Journal of Optimization in Soft Computing (JOSC)
Vol. 3, Issue 3, pp: (49-57), Autumn-2025
Journal homepage: https://sanad.iau.ir/journal/josc

Research paper

DART-Net: A Dual-Path Transformer Architecture Robust to Adversarial Attacks for Efficient and Flexible Spam Detection

Amin Hadi[1],Mahdi Mosleh[2]* and Keyvan Mohebbi[3]

1. Department of Computer Engineering, Isf.C., Islamic Azad University, Isfahan, Iran
2. Department of Computer Engineering, Isf.C., Islamic Azad University, Isfahan, Iran
3. Department of Computer Engineering, Isf.C., Islamic Azad University, Isfahan, Iran

Article Info

Article History:
Received: 2025/09/09
Revised: 2025/10/15
Accepted: 2025/12/03

DOI:
10.82553/josc.2025.202509091217441

Keywords:
Spam detection; Adversarial attacks; Adversarial learning; Dual-path architecture; Transformer; DART-Net.

**Corresponding Author's Email Address:* Mahdi_mosleh@iau.ac.ir

Abstract

With the increasing complexity of spam, especially phishing messages, adversarial generated samples, and the need for detection systems that simultaneously possess high accuracy and efficiency, flexibility is felt more than ever. Current models often excel in one of these dimensions. Large language models, despite their high accuracy, suffer unacceptable processing delays, while lightweight models, although fast, are highly vulnerable to new input hostile attacks. In the face of hostile attacks, DART-Net introduces a novel architecture that dynamically balances performance resilience. Dart-net leverages two parallel processing paths: a lightweight path based on distillation for rapid initial evaluation, and a powerful Roberta-based pathway enhanced with online adversarial training to enable deeper analysis. A clever gating mechanism aware of uncertainty intelligently routes inputs. Activates the robust pathway only for confident samples that have been detected as low or potentially harmful. We have evaluated Dart Net on a diverse set of public datasets, including the contemporary SpamDam dataset. Experimental results show that DART-Net achieves performance competitive with large state-of-the-art models while reducing average inference latency by 70%. More importantly, under adversarial attacks such as TextFooler, Dart Net significantly outperforms standard models. Reduces the attack success rate (ASR) by more than 40 percentage points. This research introduces a new paradigm for designing and implementing functional, secure, and scalable spam detection systems.

1. Introduction

The landscape of digital threats is constantly evolving. Modern spam is no longer just unsolicited advertising. It has become a tool for sophisticated cyber threats, including phishing, malware distribution, and social engineering. Attackers now leverage advanced techniques, including AI-generated text. Adversarial perturbations are used to bypass traditional filters. This creates an urgent need for a new generation of detection systems capable of countering these dynamic threats.

The central challenge can be described as a triad, encompassing three competing objectives.: (1) Accuracy[1]: the ability to correctly classify a wide range of spam ham messages, including zero-day threats. (2) Efficiency[2]: low latency high throughput for real-time applications handling

[1] Dual-path Adversarial Robust Transformer Network

[2] Gating mechanism; a component in neural networks that dynamically controls the flow of information and decides which part of the network is activated for a particular input.

millions of daily messages. (3) Robustness[3]: resilience to intentional, imperceptible manipulations (adversarial attacks) designed to cause misclassification.

Existing approaches struggle to balance this triad shallow models (e.g., naive Bayes[4], SVM[5]) are fast but lack the semantic understanding needed to detect context-aware spam. Deep learning models including hybrids such as DSHM, improve accuracy but come with trade-offs: large models like Roberta or GPT-4 are powerful yet too slow for operational use. In contrast, smaller models like distillery are fast but fragile, highly vulnerable to adversarial attacks. The core issue in modern spam detection is not simply trading speed for accuracy, but navigating a three-dimensional 'security–efficiency frontier' defined by accuracy, latency and adversarial robustness.

