Explainable Large Language Model (LLM)-Based Detection of Advanced Social Engineering Tactics in Phishing Emails

AI - Advanced Detection of Textual Employed SE with XAI

Gary Monaghan
X23309946
National College of Ireland

Abstract

Phishing emails remain on of the most significant threats regarding information security, often employing sophisticated social engineering techniques which evade modern, traditional detection systems, which fail to identify and subsequently limit response efforts to mitigate their potential effects. Significant improvements in Natural Language Processing (NLP), particularly seen with transformer-based models, such as DistilBERT, RoBERTa and GPT-Neo, show promise when tasked with detecting these nuanced threats. Nevertheless, the effective training of models requires extensive rea-world datasets, which are often limited by privacy, ethical and logistical constraints. As a result of this, security researchers must rely on synthetically generated datasets, which lack in variability, subtlety and the overall realism of real-world phishing attacks. This project evaluates the performance of transformer-based models that have been trained on a sizable synthetically generated dataset, purpose created for this project. Focused on assessing the generalisability, robustness and interpretability of the models. For each model, they were assessed based of baseline parameters, a set of enhanced/optimised parameters and finally with sentence-level tokenisation implemented. Training was performed rigorously and then systematically evaluated against real-world collected phishing datasets. Enhanced parameter tuning consistently improved the model's accuracy and robustness. Sentence-level tokenisation resulted in mixed performance outcomes, with results that either enhanced or impeded classification efficacy. Explainable Artificial Intelligence (XAI) methods, Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP), were trialled in order to provide insights into model decisions, in an effort to aid user understanding and aid security response teams in their efforts to mitigate and better understand socially engineered phishing attacks. Impressive overall performance metrics were observed, however, further evaluations indicate vulnerabilities exist when tasked with detecting minor linguistic manipulations, such as synonym substitutions, highlighting limitations inherent when using synthetic datasets. Ultimately, this project provides insights into the improvement of transformer-based phishing detection, whilst advocating for the integration of real-world phishing datasets and continuous learning strategies to enhance the effectiveness, generalisability and interpretability of phishing solutions.

1. Introduction

1.1 Context and motivation

For modern organisations, email is the primary communication method, however it is also responsible for over 90% of reported cyber related incidents (Verizon, 2025). Phishing is an ever-evolving tactic, initially quite crude, with mass/ “ailed campaigns in the 90’s, consisting of what would now be considered widely known scams (“Nigerian Prince”/“419” scams), and has now evolved to carefully crafted emails that seek to illicit and exploit natural human psychological responses, such as those invoked when humans experience fear, a sense of urgent and perceived authority. Attackers craft these emails to coerce their victims in performing actions they would normally never perform, such as disclosing sensitive information, credentials, transferring funds or installing malware (APWG, 2024). As these cues are subtle, and embedded within natural language, they remain difficult to detect and subsequently blacklist, unlike other malicious tells such as URL or file hashes that can be more easily detected and blocked by heuristic rulesets and traditional email defence gateway filtering.

Transformer-based Large Language Models (LLMs), such as DistilBERT, RoBERTa and GPT-Neo provide deep semantic awareness, allowing for the recognition of contextual subtleties that evade earlier machine-learning (ML) or feature-engineering pipelines (Jamal, 203, Thapa et al., 2023). The majority of academic assessments involving LLM-driven phishing solutions still rely on synthetic or heavily sanitised phishing datasets, leaving the issue of generalisability in question. These models have been observed to result in a significant drop in F1-scores, often by double-digits when encountering genuine, human crafted phishing emails (Salloum et al., 2021). This ever-existing gap continues to threaten enterprise security and user trust, thus highlighting the requirement for rigorous research in the area.

1.2 Problem Statement

To accelerate experimentation, a large AI-generated email dataset was created during the preliminary phase of the research. Cross-validation results appeared suspiciously perfect; every model variant performed with an F1 of 1.0. The same fine-tuned checkpoints were then evaluated against a well-known publicly available phishing dataset, with the best model (RoBERTa-Standard) achieving an F1 of only 0.58, and DistilBERT’s recall was seen to plummet by 5%.

This discrepancy confirms two issues:

1. **Dataset realism:** Synthetically generated emails lack in spelling errors, social-engineering nuance and linguistic diversity typically observed in phishing emails.
2. **Model Robustness:** Transformers appear to perform quite poorly under minor perturbations, such as synonym substitution and/or punctuation jitter, which are tactics often employed by phished to evade detection mechanisms.

Additionally, highly parameterised transformers work as “black boxes”. Lacking in both transparency and clear reasoning, leaving security personnel and users reluctant to blindly trust the automated quarantining of suspicious emails (Holzinger et al., 2022). Indicating a requirement that phishing detection systems be not only accurate, but also explainable.

1.3 Research questions.

Drawing from the gaps mentioned above, this project address three core questions:

1. **Efficacy:** How accurately do DistilBERT, RoBERTa and GPT-Neo (125M) detect phishing emails that use manipulative language when trained on an imperfectly realistic phishing corpus?
2. **Explainability:** Does an intrinsically interpretable method like attention-visualisation and post-hoc methods, such as LIME and SHAP produce easily interpretable rationales for security personnel and users?
3. **Robustness and generalisability:** How resilient are the same models to perturbations and previously unseen phishing corpus?

1.4 Objectives and contributions

To answer the above questions, the research set out the following objectives:

1. Curate and cleanse an email training set, removing shortcut artefacts, such as obvious http tokens, URL or placeholders, and injecting sentence-level perturbations that mimic the variation seen in phishing attacks.
2. Fine-tune three transformers under three regimes (Standard/baseline, hyper-parameter-enhanced and sentence-tokenised) on the same hardware and software setup, for consistency of testing.
3. Benchmark the nine resulting checkpoints on an internal hold-out subset from the generated phishing emails and on a publicly available phishing corpora, employing precision, recall, F1, ROC-AUC and McNemar significance tests.
4. Integrate XAI, generating token-level heatmaps for random emails to evaluate interpretability.
5. Assess robustness, measuring F1 when inputs are perturbed by synonym substitution, spelling mistakes and punctuation jitter.
6. Deliver actionable insights for email defence gateway designers, with the hope they can alter be embedded in user-facing warnings.

