

Explainable Detection of AI-Enhanced Phishing Emails and URLs

MSc Cybersecurity and Digital Forensics
Md Obaydul Haque | Student ID: 2435442 | Supervisor: Dr Amit Naik | 2025/26

Background and Research Problem

Cybersecurity problem

Phishing exploits trust, urgency and contextual plausibility.

GenAI Uplift

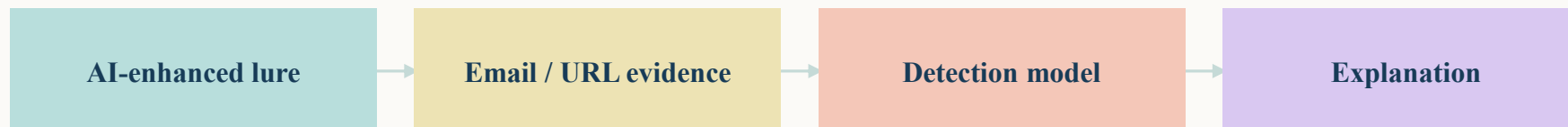
LLMs can generate fluent, personalized and credible lures at scale.

Detection challenge

Traditional language cues become less reliable as messages become more polished

Core problem in one line

- High benchmark accuracy does not necessarily imply real-world robustness, particularly under distribution shift or changing attack behavior.



Aim, Objectives and Evaluation Question

Project aim

To design and evaluate an explainable machine-learning framework for phishing-related email and URL detection.

Evaluation question

Which models perform best, and do their explanations reflect plausible security indicators rather than dataset-specific shortcuts?

Objectives

- Review phishing, GenAI, transformer and XAI research
- Compare classical ML and DistilBERT for email classification
- Evaluate ML models for phishing-URL classification
- Explain model decisions using SHAP and LIME
- Evaluate using multiple security-relevant metrics



Literature Synthesis and Research Gap

Phishing detection has evolved from reactive rules towards data-driven contextual models.



Gap area	Why it matters
Benchmark performance	High internal scores may fall under distribution shift
Modality fragmentation	Email text and URL evidence are often studied separately
Dataset bias	Models can learn source artefacts instead of transferable attack cues
Interpretability	Accuracy does not show whether reasoning is security-relevant

Research Gap

Limited integrated evidence compares email and URL models while auditing meaningful versus dataset-specific signals.

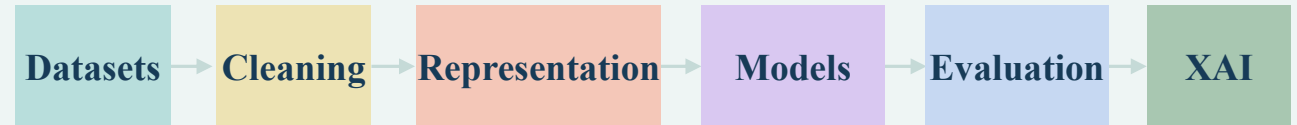
Literature implication

Compare model families + multiple metrics + explanations

Methodology and Experimental Framework

Area	Choice
Research methodology	Design Science Research (DSR)
Research approach	Positivist and deductive
Study design	Quantitative comparative experiment
Data split	Stratified 70% / 15% / 15%
Model selection	Validation F1-score
Final evaluation	Held-out test set

Framework flow



Design logic

Build and evaluate an artefact connecting email/URL data, multiple model families and explainability.

Reproducibility

Fixed random state and seed = 42

Ethical boundary

Public datasets only; no phishing campaigns or malicious payloads were created or deployed.

Data, Features and Models

Dataset	Used for	Size after cleaning	Class distribution
Email corpus	Email/spam classification	5,695 emails	4,327 legitimate / 1,368 spam
URL dataset	URL classification	57,405 URLs	26,908 legitimate / 30,497 phishing

Email Models	
TF-IDF	Logistic Regression
TF-IDF	Linear SVM
Tokens	DistilBERT

URL Models	
98 features	Gradient Boosting
98 features	Random Forest

Data Split

70% train | 15% validation | 15% test

Email Representation

TF-IDF words/bigrams; DistilBERT contextual tokens.

