sanction.	A brand/institution is mentioned only as context.

Tactic	Present when	Do not label present solely because
Impersonation	The sender claims to be, act for, or present credentials/account ownership of a named person or organisation, or identity mismatch evidence is present.	A named brand appears without an identity claim or spoofing evidence.
Persuasion	Reward, scarcity, social proof, reciprocity, or comparable influence framing is used to shape action.	Plain product/service information without an influence cue.

Adjudication and reporting. The two annotators went through the 72 messages with at least one disagreement together and agreed a final value for each disputed tactic; values for tactics on which they already agreed were kept. Both first-pass files are retained. The adjudicated labels are the reference for Table 5, Table F1, Table F2, and Table 8. A third, independent adjudicator was not used, and the remaining 180 test messages were not annotated (Limitations).

C Near Duplicate Control and Leakage Resistant Split

The three-tier grouping and stratified group split described below were applied to all reported classification results; the exact-match-only and template-key protocols appear in the Discussion (Classification) and in Tables C1 and C2 for comparison only, not as primary evidence.

Table C1: Duplicate structure of the 2,000-message corpus under the three grouping tiers (near-duplicate tier: strict template key plus character n-gram cosine similarity >= 0.85, transitive closure).

Grouping tier	Groups	Singleton groups	Multi-message groups	Messages in multi-message groups	Largest group
Exact match only	2,000	2,000	0	0	1
Historical template key	2,000	2,000	0	0	1
Strict template + near-duplicate	1,864	1,820	44	180	27

Table C2: Cross-split leakage audit of the 300 test messages against the 1,700 training and validation messages under each split scheme.

Split scheme	Test messages with exact twin in train	Test messages with near-duplicate in train	Cross-split pairs above threshold	Max cross-split similarity
Exact-match removal only	1	27 (9.0%)	193	0.999
Historical template grouped	1	27 (9.0%)	193	0.999
Revised near-duplicate grouped	0	0 (0.0%)	0	0.833

Messages were grouped on three tiers before splitting. Tier 1 joins byte-identical texts. Tier 2 reproduces the template key of the earlier pipeline: lower-cased text with links, e-mail addresses, and phone placeholders replaced by typed tokens and punctuation removed, numbers kept. Tier 3 adds a stricter template key in which digits are also replaced by a placeholder, and joins every pair of messages whose character 4-5-gram TF-IDF vectors (padded at word boundaries, minimum document frequency 2, sublinear term frequency) have cosine similarity of at least 0.85. Groups are the connected components of the resulting graph, so a template family stays together. The similarity vectoriser was fitted on all 2,000 messages without labels; it constrains the split and feeds no classifier.

The leakage-resistant split (1,400, 300, and 300 messages for training, validation, and test) assigns whole tier-3 groups to a split. Groups are placed largest first into the split with the largest unfilled share of its quota for the

group's dominant class-by-channel stratum, and a repair step ensures that validation and test each contain both classes. All TF-IDF baselines, BERT, and RoBERTa were retrained on this split, and the performance tables report only these results. The earlier template-grouped split is kept for comparison (Table C2).

D Statistical Evaluation Procedure

Per-message predictions, confidence intervals, and paired significance tests are reported in the Results. For every model and test message, a row stores the split scheme, the stage, the system, the message ID, the duplicate-group ID, the true label, the predicted probability, the predicted label (threshold 0.50), and whether the prediction was correct. Tactic predictions, evidence spans, reference labels, and LMS are stored separately for the LMDT evaluation.

Point estimates are reported with 95% percentile bootstrap intervals: 2,000 resamples of the test messages with replacement, without stratification, drawn from a separate seeded random stream for each system (seed 42 plus 1,000 times the system index). Wilson intervals are given alongside for the rate metrics (accuracy, sensitivity, specificity, precision). Table D1 lists the bootstrap intervals for accuracy, macro-F1, AUC, MCC, sensitivity, and specificity. For small subgroups the intervals are wide and are read as indicators of instability.

Table D1: Stage 1 test-set metrics with 95% percentile bootstrap intervals (2,000 resamples; n=300).