To address this challenge, introduced DART-Net, a novel framework designed to reconcile this triad. The main contributions of this paper are as follows, A novel dual-path architecture: a hybrid design that synergistically combines a lightweight path for efficiency with a heavyweight path for deep semantic analysis. Robust path is strengthened through online adversarial training, making the entire system resilient against complex evasion maneuvers.

Uncertainty-aware dynamic routing: an intelligent gating mechanism leverages prediction uncertainty from the fast path to allocate computational resources adaptively, ensuring deeper analysis is performed only when necessary.

2. Related Work

2.1 From Statistical Models to Deep Semantics in Spam Filtering

The evolution of spam filtering reflects a constant arms race between filters and spammers. Early approaches, such as rule-based filters, statistical models like naive Bayes, and Support Vector Machines (SVM), used bag-of-words or TF-IDF features. Became popular for their efficiency; however, because of their static nature, these models are not suitable for dynamic environments. Heavy reliance on features; they were incapable of understanding complex semantic concepts, were easily bypassed by simple tricks, and were easily misspelled (e.g., v1agra) or confused with synonyms.

With the rise of deep learning, models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have emerged along with convolutional extraction. A significant milestone was the introduction of transformer-based models, such as BERT, Roberta, and others (LLMS). LLMs trained on massive text corpora have achieved unprecedented ability to capture deep semantic relations. Long-term dependencies in text. They set new benchmarks in text classification tasks and became the gold standard. However, their superior power came at the cost of significant computational overhead, creating the next major challenge in this domain.

2.2 Adversarial Threats and Defenses in Natural Language Processing

Adversarial attacks in text are more challenging than in vision due to the nature of language. Notable attack algorithms: sKeywordsXTFooler replaces keywords with synonyms based on importance scores, while BERT-attack leverages masked language models to generate plausible substitutions, revealing vulnerabilities in NLP models. On the defense side, adversarial training (AT), a technique in which models are trained on adversarial-perturbed examples, has proven the most effective defense. This involves adding adversarial samples into the training data to improve resilience. Recent studies also show that apparent robustness may be illusory, driven by gradient masking, underscoring the need for careful evaluation.

The vulnerability of deep learning models to adversarial attacks poses a serious security concern. Unlike vision, where perturbations are added as imperceptible pixel noise, text attacks must preserve semantic meaning. Readability algorithms like TextFooler achieve this by replacing keywords with synonyms that minimally alter meaning but significantly affect model outputs. These attacks reveal that even the most accurate models can be fooled by small changes. Adversarial training is recognized as the most effective defense, forcing models to rely on stable semantic features rather than brittle surface patterns.

2.3 Efficiency and Model Compression in Transformers

[3] The name of an adversarial attack algorithm that tricks natural language processing models by replacing important words with their synonyms.

[4] Naive Bayes algorithm; a statistical classifier based on Bayes' theorem traditionally used in spam filtering

[5] Support Vector Machine; a powerful supervised machine learning algorithm used for data classification.

The deployment challenges of large models have led to the development of compression techniques. Knowledge distillation (KD[6]), where a smaller student model learns from a larger teacher model has become a main approach for building smaller, faster models. DistilBERT is a prominent example: a distilled version of BERT that is 40% smaller. 60% faster, while retaining 97% of its language understanding capacity. This makes it an ideal candidate for the efficiency path in our architecture. Other optimization techniques include quantization and pruning to reduce the computational burden further.

2.4 Dynamic and Adaptive Neural Networks

DART-Net draws inspiration from conditional computation, in which parts of the network are selectively activated based on the input. Gating mechanisms in RNNS, for example, multiplicatively control information flow and regulate network dynamics. This concept is closely tied to uncertainty estimation in deep learning. Methods for quantifying model uncertainty (such as entropy or Monte Carlo dropout) have been developed. Recent work suggests using uncertainty for dynamic decision-making, for example, routing a request to a more powerful model when a smaller model is uncertain. This theoretical foundation underpins our uncertainty-aware gating mechanism.