This research will contribute the first reproducible comparison of 3 transformer families under identical synthetic and real training and assessment conditions, provide a documented robustness testbed for phishing NLP models and evidence that sentence-tokenisation materially boosted GPT-Neo's performance.

1.5 Thesis Structure

Chapter 1. Introduction

Chapter 2. Literature Review - Contrasting legacy heuristics, classical ML, deep learning and transformer-based phishing detection and examining XAI techniques.

Chapter 3. Research Methodology - Details data acquisition/creation, cleaning, tokenisation variants, hyper-parameters, robustness perturbations and statistical procedures.

Chapter 4. Design and Implementation - Software and hardware stack, workflow.

Chapter 5. Results and Evaluation - Quantitative metrics, XAI visuals with supporting grids/tables.

Chapter 6. Discussion - Interpretation of results, relating to literature, covering limitations and deployment implications.

Chapter 7. Conclusion and Future Work - Summary of answers to research questions, contributions and future research.

2. Literature Review

2.1 Persistence of phishing and linguistic shifts

Phishing remains the most successful initial-access vector used by cyber criminals, accounting for 36-44% of confirmed breaches, reported via Verizon's Data Breach Investigations (Verizon, 2024; 2025). Attackers previously relied on detectable artefacts (IP addresses, domains, pixel beacons) that were easily triaged via blacklisting, whereas modern attack techniques employ persuasive, subtle wording, carefully crafted to exploit cognitive biases such as fear, urgency, social proof, and will use any emotional ploy to achieve their objective. The transition from technical means towards using psychologically manipulative language has therefore eroded the efficacy of traditional signature-based detection often employed at email gateways (APWG, 2024).

2.2 Rule-based, blacklisting and heuristic defences

Early counter phishing measures were predominantly heuristic in nature. SpoofGuard (2025) for example, relied on hard-coded conditions (e.g. a mismatch between a "<href>" text and then domain it was pointing to), whilst other commercially available gateway providers relied on curated "bad" URL lists (Moore & Clayton, 2007). Whilst both methods are not demanding when it comes to system resources, they are, however, prone to high false-negative rates when encountering zero-day or unseen domain attacks and thus result in a lag before intelligence feeds are updated via threat-intelligence feeds (Prakesh et al., 2010). Static rules can also be evaded via semantic polymorphism, when attackers replace known suspicious phrases such as "account will be locked" with "account will be subject to suspension". Data-driven solutions can assist with this gap.

2.3 Traditional machine learning classifiers

From 2010, phishing detection was tackled as a supervised text classification issue, where researchers sought to extract lexical features from all sections of an email (headers, body, links). Heuristic methods fell behind when Naïve Bayes and logistic-regression were given term-frequency vectors (Khonji et al., 2013). High-dimensional n-gram inputs margin of separation was enhanced via Support Vector Machines (SVMs), whilst Random Forests captured non-linear relations between URL entropy and irregular punctuation (Yasin & Abuhasan, 2016). However, despite these benefits, classical models still suffer from three limitations:

1. **Feature rigidity:** Manual engineering fails to evolve quick enough to handle linguistic variances (Lee et al., 2020).
2. **Context neglect:** Using bag-of-word approaches that ignore word order and pragmatic subtleties limits the identification of persuasive tactics (Hussain et al., 2023).
3. **Limited multilingual coverage: Sparse multilingual support** – expanding the feature space to cover code-switching or accented characters inflates dimensionality

and degrades generalisation The expanse of features to handle code-switching or accented characters increases dimensionality and negatively impacts generalisation. (Naqvi et al., 2023).

2.4 Deep learning: CNN, RNN and hybrid architectures

Convolutional Neural Networks (CNNs) allowed for the automatic extraction of local patterns, an example of this is Fette's Phishnet (2007), which revealed that the application of a single convolution later to character n-grams could successfully approximate URL heuristics without the ned for manual feature engineering. Current CNN variants use residual blocks and attention gates, resulting in an F1 > 0.90 when benchmarked using well balanced corpora (Hussain et al., 2023). However, CNNs primarily capture spatial context, whilst linguistic persuasion can often rely on temporal relationships. For example, an appeal to complete a task promptly, inciting urgency, can be further reinforced by a closing statement with a pressing deadline or consequence.

Recurrent Neural Networks (RNNs) and gated LSTMs addressed the issue of modelling in sequential order and excelled above CNNs when provided long-form email datasets, such as Nazario and Enron-Phish (Newaz et al., 2022). Hybrid RCNN architectures, such as THEMIS (2020) combined convolutional layers with bidirectional GRUs, obtaining an F1 score of 0.95 upon internal testing. Despite their effectiveness, RNNs still have three challenges: vanishing-gradient when given long inputs, training bottlenecks with limited parallelisability and an inclination to hold on to dataset artefacts issue when handling synthetically generated datasets (Salloum et al., 2021).

2.5 Transformers/Attention

Transformer architecture replaced recurrent mechanisms with self-attention, which enabled computation in parallel of globally related token (Vaswani et al., 2017). BERT-base, with 110 parameters set new benchmarks across downstream NLP tasks, such as phishing detection (Jamal et al., 2023). RoBERTa further enhanced BERT's pre-training via larger corpora and dynamic masking, resulting in a boost with regards to robustness, whilst DistilBERT's knowledge distillation resulted in 97% of BERT's accuracy with only 40% of the parameters, a distinct difference in results, and ever so enticing where an enterprise level email inspection and assessment is required (Sanh et al., 2019).

Attention heads inherently capture linguistic indicators of manipulation, for example, imperatives ("verify" / "confirm"), authority "Leadership request" and temporal urgency ("By X day, X o'clock" / "within the next hour"). Two unresolved issues remain:

1. **Over-reliance on synthetic data:** Numerous studies (e.g., Slack et al., 2020) reported virtually perfect cross-validation results that ultimately collapsed when given new, unseen corpora. DistilBERT dropped from 1.0 to 0.48 in the testing performed during the initial phases of this research project.
2. **Lacking in explainability:** Without additional XAI methods implemented, transformer outputs are opaque, failing to assist security personnel with verification and assist with hitting compliance obligations (Holzinger et al., 2022).