Model	Accuracy (%)	Macro F1 (%)	AUC (%)	MCC	Sensitivity (%)	Specificity (%)
TF-IDF+RF	94.0 [91.3, 96.3]	94.0 [91.3, 96.3]	98.3 [97.1, 99.3]	0.881 [0.827, 0.928]	92.0 [87.3, 96.0]	96.0 [92.6, 98.7]
TF-IDF+SVM	95.3 [93.0, 97.7]	95.3 [93.0, 97.7]	99.4 [98.8, 99.8]	0.907 [0.859, 0.953]	94.7 [90.7, 98.0]	96.0 [92.6, 98.7]
BERT	97.7 [96.0, 99.3]	97.7 [96.0, 99.3]	99.7 [99.2, 100.0]	0.954 [0.923, 0.987]	95.3 [91.8, 98.6]	100.0 [100.0, 100.0]
RoBERTa	95.0 [92.3, 97.3]	95.0 [92.3, 97.3]	99.3 [98.4, 99.8]	0.905 [0.857, 0.948]	90.0 [85.0, 94.6]	100.0 [100.0, 100.0]

Matched per-message prediction arrays for all four Stage 1 classifiers are retained for the same 300 held-out test messages; the exact McNemar results reported in the Results section are computed directly from these arrays and are reproducible from the archived predictions alone.

For each pair of Stage 1 classifiers, the 2 x 2 discordance table is built from the same 300 test messages: b counts the messages that only the first-named model classifies correctly and c counts those that only the second-named model classifies correctly; Results gives the counts with BERT named first. Two-sided exact binomial McNemar tests are reported with a Holm adjustment across all six pairwise comparisons. The prediction file and the notebook regenerate every interval and p-value. For the LMDT, the two rule versions are paired on the same adjudicated messages (Appendix F).

E Linguistic Manipulation Score Sensitivity and Ablation

Sensitivity and ablation analyses were run on the 120-message human-labelled subsample of the frozen test set (below). External validation against an independently rated severity outcome was not possible because no such outcome was collected, so LMS claims remain restricted to descriptive tactic aggregation (Results, Discussion). The weights 0.30, 0.25, 0.25, and 0.20 are heuristic defaults, not evidence-based estimates.

Table 8: LMS sensitivity and ablation analyses, computed against the human-derived reference aggregation on the 120-message adjudicated subsample.

Analysis	Procedure	Report
Equal-weight baseline	Compare the current weights with 0.25 for each tactic.	Spearman agreement with the human-derived reference is 0.337 at equal weights and 0.337 at the proposal weights; Kendall tau between the two rankings is 0.966; distinct score values fall from 9 to 5.

Analysis	Procedure	Report
One-factor sweep	Halve or double one weight at a time and rescale the others to sum to 1.	Spearman agreement ranges from 0.328 to 0.340 across the eight settings, within 0.01 of the 0.337 obtained at the proposal weights; the largest change is urgency halved (-0.008).
Component ablation	Set each tactic contribution to zero in turn and compare the resulting LMS to the full score.	Removing persuasion changes the Spearman agreement with the human-derived reference the most, from 0.337 to 0.199 (-0.137); removing urgency has a smaller effect (-0.061); removing authority or impersonation changes it by less than 0.02 in either direction. All five resulting 95% bootstrap intervals overlap the full-score interval, so no single component is individually decisive, but persuasion and urgency contribute more to the observed agreement than authority or impersonation.
Simplex sampling	2,000 Dirichlet(1,1,1,1) draws over the full non-negative weight simplex.	Kendall tau ranking stability against the proposal weights (5th percentile 0.829) and Spearman agreement with the human-derived reference (range [0.320, 0.345], mean 0.334, versus 0.337 at the proposal weights) across all 2,000 draws; the score is not highly sensitive to the exact weighting within the simplex.

No independently rated severity, prioritisation, or behavioural outcome was collected in this study; LMS claims are accordingly restricted to descriptive tactic aggregation throughout. Any future data-driven weighting must be selected on training/validation data and evaluated once on the held-out test set; it must not be tuned against the same test outcomes used for final reporting.

F Refined Authority and Impersonation Rules (v2) and Their Evaluation

These refinements (v2) were implemented, frozen, and evaluated against the same two-annotator adjudicated 120-message subsample as the original rules (v1; Results). Against human judgment, v2 records zero true positives for both authority and impersonation on this subsample (0 of 3 and 0 of 7 positive instances respectively), a result not captured in Table 5, which reports only the deployed v1 rules. The exact McNemar test on the accuracy-level discordant pairs is not significant for either tactic at this sample size (impersonation raw p=0.063, authority raw p=0.125; Holm-adjusted p=0.125 for both across the two changed tactics), though v2 does raise overall accuracy for both (impersonation 94.2% vs 90.0%; authority 97.5% vs 93.3%) because it removes the false positives v1 raised on legitimate messages. The positive-class F1 loss for authority is the larger effect in absolute terms (20.0% to 0.0%), though with only 3 positive instances in the sample this is better read descriptively than as a resolved statistical difference. Read together, v2 is not a clear improvement on either tactic and is not adopted; v1 is retained because it keeps materially more recall on both, which is what determines its selection on the validation split (Methodology).