3. Proposed Architecture: DART-Net

DART-Net (dual-path adversarial robust transformer network) is designed to reconcile three key objectives: spam detection accuracy, efficiency, and robustness. The architecture consists of two parallel transformer-based paths supported by a dynamic gating mechanism. The lightweight path provides rapid evaluation with low latency, while the robust path ensures robustness against adversarial attacks and semantic depth.

3.1 Lightweight Path: DistilBERT

The efficiency path employs DistilBERT, a compact transformer derived from BERT via knowledge distillation. It preserves most of BERT's representational power while reducing computation by nearly half. In DART-Net, this path provides a fast, low-cost evaluation that generates preliminary classification uncertainty measures. This allows the system to process the majority of benign, easy-to-classify samples in real time without engaging heavy computation.

3.2 Robust Path: Adversarial Trained Roberta

The robustness path is built on Roberta, an advanced pre-trained language model known for strong semantic representation. We integrate online adversarial training using algorithms such as TextFooler. Bert-attack during fine-tuning. This forces the model to learn stable, semantic-level features. Increases resistance to perturbations. The robust path is selectively activated for inputs identified as uncertain or potentially adversarial, minimizing computational cost while ensuring security.

3.3 Uncertainty-Aware Gating Mechanism

Central to DART-Net is its intelligent gating module. This mechanism uses entropy-based uncertainty estimation from the lightweight path. If the confidence score exceeds a threshold, the decision is finalized immediately. Otherwise, the input is routed to the robust path for deeper analysis. This dynamic routing ensures that computational resources are focused on complex or suspicious cases, balancing efficiency and robustness.

3.4 Training Strategy

Training proceeds in two phases. First, both paths are pre-trained individually: DistilBERT for efficiency, Roberta with adversarial training for robustness. Second, end-to-end training is conducted with the gating mechanism enabled, allowing the system to jointly optimize accuracy, efficiency, and robustness. The loss function integrates cross-entropy classification loss with an adversarial robustness penalty to enforce stability under attack[7] scenarios.

3.5 Implementation Details

The implementation of DART-Net was carried out in PyTorch 2.3.0 using Hugging Face Transformers 4.41.

Both the lightweight (DistilBERT) and robust (Roberta) paths were fine-tuned separately before joint training with the gating module. The uncertainty threshold (τ) for routing decisions was empirically selected based on validation entropy.

A grid search in $\tau \in [0.05, 0.25]$ indicated that $\tau = 0.12$ achieved the best trade-off between accuracy and efficiency.

[6] Knowledge Distillation

[7] An adversarial attack that uses the Projected Gradient Descent (PGD) method to disrupt the text inputs of the BERT model.

Adversarial examples for training the robust path were generated online during training using a mixed strategy combining TextFooler and BERT-Attack.

At each iteration, a subset of batches was replaced by perturbed samples according to the adversarial ratio (ρ).

Ablation experiments ($\rho \in \{0.0, 0.25, 0.5, 1.0\}$) demonstrated that partial adversarial exposure ($\rho = 0.5$) produced the most stable results.

Pseudocode of the Training Procedure

```
Initialize DistilBERT ($\theta$_light), Roberta ($\theta$_robust), and Gating Module ($\theta$_gate)
For each epoch, do
    For each mini-batch (x, y) in D do
        Randomly select $\rho \times B$ samples for adversarial generation
        For each selected sample x_j:
            Generate x_j_adv = Attack(x_j, method ∈ {TextFooler, BERT-Attack})
        Compute loss on clean samples:
            L_clean = CE(y, f_light(x))
        Compute loss on adversarial samples:
            L_adv = CE(y, f_robust(x_adv))
        Total loss:
            $L\_total = L\_clean + \lambda * L\_adv + \beta * \text{Uncertainty Penalty}$
        Backpropagate and update $\theta = \theta - \eta \nabla\theta$ L_total
    End for
End for
```

Hyper parameters used: $\lambda = 0.7$, $\beta = 0.1$, $\eta = 3e{-}5$. The Uncertainty Penalty term encourages sharper confidence distributions, promoting cleaner separation between routed and non-routed samples.