URL Representation

Lexical, structural and host-based numeric features

Scope note

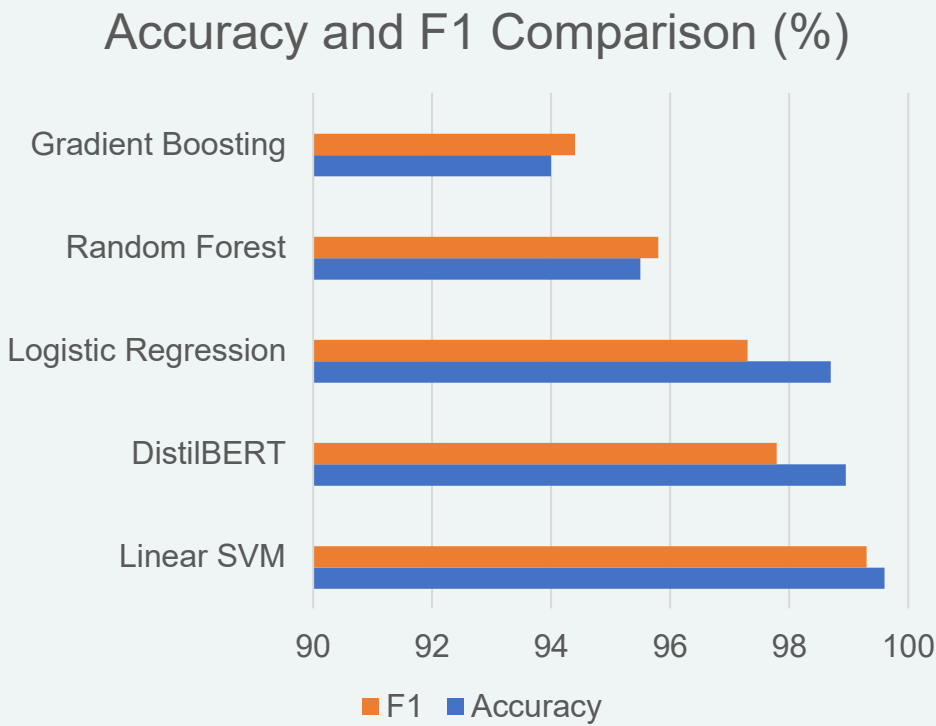
Email positive class represents spam broadly, not a separately verified AI-generated phishing class.

Comparative Model Performance

Track	Model	Accuracy	Recall	F1	ROC-AUC
Email	Linear SVM	99.60%	99.50%	99.30%	0.9998
Email	DistilBERT	98.95%	96.60%	97.79%	0.9994
Email	Logistic Regression	98.70%	95.60%	97.30%	0.9998
URL	Gradient Boosting	94.0%	95.0%	94.4%	0.9854
URL	Random Forest	95.5%	96.5%	95.8%	0.9916

Key finding

- Linear SVM was the strongest email model; Random Forest was strongest for URL classification.
- DistilBERT performed strongly but did not outperform Linear SVM.



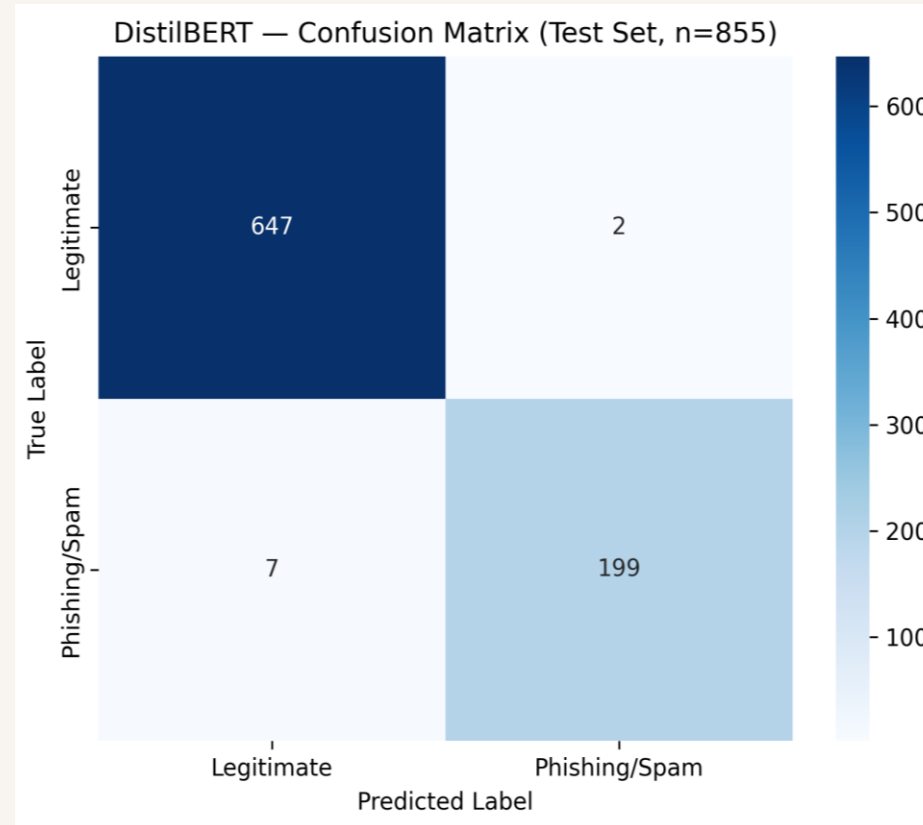
DistilBERT Fine-Tuning and Final Evaluation