Table F1: LMDT v1 versus v2 per tactic on the 120-message adjudicated subsample. Positive-class F1 with 95% bootstrap CI (2,000 resamples); b = v1 correct and v2 incorrect, c = v1 incorrect and v2 correct; p = exact two-sided McNemar test (raw).

Tactic	n+	v1 F1 [95% CI]	v2 F1 [95% CI]	Accuracy v1 / v2	b	c	p
Urgency	13	48.6 [26.1, 66.7]	48.6 [27.3, 66.7]	84.2 / 84.2	0	0	1.000
Authority	3	20.0 [0.0, 54.5]	0.0 [0.0, 0.0]	93.3 / 97.5	1	6	0.125
Impersonation	7	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	90.0 / 94.2	0	5	0.063
Persuasion	22	45.5 [30.3, 60.3]	45.5 [29.0, 58.8]	70.0 / 70.0	0	0	1.000

Table F2: Confusion counts, precision, and recall of LMDT v1 and v2 against the adjudicated labels (n=120). Precision is n/a where a version flagged no message.

Tactic	Version	TP	FP	FN	TN	Precision	Recall
Urgency	v1	9	15	4	92	0.375	0.692
Urgency	v2	9	15	4	92	0.375	0.692
Authority	v1	1	6	2	111	0.143	0.333
Authority	v2	0	0	3	117	n/a	0.000
Impersonation	v1	0	5	7	108	0.000	0.000
Impersonation	v2	0	0	7	113	n/a	0.000
Persuasion	v1	15	29	7	69	0.341	0.682
Persuasion	v2	15	29	7	69	0.341	0.682

Table F3: Positive-class F1 of LMDT v1 and v2 on the 300-message validation split, scored against the Qwen3-4B pseudo-labels (the only labels available for that split).

Tactic	v1 F1	v2 F1
Urgency	0.269	0.269
Authority	0.194	0.000
Impersonation	0.000	0.000
Persuasion	0.213	0.213

Detector	Refinement in v2	Effect on false positives
Authority	An institution or role term must share a sentence with a mandate, compliance, or directive cue and sit within 120 characters of it, or appear inside an institutional self-identification; role mentions introduced by a personal possessive are excluded.	A neutral mention of a role or institution no longer fires the detector; v1 fired on any such mention.
Impersonation	Sufficient on their own: a brand paired with a service or department role, a first-person identity claim, or a forged sender address. Conditional on an account-action request also being present: a bare brand token from the unambiguous-brand list, a generic salutation, or an automated-mail phrase. Brand tokens in order or delivery context are excluded.	A bare brand token or a generic salutation alone no longer fires the detector; v1 treated an optional role suffix as making a bare brand sufficient, and its salutation pattern fired on legitimate bulk mail.
Urgency, persuasion	Unchanged between v1 and v2; held constant as controls.	Outputs of the two versions are identical (Table F1).

The comparison of v1 and v2 was run per tactic on the same 120 adjudicated messages. Table F1 gives positive-class F1 with bootstrap intervals, accuracy, and McNemar counts. Table F2 gives the confusion counts,

precision, and recall of both versions. Table F3 gives the validation-split scores on which v1 was retained. Extending the comparison to the full frozen test set requires labels for the remaining 180 messages.

G Scope of Novelty and Related Work Comparison

The framework's claimed novelty is deliberately narrow. It does not claim that transformer-based binary classification is novel, nor that the weighted LMS is a validated risk score. Its contribution is the integration of a conventional transformer detector with deterministic, regex-based manipulation-tactic flags, character-span evidence, and a configurable descriptive score for analyst and awareness workflows, evaluated end-to-end against a 120-message, two-annotator adjudicated subsample of the test set with leakage-resistant splitting, retained per-message predictions, and paired significance testing.

Table 9. Related-work comparison and the scope of this framework's contribution.

Study already cited	Main emphasis	Relation to this framework
Ibrahim et al. [4]; Meléndez et al. [5]	Transformer/RoBERTa phishing classification.	Establishes a binary-detection baseline; LMDT adds a separate tactic-evidence layer rather than claiming a new classifier architecture.
Acharya et al. [2]	Lightweight transformer with token-level XAI highlights.	Related explainability approach; LMDT aims to name four pre-defined tactics and return deterministic rule-linked spans.
This work	RoBERTa gate plus regex-based LMDT and configurable, descriptive LMS across email and SMS.	Integration: binary detection plus auditable tactic flags/spans and a descriptive composite, evaluated against a leakage-resistant split and an independently human-annotated tactic ground truth.