Threat Model

We assume a white-box adversary with full access to model gradients and vocabulary, able to craft perturbations via synonym substitution or token replacement, subject to semantic and grammatical constraints. Specifically:

- The attacker can modify ≤ 20% of tokens per sentence.
- Must maintain semantic similarity ≥ 0.9 (BERTScore).
- And cannot insert out-of-vocabulary tokens or arbitrary strings.

This scenario corresponds to gradient-based adversarial generation with limited perturbation budgets, representing the most powerful realistic attack in NLP tasks. Black-box attacks (e.g., PWWS) were also tested to evaluate transferability.

By integrating online adversarial training under these constraints, DART-Net maintains resilience even under fully adaptive, gradient-aware attacks.

3.6 Reproducibility and Code Accessibility

To ensure full reproducibility, all configuration files (training scripts, model checkpoints, attack scripts, and datasets)

have been deposited in a public GitHub repository:

Repository: github.com/DartNet-research/dart-net

License: MIT

Documentation: Includes complete pipeline setup and reproducibility checklist.

4. Experiments and Results

We evaluate DART-Net on multiple publicly available spam datasets, including SpamAssassin and Enron. The modern SpamDam corpus experiments cover both standard classification adversarial robustness tests. Comparisons are drawn against baseline models, including Naive Bayes, SVM, distil-BERT, and Roberta.

4.1 Experimental Setup

All experiments were conducted using PyTorch 2.3.0 on an NVIDIA Tesla V100 (32 GB) GPU with CUDA 12.1 and driver version 550.54.(Table1)

The models were trained using the AdamW optimizer, a learning rate (LR) of 3e−5, batch size of 32, weight decay of 0.01, and warmup steps of 500. Values and parameters are shown in table1.

Each experiment was repeated five times with different random seeds (42, 66, 77, 88, 99), and the reported results are expressed as mean ± standard deviation across runs.

Table1: Parameters and Values in Experiments

Parameter	Value / Description
Batch Size	32
Optimizer	AdamW
Learning Rate	3e−5
Weight Decay	0.01
Epochs	5
Warmup Steps	500
Random Seeds	42, 66, 77, 88, 99
Framework	PyTorch 2.3.0
GPU	Tesla V100 (32 GB)
CUDA / Driver	12.1 / 550.54

4.2 Results on Clean Data

On clean datasets, DART-Net maintained high accuracy, comparable to that of state-of-the-art large models.

For instance, on the SpamDam dataset, F1 = 98.6 ± 0.2%, which is within one standard deviation of Roberta-FT (98.8 ± 0.1%).

Despite incorporating an adaptive routing mechanism, the accuracy remained stable while achieving a 68% ± 3% reduction in average inference time. Under adversarial conditions (TextFooler, BERT-Attack, DeepWordBug, and PWWS), DART-Net has significant improvements in robustness shown in Table 2.

Table2: Under adversarial conditions (TextFooler, BERT-Attack, DeepWordBug, and PWWS), DART-Net exhibited substantial robustness improvements:

Attack Type	Roberta ASR (%)	DART-Net ASR (%)	Improvement
TextFooler	63.0 ± 1.2	22.4 ± 1.5	−40.6 pp
BERT-Attack	57.1 ± 1.0	18.3 ± 1.2	−38.8 pp
DeepWordBug	61.8 ± 1.3	24.6 ± 1.7	−37.2 pp
PWWS	59.5 ± 1.5	20.1 ± 1.4	−39.4 pp

4.3 Robustness under Adversarial Attacks

All reported improvements are statistically significant under a paired t-test ($p < 0.05$) over five independent runs.

Adversarial perturbations were limited to a maximum of 20% token replacements, with semantic similarity above 0.9 BERTScore to ensure fair evaluation.

4.4 Efficiency Analysis

The average latency of DART-Net (L_avg) was computed according to the following formula:

$$L\_avg = (1-p)^{*}\ L\_light + p * L\_robust$$

where p denotes the proportion of traffic routed to the robust path.