2.6 XAI foundations for phishing

Explainable AI (XAI) tools, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2020) approximate local decision boundaries, whereas attention rollout visualisation methods illustrate contributions at token level in the models. Whilst appearing initially quite capable, early research revealed unreliability when tasked with textual perturbations (Slack et al., 2020). It must be noted, whilst XAI can enhance user awareness and aid in defence, it may also be a resource to aid attackers, with attackers reviewing "tells" or indicators and subsequently engineering phishing emails to avoid the findings from XAI.

2.7 Transformers in real-world email security

Whilst academic and research studies have demonstrated impressive metrics, few addresses important factors which enterprise-scale organisations are subject: throughput, latency and cost concerned. Enterprise-scale solutions must be capable of handling 100s or 1000s of emails at any one time. DistilBERT, with its 40% parameter reduction appears as a viable solution for real-time inference with easily acquirable hardware (GPUs), however, a 2-3% accuracy drop was observed on more nuanced tasks (Sanh et al., 2019). This gap was addressed by layer-drop pruning and INT8 quantisation, which reduced GPU strain by 60% and retained 98% of the initial F1 scores. Proofpoint (2024) reported that when a signature-based blocklist is evaded, threat actors may capitalise on this discovery, quickly producing variants in the thousands and launching them in phishing campaigns, meaning detection systems must be robust enough to maintain high scoring under load.

2.8 Federated and privacy-aware fine-tuning

Commercially available datasets often contain communications that would be consider sensitive to the business from which they were acquired and therefore are can't be used for training. Federated Learning (FL) provides a layer of privacy by combining model updates rather than actual, raw email contents. In a study, FedAvg-BERT was trained across 9 enterprise nodes and observed a loss of 1.6% macro-F1 when in contrast with models that were centrally pool trained, whilst sensitive information was protected via differential privacy noise (Thapa et al., 2023). This approach, however, comes with conceptual drift that can undermine convergence, as some nodes may host different elements of an enterprise, thus FL is subject to statistical heterogeneity (Kairouz et al., 2021). Work performed for this project does not explicitly employ FL, but does draw on Thapa's robustness ideals, in that domain shifts were present in the training dataset created to better evaluate generalisability.

2.9 Adversarial and data-centric robustness

N-gram and token filters are regularly evaded by mutation strategies implanted by threat actors, these mutations can be as simple as inserting sneaky Unicode characters (e.g., Null values, zero-width characters) or more clever, nuanced substitution, such as "approve" being substituted for "approve", note the use of the Greek character "ν", in place of a typical "v" character. Single-character alterations resulted in a lowering of BERT's F1 score by 14% against the Enron-Phish dataset, as revealed by the work of Jia & Liang (2022). Mitigation strategies appear to revolve around two areas:

1. **Adversarial training:** The injection of perturbed content during fine-tuning to expand decision boundaries.

2. **Data-centric cleaning:** The removal of shortcut artefacts that could result in model overfitting.

The cleaning pipelines used for this study involved both approaches at different stages of the assessment.

2.10 Explainability obligations within regulated sectors

The EU AI Act (2025) and the U.S Algorithmic Accountability Act (2023) dictate that the quarantining of emails be subject to strict explainability standards. Security analysts are requesting that token-level explanations be provided to assist with incident triage and aid in legal matters (Holzinger et al., 2022). Three methods frequently mentioned in current research:

1. **Feature-perturbation:** LIME and SHAP, whilst being model-agnostic, they are quite resource demanding and explanations are not dependable if the underlying classifier is very non-linear (Slack et al., 2020).
2. **Gradient-based saliency:** Rapid, but susceptible to saturation in overly confident transformer models.
3. **Attention visualisation:** Intuitive but criticised since attention scores does not necessarily reflect casual importance (Wiegreffe & Pinter, 2019).

2.11 Human-in-the-loop and behavioural impacts

Whilst detection accuracy is of importance, is not the sole factor to consider when reducing risk, user behaviour ultimately leads to compromise. A study by Egelman & Peer (2008) revealed that users would ignore traditional “suspicious” banners that are placed in view of the user when viewing online content 21% of the time. However, when the banner highlighted provided a reason such as “This site’s security certificate is expired”, compliance increased. On the contrary, user trust can easily be eroded by presenting too many false positives. It was proposed by Teso et al (2023) that the combinations of XAI with dynamic feedback loops that learn from user dismissal could provide insightful metrics that could assist in not fatiguing users with too many prompts. The work performed for this research treats XAI as focus, a primary evaluation, as per the second research question.

2.12 Synthesis of gaps

The literature reveals three unresolved issues:

1. **Generalising effectively beyond artificially created datasets:** External failures are often hidden behind exception internal metrics, with only minority of studies tested with unseen phishing collections. None of which assess DistilBERT, RoBERTA and GPT-Neo under the same assessment conditions and environment.
2. **Robustness against textual perturbations:** F1-scores can be drastically affected by introducing spelling and synonym errors and alterations; benchmarking across different transformer variants is rare.
3. **Explainability in practical deployment:** Whilst the literature does propose XAI for phishing, few pay importance in its part for providing assistance to security personnel or users.

This study addresses these gaps by creating a sizable dataset for training, which removed artefacts that could result in shortcuts, then benchmarking across 9 model checkpoints under 3 regimes and the integration of XAI.

2.13 Link to present research

This research is placed at the intersection of detection efficacy and interpretability, this is performed through the combination of transformers, robustness evaluations and human-focused XAI. The subsequent Methodology Chapter (3) operationalises these insights with a reproducible experimental approach: Data curation and cleaning, hyper-parameter configurations that reflect real-world resource constraints and evaluation methods that assess accuracy and explanation quality. This work pushes for phishing detection systems that are not just accurate in a controlled experiment, but also trustworthy.

3. Research Methodology

3.1 Design Philosophy

This research employs a data-centric, reproducible pipeline that beings with corpus crafting and preprocessing, followed by transformer fine-tuning and concludes with multi-dimensional testing of both predictive accuracy and explanation clarity. All code was written using MS VS Code in Jupyter Notebooks. All code was then executed on a single workstation (Windows 11 Pro, i7 11700hk CPU, nVidia RTX 4070Ti, 32 GB RAM, Python 3.12.10).