Final Configuration

- **Model:** distilbert-base-uncased
- **Learning rate:** 3×10^{-5}
- **Batch size:** 16
- **Epochs:** 5
- **Max length:** 256 tokens
- **Seed:** 42

Key Finding

Strong performance, but no improvement over Linear SVM.
2 false positives | 7 false negatives

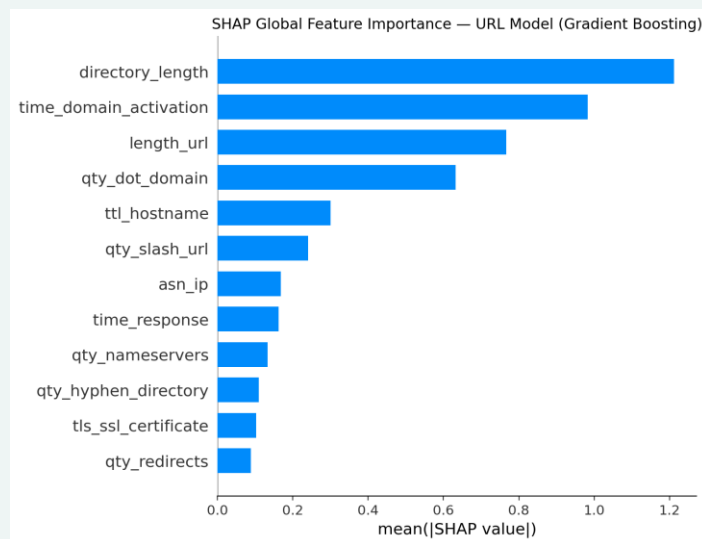


Held-Out Test Results

Accuracy	98.95%
Precision	99.00%
Recall	96.60%
F1-score	97.79%
ROC-AUC	0.9994
Test samples	855

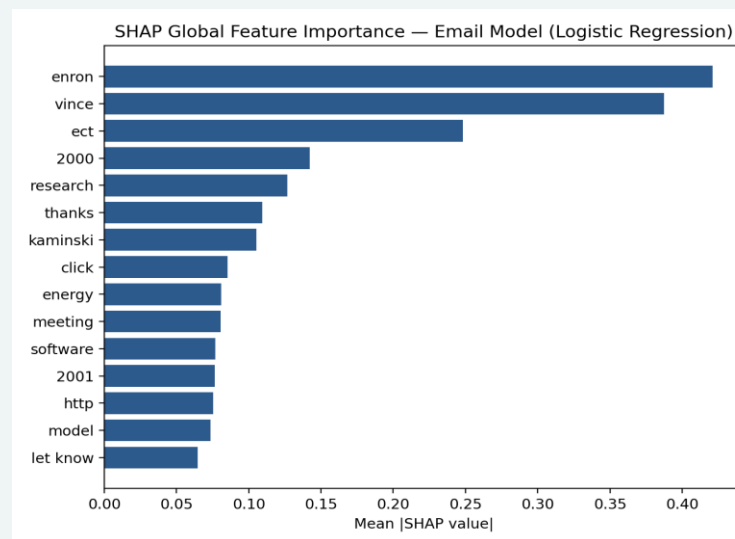
Explainability Findings: What Did the Models Learn?