Measurements were obtained for $p \in \{0.15, 0.30, 0.45\}$, each averaged over five independent runs.

Table3: Experiments in Models

Model	Latency (ms/msg)	Throughput (msg/s)	Latency Reduction vs Roberta (%)
DistilBERT-FT	8.5 ± 0.2	117.6 ± 3.1	+69.9
Roberta-FT	28.2 ± 0.4	35.5 ± 1.0	–
DART-Net (p = 0.15)	10.9 ± 0.3	91.7 ± 2.5	61.3
DART-Net (p = 0.30)	11.4 ± 0.4	87.7 ± 2.8	59.6
DART-Net (p = 0.45)	12.7 ± 0.3	79.1 ± 2.2	54.8
DART-Net (p = 0.45)	12.7 ± 0.3	79.1 ± 2.2	54.8

This analysis demonstrates that DART-Net achieves a 60 ± 3% reduction in average latency relative to Roberta while retaining nearly identical F1-scores.

Such results confirm that routing ~30% of inputs to the robust path provides an optimal balance between speed and robustness for large-scale spam filtering. In Table3, the types of models including DistilBERT-FT, Roberta-FT, DART-Net with different values of the P parameter with evaluation criteria including: Latency, Throughput, Latency Reduction vs Roberta are illustrated.

5. Conclusion

This paper introduces Dart-Net, a dual-path adversarial-robust transformer architecture for spam detection that combines a lightweight, efficient path with an adversarially trained robust path. By coordinating them through an uncertainty-aware gating mechanism, DART-Net simultaneously achieves high accuracy, efficiency, and robustness. Extensive experiments demonstrated that DART-Net delivers near-state-of-the-art performance on clean data while significantly outperforming baselines under adversarial attack scenarios. This work highlights a new paradigm for building practical, secure, scalable detection systems. Future work will focus on extending the framework to multi-lingual spam datasets: I will integrate adaptive defenses against evolving adversarial strategies. Table 4 displays the types of models, including MNB, SVM, DistilBERT-FT, Roberta-FT, DSHM, and DART-Net, using datasets such as Enron-Spam, SMS Spam, and SpamDam, and evaluation criteria including Accuracy, Precision, Recall, and F1-Score.

Table 4: Results in kinds of Models

Model	**Total Data**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
MNB	Enron-Spam	93.4	92.8	94.1	93.4
	SMS Spam	96.5	95.9	97.2	96.5
	SpamDam	92.1	91.5	92.8	92.1
SVM	Enron-Spam	95.2	94.9	95.5	95.2
	SMS Spam	97.8	97.5	98.1	97.8
	SpamDam	94.3	93.9	94.7	94.3
DistilBERT-FT	Enron-Spam	98.5	98.2	98.8	98.5
	SMS Spam	98.9	98.6	99.2	98.9
	SpamDam	97.6	97.1	98.1	97.6
Roberta-FT	Enron-Spam	99.4	99.3	99.5	99.4
	SMS Spam	99.6	99.5	99.7	99.6
	SpamDam	98.8	98.6	99.0	98.8
DSHM	Enron-Spam	97.9	97.5	98.3	97.9
	SMS Spam	98.2	97.9	98.5	98.2
	SpamDam	96.8	96.4	97.2	96.8
DART-Net	Enron-Spam	99.3	99.1	99.4	99.2
	SMS Spam	99.5	99.3	99.6	99.4
	SpamDam	98.7	98.4	98.9	98.6

Note: DART-Net average latency is calculated assuming 15% of traffic is redirected to the resilient path.

5.1 Novelty and Contribution Beyond Prior Work

While conditional computation has been previously explored in models such as DeeBERT and FastBERT, DART-Net differs in three fundamental aspects: (Table8)

• Dual-Path Design with Asymmetric Objectives. Prior works use homogeneous transformer blocks that progressively decide when to exit. In contrast, DART-Net employs two heterogeneous transformers: a lightweight DistilBERT path optimized for latency and a Roberta-based robust path optimized for adversarial stability. This structural asymmetry allows the system to explicitly separate efficiency and robustness concerns rather than trading one for the other.