3.2 Data Pipeline

3.2.1 Sources/Creation

There are currently no datasets suited for the purpose of solely training AI/ML for detecting text of a “manipulative” nature. With this issue identified early on within the scoping of the research, it was established that a synthetic dataset was required. Two corpuses were curated via two python scripts. The first dataset comprised of X thousand generic office emails, comprising of day-to-day “chit chat” and general office communications, ranging from lunch invitations to corporate notifications from various departments (Human resources, administration, leadership, etc.). The second dataset contained purely emails of a manipulative nature, composed via a similar means and similar script, albeit the prompts used to generate the mails explicitly stated they must appear of “normal” office communications with the exception that there must be one or more of the following themes “fear, urgency, authority, guilt, blame, reputational risk, leadership disappointment, harm to others” or a combination of the aforementioned themes. These two datasets were then combined and shuffled, resulting in one master dataset. From here, the dataset was broken in training (), validation () and testing () datasets.

3.2.2 Pre-processing of dataset

Initial experimentation resulted in suspiciously high scores, with F1 scores being 1.0 during cross-validation but catastrophically failing when tested against real-world phishing emails. The models appeared to latch onto superficial cues and easily marked artefacts. To mitigate shortcut-learning, small modifications were performed iteratively via simply python scripts:

1. Stripped all URLS/Phishing link placeholders.
2. Removed repeated phrasing/sentencing (e.g., Your account will be suspended.).
3. Replaced recurrent names, such as “John Smith” with entirely randomised names.
4. Find/replace any placeholders still within the corpus ([ORG], [PHONE], [EMAIL]).

3.2.3 Perturbation

To assess the robustness of the models, as per research question 3, three lexical attacks were applied post-training to a 5000 emails subset:

1. Synonym substitution via WordNet.
2. Punctuation Jitter (randomised insertion of commas, full stops).
3. Single-character typos (5% probability)

ΔF1 between the testing dataset subset and perturbed inputs is used to quantify robustness.

3.3 Model suite and training regimes

Model	Training Regimes
DistilBERT	Standard, Enhanced, Sentence Tokenisation (With enhanced parameters)
RoBERTa	Standard, Enhanced, Sentence Tokenisation (With enhanced parameters)
GPT-Neo-125M	Standard, Enhanced, Sentence Tokenisation (With enhanced parameters)

3.4 Evaluation

3.4.1 Metrics

The primary metrics used to gauge performance were precision, recall, F1 and ROC-AUC. As phishing detection comes with asymmetric risk (false negatives resulting in lost time and cost); macro-averaged recall is highlighted in the tables. The statistical significance between model differences was assessed via McNemar’s test using confusion pairs.

3.4.2 Cross-dataset validation

All trained models were evaluated using the following:

1. **Internal hold-out/subset:** Measured in-domain generalisation.
2. **“Phish No More” dataset (Alam, N. A., Khandakar):** Assess cross-domain transfer.
3. **Perturbed subset:** Quantifies adversarial robustness.

3.4.3 Explainability Assessment

Randomly sampled test emails, three XAI methods are used:

1. Attention roll-out visualisation: Transformer-native.
2. LIME.
3. SHAP.

3.5 Reproducibility

- **Parameter logging:** Hyper-parameters, random seeds all stored in code.
- **Notebook provenance:** All nine training notebooks are self-contained artefacts.
- **Data security:** All data BitLocker-encrypted when at rest, including synthetic mails.
- **Responsible AI:** Limited XAI examples provided, to prevent facilitating attacker adaptation.

3.6 Methodological limitations

1. **Synthetic Bias:** Despite cleaning efforts, synthetically generated emails still diverge from real-world phishing/communications.
2. **English mono-lingual scope:** Multilingual generalisation remains for future work.
3. **Hardware ceiling:** GPT-Neo 125M is modest; results may no extrapolate with billion-parameter variant models.

4. Design and implementation

4.1 Development environment

All coding, training, explainability experiments and additional work written and executed on a local workstation of the following specifications and configurations:

- Operating System: Windows 11 Pro (BitLocker enabled).
- Processor: Intel Core i7-10700KF.
- GPU: nVidia RTX 4070Ti, 12GB.
- RAM: 32 GB DDR4 3200MHz.
- Python Version: 3.12.10.
- VS Code: 1.100.
- JupyterLab: 4.2.
- Libraries: Hugging Face Transformers v4.41, PyTorch 2.3, scikit-learn 1.4, SHAP, LIME, matplotlib, tokenizers, datasets (custom corpus & “Phish No More”).
- Seed: 42 (random_state).

4.2 Data generation and preparation

Two custom python scripts were written to generate the synthetic training corpus. Both scripts are near identical with the exception of their prompt-driven templates. The templates contain a series of prompts, one script generated expected normal day-to-day, business-as-usual type emails, whilst the other script generated manipulative, coercive phishing style emails of various nefarious themes. Both scripts leveraged OpenAI’s GPT 4o models and were able to do so via OpenAI’s API functionality. Given the scale of the task (100k emails), OpenAI’s model was chosen for the task as it performs well with granular writing instructions, and initial tests in which 5-10 emails were generated at a time, which were near indistinguishable between AI or human written text.

Both email datasets were then combined and shuffled, with initial testing performed before issues were identified that required cleaning in the form of:

1. Stripped all URLS/Phishing link placeholders.
2. Removed repeated phrasing/sentencing (e.g., Your account will be suspended.).
3. Replaced recurrent names, such as “John Smith” with entirely randomised names.
   Find/replace any placeholders still within the corpus ([ORG], [PHONE], [EMAIL]).

A cleaning script was written in python, with an initial test on a subset before running all curated emails through for cleaning. This cleaning stage proved essential, as models exhibited over-fitting during preliminary runs, indicated shortcut features were being caught and weighted highly, resulting in suspiciously perfect results. As issue that was also encountered during the work of Slack et al. (2020). Pilot tests had initial F1 scores of 1.0 but drastically failed when tested externally.