URL Model: Security-Relevant Signals



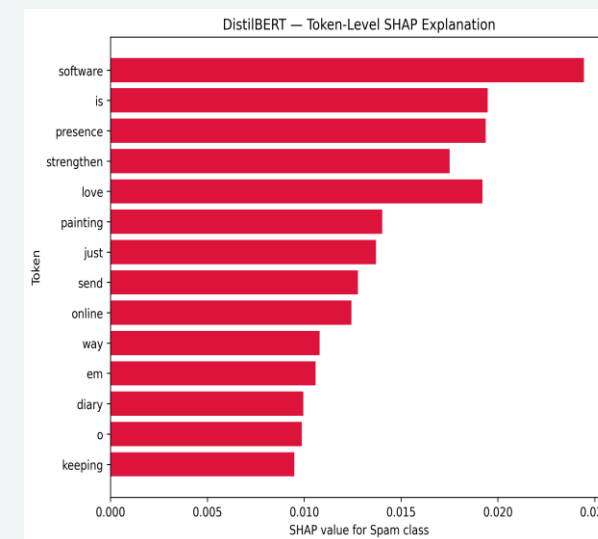
Top features: directory length, domain activation time, URL length and subdomain patterns.

Email Baseline: Dataset Shortcuts



Influential terms included enron, vince, ect and kaminski.

DistilBERT: Local Token Attribution



Software, online and send influenced the selected spam-labelled prediction.

Key XAI Finding

- Explainability revealed information that headline accuracy alone did not show.

Key Findings, Limitations and Future Work

Key finding 1

Linear SVM outperformed DistilBERT on the email dataset.

Key finding 2

Random Forest was the strongest URL model.

Key finding 3

XAI exposed dataset-specific shortcuts hidden by headline accuracy.

Limitations

- Email positive class is broad spam, not verified AI-generated phishing
- Evaluation was primarily within selected datasets
- Enron-specific terms indicate possible shortcut learning
- SHAP/LIME attribution does not establish causation

Future Work

- Use a contemporary, verified AI-phishing email corpus
- Perform cross-dataset and temporal validation
- Combine email-text and URL evidence
- Test robustness under concept drift and operational conditions

Project Contribution

- An integrated, reproducible framework comparing email and URL models with XAI-based auditing.

Conclusion and Demo

Key Takeaways

- Linear SVM was the strongest email model — 99.6% accuracy
- Random Forest was the strongest URL model — 95.5% accuracy
- DistilBERT achieved 98.95%, but did not improve on SVM
- XAI revealed meaningful signals and dataset-specific shortcuts
- Real-world robustness to AI-generated phishing remains unproven

What the project shows

Performance alone is not enough; explanation and generalisation matter

Demo Flow



References

- Schmitt, M. and Flechais, I. (2024) “Digital deception: Generative artificial intelligence in social engineering and phishing”, *Artificial Intelligence Review*, 57, 324.
- Opara, C., Modesti, P. and Golightly, L. (2025) “Evaluating spam filters and stylometric detection of AI-generated phishing emails”, *Expert Systems with Applications*, 276, 127044.
- Safi, A. and Singh, S. (2023) “A systematic literature review on phishing website detection techniques”, *Journal of King Saud University - Computer and Information Sciences*, 35(2), pp. 590–611.
- Al-Subaiey, A. et al. (2024) “Novel interpretable and robust web-based AI platform for phishing email detection”, *Computers & Electrical Engineering*, 120, 109625.
- NCSC (2024) The near-term impact of AI on the cyber threat.
- Meléndez, R., Ptaszynski, M. and Masui, F. (2024) “Comparative investigation of traditional machine-learning models and transformer models for phishing email detection”, *Electronics*, 13(24), 4877.
- Rashid, F. et al. (2024) “Phishing URL detection generalisation using unsupervised domain adaptation”, *Computer Networks*, 245, 110398.
- Calzarossa, M.C., Giudici, P. and Zieni, R. (2023) “Explainable machine learning for phishing feature detection”, *Quality and Reliability Engineering International*.
- Vrbančič, G. (2020) Phishing websites dataset. Mendeley Data, Version 1.
- ENISA (2025) ENISA Threat Landscape 2025.

Thank You...