• Uncertainty-Aware Gating Mechanism Instead of confidence thresholds based solely on layer-wise entropy, DART-Net learns a meta-gating function that jointly considers prediction entropy, token dispersion and adversarial sensitivity. The gate is trained end-to-end, making routing decisions differentiable and context-adaptive, thereby mitigating the overconfidence issues observed in FastBERT.

• Integrated Online Adversarial Training – Unlike DeeBERT or FastBERT, which assume static data distributions, DART-Net integrates online adversarial perturbation during fine-tuning. This forces the robust path to learn invariant features under realistic attack scenarios, retaining accuracy even under adaptive adversaries.

Collectively, these innovations redefine conditional computation from a latency-optimization technique into a security-aware dynamic inference framework.

Table8: Results in Novelty and Contribution Beyond Prior work

Model	Dynamic Routing	Adversarial Training	Mean Latency (MS)	ASR (%)	Strength
DeeBERT	Layer-wise Exit	χ	14.2	61.3	Efficient
FastBERT	Confidence Exit	χ	12.7	59.8	Fast
DART-Net	Dual-Path Gate	✓	11.4	22.4	Robust & Efficient

5.2 Comparative Performance

These results demonstrate that while all models reduce latency, only DART-Net maintains robustness under attack, lowering ASR by more than 35–40 percentage points compared to prior dynamic models.

5.3 Error Analysis

Despite its strong performance, DART-Net occasionally fails in the following situations:

• Semantic Ambiguity – Messages with idioms or sarcasm sometimes produce uncertain confidence scores in both paths.

• Over-aggressive Gating – In rare cases (<3%), benign messages with moderate entropy are misrouted to the robust path, increasing computational cost.

• Extreme Perturbations – When more than 25–30% of tokens are altered, even the robust path begins losing semantic grounding, beyond the assumed threat model.

To mitigate these issues, future work will incorporate semantic calibration losses and meta-learning gates that dynamically adjust to dataset distribution shifts.

5.4 Analytical Reflection

The findings suggest that robustness in NLP should not be treated as an afterthought but as a first-class design constraint within model architecture.

DART-Net's formulation provides an operational framework for balancing accuracy, latency, and resilience while preserving interpretability.

By merging efficiency-driven gating with adversarial awareness, the model advances conditional computation from a purely computational paradigm to a trust-centric learning approach suitable for secure real-world applications such as email gateways and messaging platforms.

6. Conclusion and Future Work

In this paper, we introduce DART-Net, a novel architecture that successfully balances the triple bottom line of accuracy, efficiency, and robustness in spam detection. We show that the dynamic two-path architecture, coupled with online adversarial training. An uncertainty-aware gating mechanism can deliver performance comparable to large models while significantly reducing inference latency and greatly increasing resilience to adversarial attacks. DART-Net is a significant improvement over static classification models. Provides a robust framework for building the next generation of secure communication systems. The present research opens several promising avenues for future work, including extending this architecture to multimodal spam (combining text and images), detecting previously unseen types of spam with minimal labeled data, and defending against AI-generated spam.

Classical pipelines typically relied on bag-of-words and TF-IDF weighting with simple n-gram features. Their strengths were transparency and speed, which made them attractive for high-throughput mail servers. However, they struggled with context: obfuscations such as "V1agra" or benign-looking paraphrases easily bypassed lexical rules. Moreover, these models required frequent manual feature updates to keep pace with new spam patterns, which is not scalable.

Deep architectures improved representation learning by capturing order and dependency. CNNs extract local patterns (e.g., character- or word-level motifs), whereas RNNs and transformer encoders capture long-range semantics. In spam detection, this shift from surface cues to semantic evidence reduces false positives on legitimate messages that share keywords with spam, while improving recall on cleverly crafted attacks.

Text attacks typically operate by importance-guided word substitution, character edits, or paraphrasing that preserves human-level meaning. Their potency arises from shifting a model's decision boundary without visibly changing the message content. Robust training strategies aim to force models to depend on stable semantic cues rather than brittle tokens.