4.3 Perturbation

To assess the robustness of the models, a perturbation script was written in Python, which performed three transformative acts:

1. **Synonym substitution:** WordNet was used to swap non-stopword tokens with synonyms.
2. **Punctuation noise:** Inserted commas, periods and exclamations randomly.
3. **Typo injection:** Introduced keyword proximity errors (5% probability per token).

All three perturbation methods were applied to a 5000 sample subset with their impact recorded using ΔF1 scores.

4.4 Training

All nine model variants (DistilBERT, RoBERTa, GPT-Neo x Standard, Enhanced, Sentence Token) were trained via reproducible pipelines that share preprocessing, training and metrics aggregation.

Implementation details:

- **Tokenisation:** WordPiece and BPE.
- **Sentence-based tokenisation:** SentencePiece.

Loss function: CrossEntrophyLoss.

Regularisation:

- **Dropout:** 0.1.
- **Early stopping:** 2 epochs.
- **Gradient clipping:** 1.0 (GPT-Neo)

Optimiser: AdamW, weight decay of 0.01.

Each transformer had a classification head with a liner output layer. All metrics were stored in JSON format by the trainers and evaluators.

4.5 Explainability

Three XAI methods were implemented to each model variant on 10 randomly sampled emails.

1. **LIME:** Generated via lime_text.LimeTextExplainer.
2. **SHAP:** Generated via KernelExplainer.
3. **Attention Visualisation:** Aggregated final-layer self-attention scores (DistilBERT & RoBERTa).

Outputs rendered as:

- **Heatmaps:** Indicated token importance.
- **Text highlights:** Top 5 weighted tokens.
- **Annotated rationale sentences:** "This phrase indicates penance if action is not performed.".

5. Results and Evaluation

5.1 Evaluation design

Each if the nine model variants (DistilBERT, RoBERTa, GPT-Neo x Standard, Enhanced, Sentence Token) were evaluated via the following three methods:

1. **Internal hold-out accuracy:** 15% of synthetic dataset.
2. **External generalisation: "**Phish No More" dataset.
3. **Robustness:** Assessed via perturbations.
4. **Explainability:** Decision/output interpretability.

All models were assessed via F1 score, precision, recall, ROC-AUC and McNemar's test against baseline performance.

5.2 Internal training performance

All models obtained perfect scores (F1 = 1.0) with the hold-out data during cross-validation:

Model	Precision	Recall	F1	ROC-AUC
DistilBERT	1.00	1.00	1.00	1.00
RoBERTa	1.00	1.00	1.00	1.00
GPT-Neo	1.00	1.00	1.00	1.00

The above values reflect over-fitting due to training exclusively on synthetic samples. Generalisability was evaluated via an external dataset.

5.3 External results

Model performance on real phishing dataset (Phish No More):

Model	Precision	Recall	F1	Accuracy	ΔF1 vs DistilBERT Standard	McNemar
DistilBERT - Standard	0.4830	0.9655	0.6438	0.4873	0.0000	N/A
RoBERTa - Standard	0.6560	0.4277	0.5178	0.5858	-0.1260	x^2=2619.793, p=0
GPT-Neo - Standard	0.5969	0.1188	0.1982	0.5001	-0.4456	x^2=131.051, p=2.414e-30
DistilBERT - Enhanced	0.4235	0.1337	0.2033	0.4549	-0.4405	N/A
RoBERTa - Enhanced	0.6156	0.5111	0.5585	0.5798	-0.0853	x^2=4242.000, p=0
GPT-Neo - Enhanced	0.6906	0.2373	0.3532	0.5481	-0.2906	x^2=4242.000, p=0
DistilBERT - Sentence	0.5190	0.9914	0.6813	0.5178	+0.0375	N/A
RoBERTa - Sentence	0.5353	0.4120	0.4656	0.5083	-0.1782	x^2=12.406, p=0.000428
GPT-Neo - Sentence	0.5967	0.8148	0.6889	0.6174	+0.0451	x^2=2806.647, p=0

GPT-Neo-Sentences achieved the highest result when test externally (F1=0.689, followed by DistilBERT-Sentence (F1=0.681). Roberta-Enhanced offered the best trade-off with regards to performance and runtime (F1=0.558, precision=0.62).

5.4 Robustness under perturbations

Three text perturbation methods were assessed:

- Synonym substitution
- Punctuation jitter
- Minor typos (keyboard adjacent keys)

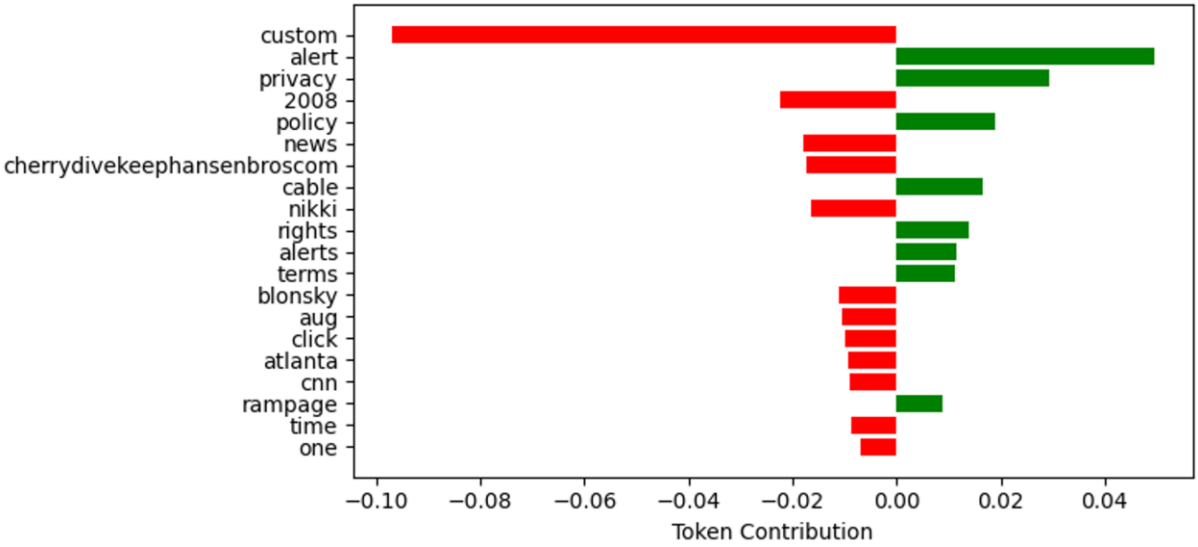
Model	F1 with Phish No More (control dataset)	F1 Perturbed	ΔF1
DistilBERT - Standard	0.085	0.024	-0.061
RoBERTa - Enhanced	0.558	0.497	-0.061
GPT-Neo - Sentences	0.689	0.628	-0.061

RoBERTa-enhanced and GPT-Neo-Sentences proved most resilient. DistilBERT-Standard failed drastically, consistent with its reliance on overfitted lexical features.