REFERENCES

1. P. M. Kavitha, B. C. Akkina, and R. S. Perikala, “A review on phishing email detection using domain analysis and transformer-based explainability”, 2026 2nd International Conference on Big Data & Machine Learning (ICBDML), 2026.
2. M. Acharya and S. Sharma, “Detecting phishing emails through stylometric analysis: A novel NLP approach”, 2025 IEEE International Conference on Advances in Computing Research On Science Engineering and Technology (ACROSET), 2025.
3. S. Datto, D. H. Tuhin, M. Ahmed, K. Redwan, M. F. A. A. Sohan, A. Shufian, B. Tamang, R. Lama, and R. Shrestha, “Enhanced phishing payload detection using fine-tuned DistilBERT and XAI-based NLP models”, TEN-CON 2025 - 2025 IEEE Region 10 Conference (TEN-CON), 2025.
4. M. Ibrahim and R. Elhafiz, “Phishing email detection using BERT and RoBERTa”, Comput., 2026.
5. R. Meléndez, M. Ptaszynski, and F. Masui, “Comparative investigation of traditional machine-learning models and transformer models for phishing email detection”, Electronics, 2024.
6. S. Bagui, D. Nandi, S. Bagui, and R. J. White, “Machine learning and deep learning for phishing email classification using one-hot encoding”, Journal of Computer Science, 2021.
7. M. Regina, M. Meyer, and S. Goutal, “Text data augmentation: Towards better detection of spear-phishing emails”, ArXiv, 2020.
8. Y. Ajitha, P. Sudha, G. Gowthami, C. Shanthini, and M. A. Ahmad, “A dual pipeline AI framework for real-time detection of phishing and financial fraud in fraud gpt”, FMDB Transactions on Sustainable Technoprise Letters, 2024.
9. M. Aljamal, R. Alqura’n, M. Aljaidi, O. S. AlJamal, A. Alsarhan, I. Al-Aiash, G. Samara, M. Banisalman, and M. Khouj, “Harnessing ml and NLP for enhanced cybersecurity: A comprehensive approach for phishing email detection”, 2024 25th International Arab Conference on Information Technology (ACIT), 2024.
10. S. Sindhu, S. Patil, A. Sreevalsan, F. Rahman, and M. N, “Phishing detection using random forest, SVM and neural network with backpropagation”, 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020.
11. A. M. H. Al-Ibrahim, “Using sequential minimal optimization for phishing attack detection”, Modern Applied Science, 2019.
12. A. Maslan, A. Rahman, U. Faruq, R. Raad, and A. Al-Jawry, “Spam email classification optimization with NLP-based naïve bayes on TF-IDF and smote”, Jurnal Nasional Teknik Elektro dan Teknologi Informasi, 2025.
13. F. Khalid, M. Hussain, I. Noreen, and Z. Karim, “A comparative study of transformer-based models for multi-class SMS phishing detection”, 2026 International Conference on Data Science, Machine Learning, and Intelligence (DataSciMI), 2026.