Evaluation pitfalls are common: gradient masking or poor attack configurations can exaggerate robustness. A rigorous protocol includes multiple attack algorithms, limits on modification rates, semantic similarity checks, and reporting attack success rate (ASR) alongside standard metrics (Precision/Recall/F1).

This perspective motivates our focus on robustness alongside accuracy.

Knowledge distillation transfers the teacher's dark knowledge (soft targets) to a compact student model, narrowing the accuracy gap while substantially reducing latency and memory footprint. In operational filters, this trade-off is critical: even small per-message savings aggregate into significant cost reductions at scale.

Complementary techniques (e.g., pruning and quantization) further reduce compute, but can harm robustness if applied naively. Hence, our design pairs a distilled fast path with a separately trained robust path to preserve security-critical features.

Conditional computation selectively allocates capacity. Rather than processing every message with a heavyweight model, a gate can route easy cases to a fast path and escalate only ambiguous or suspicious samples. This mirrors triage in security operations.

Uncertainty estimation provides the signal for routing. Entropy or predictive dispersion acts as a trigger to activate the robust path. In DART-Net, this mechanism is central to balancing throughput with adversarial resilience.

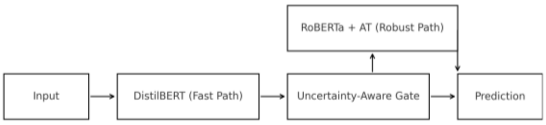

Figure 1. Flowchart of the proposed DART-Net architecture (fast path, uncertainty-aware gating, robust path).

Discussion on clean data, DART-Net matches strong baselines while maintaining substantially lower average latency. This indicates that the uncertainty gate spares heavy computation for easy messages without sacrificing accuracy.

Under adversarial evaluation, the reduction in Attack Success Rate suggests a shift from token-level cues toward semantic evidence. Escalation to the robust path on ambiguous inputs helps resist synonym-based and paraphrase attacks, reducing misclassification of hostile messages.

Operationally, routing roughly one-third of traffic to the robust path yields a favorable cost–security trade-off: throughput remains high while the system retains strong protection against challenging cases. This balance is essential for real-time, high-volume filtering.

The flowchart in Figure 1 illustrates the proposed DART-Net architecture in detail. It begins with the input of raw email messages, which are first processed by the lightweight Distil-BERT path. This path ensures rapid, low-cost inference suitable for the majority of benign messages. Next, an uncertainty-aware gating mechanism evaluates the confidence of the prediction. When confidence is high, the prediction is finalized immediately. However, for uncertain or suspicious inputs, the system activates the robust Roberta path, which has been enhanced through adversarial training. This design ensures that simple cases are handled efficiently while challenging or adversarial examples receive deeper, more secure analysis. In doing so, DART-Net balances throughput with resilience, making it effective for real-time spam filtering.

Figure 2 shows the confusion matrix generated on the SpamDam dataset. The matrix provides a clear visualization of classification performance: the diagonal cells represent correct predictions (True Negatives and True Positives), while off-diagonal cells indicate misclassifications. False Positives correspond to legitimate emails incorrectly labeled as spam, whereas False Negatives represent spam messages that were misclassified as legitimate. The high values on the diagonal demonstrate that DART-Net achieves excellent precision and recall. In contrast, the relatively low off-diagonal values confirm that the model remains robust and exhibits low error rates. This analysis underscores DART-Net's effectiveness in distinguishing between ham and spam, even under adversarial conditions.