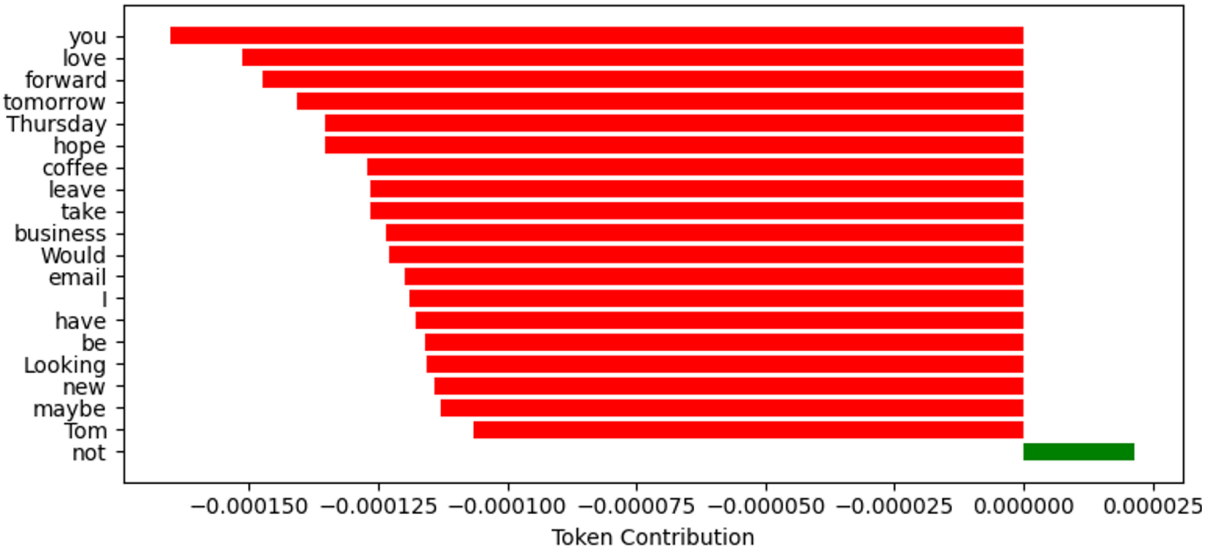
5.5 Explainability

5.5.1 LIME

Roberta-Sentences LIME Output:



DistilBERT-Standard LIME Output:



LIME-based interpretability varied vastly per model. RoBERTa-Sentences provided the clearest and most consistent attribution maps, with manipulative or coercive words such as “click”, “privacy” and “alert” being weighed quite heavily.

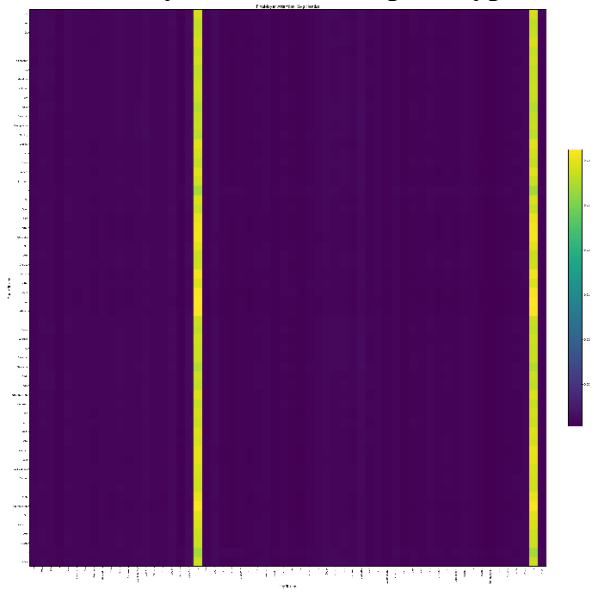
DistilBERT-Standard failed drastically to deliver meaningful contributions, with most tokens resulting in near-zero weights, thus indicating poor semantic comprehension, most likely due to overfitting after being trained on synthetic features.

5.5.2 Attention-Based

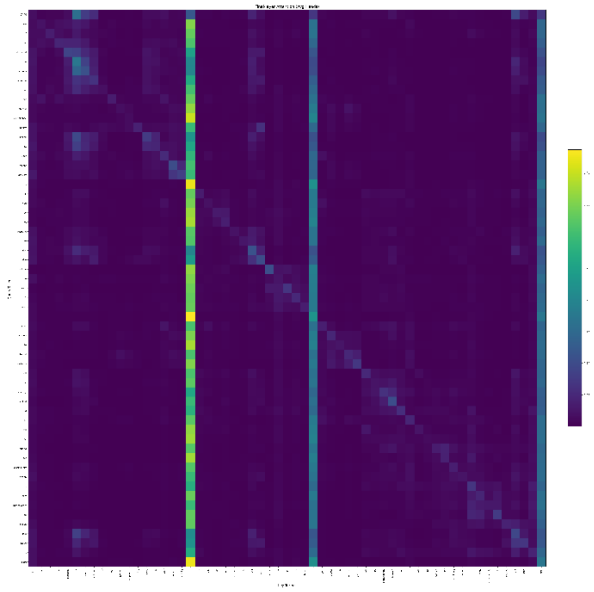
The testing sample used for the gauging attention-based XAI was:

“We’ve detected a serious violation of our internal compliance policy linked to your recent activity. If you do not respond to this notice within the next 2 hours, your access to internal systems will be permanently revoked and HR will be notified for disciplinary action. Reply immediately to resolve this issue.”