14. S. Thangavel, T. Palanisamy, J. T, K. H. Karuvelil, D. Poojitha, and N. Prabhu, “Web phishing and email phishing analysis using tabular and semi-tabular approaches with NLP techniques”, 2025 9th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2025.
15. D. O. Otieno, A. Namin, and K. S. Jones, “The application of the BERT transformer model for phishing email classification”, 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), 2023.
16. S. G. Fukarel, O. P. Edriatama, M. Aqila, N. A. P. Masaling, and A. Chowanda, “Phishing detection with BERT: A transformer-based approach for cybersecurity”, 2025 IEEE International Symposium on Consumer Technology (ISCT), 2025.
17. M. Hasan, N. Nahar, M. H. Imam, and M. U. Khandaker, “PhisNet: Intelligent detection of phishing”, The Eurasia Proceedings of Science, Technology, Engineering and Mathematics, 2025.
18. G. Lingaiah, “An intelligent deep learning framework for detecting phishing attacks through email analysis on the CEAS_08 dataset”, International Journal of Applied Mathematics, 2025.
19. C. M. Chang, M. Al-Andoli, and C. Zheng, “Hybrid phishing detection model: Integrating BERT with TF-IDF for enhanced email security”, International Journal on Robotics, Automation and Sciences, 2025.
20. S. Noorain, J. Metan, T. Sharieff, and Z. Harmain, “Transformer-driven NLP framework for high-precision phishing email identification”, 2026 6th International Conference on Intelligent Technologies (CONIT), 2026.
21. M. Hosseinzadeh, U. Ali, S. Ali, R. Abbaszadi, F. S. Gharehchopogh, P. Khoshvaght, T. Porntaveetus, and J. Lánský, “Improving phishing email detection performance through deep learning with adaptive optimization”, Scientific Reports, 2025.
22. S. Abiramasundari and V. Ramaswamy, “Cacography based ransomware email phishing attack prevention using language pack tuned transformer language model”, Scientific Reports, 2025.
23. A. Gaurav, B. B. Gupta, J. Wu, V. Arya, and K. Chui, “Lightweight cloud-based phishing email detection using BERT and deep learning”, 2024 IEEE Globecom Workshops (GC Wkshps), 2024.
24. K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, “Next-generation spam filtering: Comparative fine-tuning of LLMs, NLPs, and CNN models for email spam classification”, Electronics, 2024.
25. K. Wisaeng and B. Muangmeesri, “Deep learning for email threat intelligence: Feature importance and model performance in phishing detection”, International Journal of Innovative Research and Scientific Studies, 2025.
26. S. Tarapiah, L. Abbas, O. Mardawi, S. Atalla, Y. Himeur, and W. Mansoor, “Evaluating the effectiveness of large language models (LLMs) versus machine learning (ML) in identifying and detecting phishing email attempts”, Algorithms, 2025.
27. M. N. Suhaimee, F. Shakil, M. R. A. A. Khan, M. O. Faruq, S. R. Sultana, and S. Firdaus, “Real-time incremental transformer with continual learning for adaptive phishing email detection”, 2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS), 2025.
28. R. E. Ulfath, H. AlQahtani, M. Hammoudeh, and I. H. Sarker, “Hybrid CNN-GRU framework with integrated pre-trained language transformer for SMS phishing detection”, Proceedings of the 5th International Conference on Future Networks and Distributed Systems, 2021.
29. H. Milner and M. Baron, “Establishing an optimal online phishing detection method: Evaluating topological NLP transformers on text message data”, Journal of Data Science and Intelligent Systems, 2023.
30. O. S. Kumar, K. P. Kumar, K. Chaitanya, M. Saiteja, K. Abhilash, and B. Goud, “Enhancing protection with NLP-based phishing detection across diverse communication channels”, 2025 3rd World Conference on Communication & Computing (WCONF), 2025.
31. A. Alhuzali, A. Alloqmani, M. Aljabri, and F. Alharbi, “In-depth analysis of phishing email detection: Evaluating the performance of machine learning and deep learning models across multiple datasets”, Applied Sciences, 2025.
32. F. Agusman, “Comparative study of DistilBERT and ELECTRA-small models in spam email classification”, Jurnal Informatika, 2025.
33. H. Asliyuksek, O. Tonkal, and R. Kocaoğlu, “A comparative evaluation of a multimodal approach for spam email classification using DistilBERT and structural features”, Electronics, 2025.
34. J. W. John and M. B, “A comparative analysis of BERT and LLM models for advanced phishing detection”, 2025 IEEE International Conference on Communication Networks and Computing (CNC), 2025.
35. S. Irfan, “Enhancing email security through accurate phishing detection using deep transformer models”, 2026 World Conference on Computational Science and Technology (WcCST), 2026.
36. B. Majumder, A. R. Chowdhury, A. Hasan, D. I. Didar, B. Islam, and S. Maruf, “A multi-view feature fusion approach with XLM-RoBERTa for spam email detection”, 2025 28th International Conference on Computer and Information Technology (ICCIT), 2025.
37. B. Varshitha, B. Ganesh, D. Saiteja, D. S. Rani, and M. Jabbar, “AI-driven phishing website detection using URL and HTML features with NLP”, 2026 2nd International Conference on Computing, Communication and Green Engineering (CCGE), 2026.
38. S. Mohan, S. Sharma, M. M. Unnithan, and S. Basavaraju, “Robust phishing detection via transformer embeddings and adversarial testing with generative AI”, 2025 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM), 2025.

39. M. Verma, “Automated fraud and phishing detection through natural language processing: A deep contextual approach to combating generative threats”, International Journal for Research in Applied Science and Engineering Technology, 2026.
40. L. Naik, B. Ks, S. Ss, and S. K., “AI-driven email classification and phishing attack detection for secure communication”, 2025 9th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), 2025.
41. G. S, F. S. S, C. S, and D. S, “AI-driven phishing detection using email headers and content analysis”, 2025 9th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2025.
42. G. Sharma and R. Kumawat, “Email intelligence with NLP: Summarization, spam detection, and hybrid filtering models”, 2025 Second International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT), 2025.