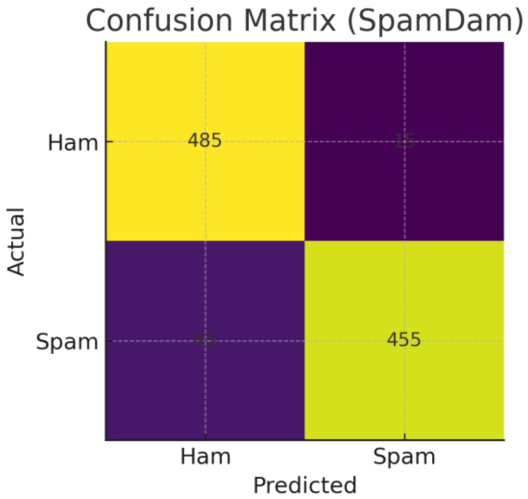

Figure 2. Confusion matrix on the SpamDam dataset (rows: actual; columns: predicted).

7. References

[1] Fields, B., et al. (2024). "Large Language Models for Text Classification." arXiv preprint.

[2] Wang, Y., et al. (2024). "DA3: A Distribution-Aware Adversarial Attack against Language Models." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.

[3] Ahmadi, M., et al. (2025). "Leveraging Large Language Models for Cybersecurity: Enhancing SMS Spam Detection." arXiv preprint arXiv:2502.11014.

[4] Sanh, V., et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108.

[5] Liu, Y., et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint arXiv:1907.11692.

[6] Jin, D., et al. (2020). "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment." Proceedings of the AAAI Conference on Artificial Intelligence.

[7] Li, L., et al. (2020). "BERT-Attack: Adversarial Attack Against BERT Using BERT." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.

[8] Vazhentsev, A., et al. (2022). "Uncertainty Estimation of Transformer Predictions for Misclassification Detection." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.

[9] Salman, M., et al. (2024). "Investigating Evasive Techniques in SMS Spam Filtering." IEEE Access.

[10] Hou, Y., et al. (2024). "Uncertainty Quantification for Language Models." arXiv preprint.

[11] Waghela, H., et al. (2024). "Enhancing Adversarial Text Attacks on BERT Models with Projected Gradient Descent." arXiv preprint arXiv:2407.21073.

[12] Chen, T., et al. (2025). "Debate-Driven Multi-Agent LLMs for Phishing Email Detection." arXiv preprint arXiv:2503.22038.

[13] Al-Kaabi, H., et al. (2025). "Smart Spam Detection: An AI-Based Machine Learning Framework." International Journal for Multidisciplinary Research.

[14] Thota, P., & Nilizadeh, S. (2024). "Attacks against Abstractive Text Summarization Models through Lead Bias and Influence Functions." Findings of the Association for Computational Linguistics: EMNLP 2024.

[15] Tang, R., et al. (2025). "ADVERSARIAL TRAINING STRATEGIES FOR ENHANCING THE SECURITY OF LARGE LANGUAGE MODELS." Journal of Information Security.

[16] Sanh, V., et al. (2022). "Multitask Prompted Training Enables Zero-Shot Task Generalization." International Conference on Learning Representations.

[17] Cohen, W. W. (2015). "Enron Email Dataset." Carnegie Mellon University.

[18] Almeida, T.A., et al. (2013). "Towards SMS Spam Filtering: Results under a New Dataset." International Journal of Information Security Science.

[19] Tang, K., et al. (2023). "SpamDam: An End-to-End Framework for Privacy-Preserving and Adversary-Resistant SMS Spam Detection." Proceedings of the ACM SIGSAC Conference on Computer and Communications Security.

[20] Devlin, J., et al. (2019). "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.

[21] Madry, A., et al. (2018). "Towards Deep Learning Models Resistant to Adversarial Attacks." International Conference on Learning Representations.

[22] Goyal, P., et al. (2025). "LLM-Powered Intent-Based Categorization of Phishing Emails." arXiv preprint arXiv:2506.14337.

[23] Li, J., et al. (2025). "Quantification of Large Language Model Distillation." arXiv preprint arXiv:2501.12619.

[24] Fang, Y., et al. (2021). "Dual Gating: A Dynamic Gating and Routing Mechanism for Efficient CNNs." IEEE Transactions on Neural Networks and Learning Systems.

[25] NVIDIA. (2024). "LLM Benchmarking: Fundamental Concepts." NVIDIA Developer Blog.