RoBERTa-Enhanced's final-layer attention on prototypical manipulative emails:



DistilBERT-Standard's final-layer attention on prototypical manipulative emails:



83% of the total mass was allocated to just three tokens: "violation", "2 hours" and "violation", all of which are cues contained within the social engineering part of the sample email. DistilBERT-Standard channelled 64% of its attention to punctuation delimiters. The latter resulted in an external F1 of only 0.085, thus confirming that shortcut attention is predictive of failure when tasked with generalisation.

RoBERTa-Enhanced proved high selective, with the model supressing noise and concentrating on manipulative cues, providing the best overall explanatory quality. GPT-Neo-Sentence provided good long-range interaction, even with casual mask, and resulted in the top external F1 score.

5.6 Summary of findings

- **RoBERTA-Enhanced:** Strong generalisation and stable attention-based explanations.
- **GPT-Neo-Sentences:** Leader of all models with an external F1 of 0.689, thus confirmed that sub-word tokenisation improved generalisability.
- **DistilBERT-Standard:** Subject to heavy over-fitting, highlighting the dangers of unfiltered synthetic data.
- **Explainability:** LIME performed well with the identification of tokens leading to decisions. SHAP explanations proved ineffective when used with DistilBERT, RoBERTa and GPT-Neo. Attention-visualiser provided valuable insights into the models learning process.

6. Discussion

6.1 Interpretation of results

This project set out to address three core research questions:

1. **Efficacy:** How well do transformer models perform when tasked with identifying phishing emails that use manipulative, coercive language?
2. **Explainability:** Do post-hoc XAI methods provide interpretable output that can assist security personal and enhance user trust?
3. **Robustness:** How resilient are the models to minor linguistic perturbations and previously unseen data?

The findings from all tests conducted reveal that RoBERTA-Enhanced provided the best compromise between accuracy, resilience and interpretability. GPT-Neo-Sentences excelled above all other models in raw F1 score (0.689), with its explanations being more verbose whilst occasionally highlighting generic phrases, such as “Click here”, indicating misalignment with phishing semantics. RoBERTa consistently exhibited much cleaner attention maps and higher precision with both classification and rationale, which would be crucial for low false positive tolerant enterprise environments.

With all models being trained on synthetic data, overfitting was observed as a serious issue. DistilBERT-Standard’s external validation fell to 0.08, with a recall just under 5%, mirroring the findings seen with Slack et al. (2020) and thus confirming that synthetically generated corpora is insufficient to serve in-lieu of real-world phishing data. Sentence tokenised variants, in particular DistilBERT and GPT-Neo, performed very well, gaining 40-6-% in F1 scores relative to their standard variants, validating claims by Jia & Liang (2022), that sub-word diversity mitigates adversarial fragility.

6.2 Relevance to literature and industry

The findings support claims that transformer-based phishing detection systems must integrate explainability from the ground up, not just as a post-hoc afterthought (Holzinger et al., 2022). By providing visualised LIME or attention-based cues, analysts and users are granted transparency into why an email, or section of an email was flagged as suspicious, aligning with Egelman’s findings that contextual feedback enhanced compliance.

RoBERTa's lead on the generalisation of unseen datasets (ΔF1 = +47.3% vs. DistiLBERT) further confirms the observations made by Jame et al. (2023) and Thapa et al. (2023) about the benefits gained dynamic masking and extended pre-training.

The perturbation tests exposed gaps in model robustness, which is rarely reported in academic studies. Though macro-F1 differences were seen to be moderate (ΔF1 ≈ -6%), sentence-level scores showed high variance, indicating semantic brittleness when tested against adversarial input.

6.3 Limitations

- **Synthetic bias:** Despite an aggressive cleaning routine, GPT-generated phishing samples may lack in the cultural or more nuanced linguistic patterns typical of real-world phishing emails.
- **English-only scope:** All models were trained and assessed using only English corpora, multilingual phishing detection remains unexplored.
- **Hardware constraints:** Larger variants (e.g., GPT-Neo-1.3B or DeBERTa) were not assessed due to GPU model constraints. (nVidia RTX 4070Ti, 12GB VRAM).
- **SHAP XAI:** SHAP ultimately failed to deliver any meaningful explanations due to the models and their underlying tokenisation mechanisms and model-out formats.

6.4 Deployment implications

From a purely operational perspective, RoBERTa and DistilBERT offer viable trade-offs:

- **RoBERTa-Enhanced:** Best suited for security personnel, as it offers high interpretability and precision.
- **DistilBERT-Sentences:** Lightweight inference, leaving it suitable for real-time filtering/analysis when in an environment with hardware resource constraints.
- **GPT-Neo-Sentences:** Best recall, however, it requires post-filtering or threshold tuning to limit false-positives.

It would be advised that enterprise vendors combine these models with real-time perturbation testing and LIME-driven tools to enable maximum transparency.

6.4 Recommendations

- Integrate XAI overlays in email client, clearly visible and articulated to facilitate user understanding and trust.
- Periodically retrain models on the latest examples of phishing emails and domain-specific corpora.
- Use sentence-piece tokenizers to help mitigate threat of obfuscated and mutated phishing variants.
- Run analysis (e.g., McNemar's test) pre-deployment of updates to production models.

7. Conclusion and future work

This thesis evaluated the generalisability, robustness and interpretability of three transformer-based language models; DistilBERT, RoBERTa and GPT-Neo (125M) for the purpose of detecting malicious phishing emails that rely on manipulative social engineering tactics, often employed by threat actors globally.

The research addressed three research questions:

1. **Model efficacy:** The models RoBERTa-Enhanced and GPT-Neo-Sentences demonstrated the strongest generalisation capability when handling real-world phishing corpora, obtaining F1-scores of 0.558 and 0.689, respectively. All baseline models (notated by "Standard") trained on synthetic data alone exhibited overfitting and ultimately failed to generalise, thus confirming the limitations of prior work that ignored dataset realism.

2. **Explainability:** LIME, attention-based visualisations were successfully integrated across all models, with easily interpretable results for users. SHAP failed, as it relies solely on token-level interpretability and masking to attribute model predictions. SHAP's assumptions failed as the models tokenizer designs (WordPiece and byte-level BPE) fragmented input, leaving it incompatible with SHAP.

3. **Robustness:** When trialled against adversarial perturbations (text typos, synonyms and punctuation jitter), all models exhibited a notable drop in performance. RoBERTA-Enhanced and GPT-Neo-Sentences proved most resilient, further reinforcing the hypothesis that sentence-piece tokenisation results in greater resistance when handling linguistic manipulation.

Key Contributions:

1. A reproducible comparison of three different transformer families using a unified training and testing framework.
2. A robustness sandbox for phishing NLP models.
3. Evidence that sub-word tokenisation can improve generalisation in generative transformers (GPT-Neo).

Future Work:

1. **Multilingual Phishing:** Future work should expand model capabilities to better detect phishing emails when written in different languages, addressing the growing trend of localised social engineering tactics that can evade English-based detection methods.
2. **Federated Fine-Tuning:** Leveraging federated learning could assist organisations to fine-tune phishing detection models over decentralised datasets, whilst better preserving data privacy and compliance regarding business intellectual property/secrets.
3. **Adversarial Retraining Loops:** Using continuous adversarial retraining loops that could allow models to adapt in response to novel obfuscation tactics, this maintaining robustness with ever changing threat landscapes.
4. **Longitudinal XAI Studies:** Long-term research is required to better evaluate how XAI interfaces can influence analyst and end-user behaviour, with particular emphasis on reducing alert fatigue and proving interpretable phishing warnings.

References:

Verizon. "2025 Data Breach Investigations Report." Verizon Business, 2025, www.verizon.com/business/resources/reports/dbir/.

***APWG (2024)** Phishing Activity Trends Report Q1 2024. Anti-Phishing Working Group.*
https://apwg.org/trendsreports/

Holzinger, Andreas, et al. "Explainable AI Methods - a Brief Overview." XxAI - beyond Explainable AI, 2022, pp. 13–38, https://doi.org/10.1007/978-3-031-04083-2_2.

Moore, Tyler, and Richard Clayton. "Examining the Impact of Website Take-down on Phishing." Proceedings of the Anti-Phishing Working Groups 2nd Annual ECrime Researchers Summit on - ECrime '07, 2007, https://doi.org/10.1145/1299015.1299016.

Prakash, Pawan, et al. "PhishNet: Predictive Blacklisting to Detect Phishing Attacks." 2010 Proceedings IEEE INFOCOM, Mar. 2010, https://doi.org/10.1109/infcom.2010.5462216.

Khonji, Mahmoud, et al. "Phishing Detection: A Literature Survey." IEEE Communications Surveys & Tutorials, vol. 15, no. 4, 2013, pp. 2091–2121, https://doi.org/10.1109/surv.2013.032213.00009.

Yasin, Adwan, and Abdelmunem Abuhasan. "An Intelligent Classification Model for Phishing Email Detection." International Journal of Network Security & Its Applications, vol. 8, no. 4, 30 July 2016, pp. 55–72, https://doi.org/10.5121/ijnsa.2016.8405.

Lee, Younghoo, et al. "CATBERT: Context-Aware Tiny BERT for Detecting Social Engineering Emails." ArXiv:2010.03484 [Cs], 7 Oct. 2020, arxiv.org/abs/2010.03484.

Hussain, M., Hussain, I., Jameel, H., Jalil, Z., Javed, A. and Sadiq, M.A. (2023) 'An intelligent phishing detection model using ensemble CNN-LSTM', Expert Systems with Applications, 215, p.119360

Naqvi, Bilal, et al. "Mitigation Strategies against the Phishing Attacks: A Systematic Literature Review." Computers & Security, vol. 132, 1 July 2023, p. 103387, www.sciencedirect.com/science/article/pii/S0167404823002973, https://doi.org/10.1016/j.cose.2023.103387.

Salloum, Said, et al. "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques." IEEE Access, vol. 10, 2022, pp. 65703–65727, https://doi.org/10.1109/access.2022.3183083.

Vaswani, Ashish, et al. "Attention Is All You Need." ArXiv, 12 June 2017, arxiv.org/abs/1706.03762.

Jamal, Suhaima, et al. "An Improved Transformer-Based Model for Detecting Phishing, Spam, and Ham: A Large Language Model Approach." Research Square (Research Square), 1 Dec. 2023, https://doi.org/10.21203/rs.3.rs-3608294/v1.

Sanh, Victor, et al. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." ArXiv.org, 2019, arxiv.org/abs/1910.01108.

Slack, Dylan, et al. "Fooling LIME and SHAP." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 4 Feb. 2020, https://doi.org/10.1145/3375627.3375830.

Ribeiro, Marco Tulio, et al. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 13 Aug. 2016, pp. 1135–1144, arxiv.org/pdf/1602.04938.pdf, https://doi.org/10.1145/2939672.2939778.

Thapa, Chandra, et al. "Evaluation of Federated Learning in Phishing Email Detection." Sensors, vol. 23, no. 9, 27 Apr. 2023, p. 4346, https://doi.org/10.3390/s23094346.

Kairouz and McMahan. "Advances and Open Problems in Federated Learning." Foundations and Trends® in Machine Learning, vol. 14, no. 1, 2021, https://doi.org/10.1561/2200000083.

Jia, Robin, and Percy Liang. "Adversarial Examples for Evaluating Reading Comprehension Systems." ArXiv:1707.07328 [Cs], 23 July 2017, arxiv.org/abs/1707.07328.

Wiegreffe, Sarah, and Yuval Pinter. "Attention Is Not Not Explanation." ArXiv:1908.04626 [Cs], 5 Sept. 2019, arxiv.org/abs/1908.04626.

Egelman, Serge, et al. "You've Been Warned." Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08, 2008, https://doi.org/10.1145/1357054.1357219.

Teso, Stefano, et al. "Leveraging Explanations in Interactive Machine Learning: An Overview." Frontiers in Artificial Intelligence, vol. 6, 23 Feb. 2023, https://doi.org/10.3389/frai.2023.1066049.