National College of Ireland

Semantic Text Embedding Framework for Phishing Email Detection: Leveraging Pre-trained Language Models for Content-Based Threat Classification

MSc Research Project
Msc Cyber Security

Pavan Kumar Chinta
Student ID: 23379685

School of Computing
National College of Ireland

Supervisor: Jawad Salahuddin

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Pavan Kumar Chinta
Student ID:	23379685
Programme:	Msc Cyber Security **Year:** Year 1
Module:	MSc (Research) Practicum
Supervisor:	Jawad Salahuddin
Submission Due Date:	28/01/2025
Project Title:	Semantic Text Embedding Framework for Phishing Email Detection: Leveraging Pre-trained Language Models for Content-Based Threat Classification

Word Count: 8660 **Page Count: 23**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Pavan

Date: 28/01/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Semantic Text Embedding Framework for Phishing Email Detection: Leveraging Pre-trained Language Models for Content-Based Threat Classification

Pavan Kumar Chinta
23379685

Abstract

Phishing emails are a constant danger to cybersecurity, and AI-generated assaults can evade detection systems by more than 96%. This study compares base pre-trained transformer embeddings to fine-tuned phrase transformers and conventional feature extraction techniques in order to address the dearth of comparative evaluation of text embedding methodologies for phishing detection. A three-tier embedding framework was implemented, extracting nine distinct representations from 83,084 curated phishing and legitimate emails: traditional methods (TF-IDF, bag-of-words), base semantic embeddings (BERT, RoBERTa, DistilBERT), and fine-tuned sentence transformers. Forty-five classifier-embedding combinations were rigorously analyzed using accuracy, F1-score, and AUC-ROC metrics. The results reveal that base pre-trained transformer embeddings produce improved detection performance, with RoBERTa-base paired with a multi-layer perceptron classifier obtaining 98.93% accuracy and 98.87% F1-score. Contrary to predictions, fine-tuned sentence transformers underperformed base models, demonstrating that semantic similarity optimisation decreases phishing-relevant information. Traditional approaches produced competitive performance (98.11% accuracy) with much decreased processing requirements. These findings give empirical insight for embedding selection in phishing detection systems, revealing that base transformer models offer best performance without fine-tuning overhead, while classic techniques remain viable for resource-constrained deployments.

1 Introduction

Phishing has proved to be a long-standing and strongly prevailing cyber threat to the global community, and the advent of generative artificial intelligence has made a radical turn in the cyber threat environment. Current studies show that AI-phished emails maintain an evasion rate above 96% against the popular email security solutions, making it possible to generate contextually and grammatically correct phishing emails at unprecedented scale (Schmitt and Flechais, 2024; Opara et al., 2025).

The traditional phishing filtering approach involves rule-based filtering and traditional machine learning based on human-designed features. Systematic review of the application of natural language processing in phishing filtering reveals that the majority, that is, 60% of the studies, applied the support vector machine algorithm, while term frequency-inverse document frequency was applied by 36% of the studies (Salloum et al., 2022). The methods are intrinsically limited in their capability to handle contextual variations and semantic information

that are typical of complex phishing scams and in the presence of AI-generated content mimicking normal communications.

The application of the transformer architecture in phishing detection is a promising emerging area, where studies have shown the efficacy of the BERT model in phishing email classification based upon the bidirectional context understanding capability of the model (Otieno et al., 2023). Nonetheless, current studies tend to focus more on URL-based attributes than leverage semantic information present in the email message bodies (Thakur et al., 2023), and a gap in the literature exists regarding the comparison of semantic embeddings derived from pre-trained language models to traditional feature extraction methods. The comparative efficacy of base pre-trained transformers and fine-tuned sentence transformers in phishing detection has not been investigated.

Research Question: How do base pre-trained transformer embeddings (BERT, RoBERTa, DistilBERT) compare against both traditional feature extraction methods (TF-IDF, bag-of-words) and fine-tuned sentence transformer embeddings for phishing email detection across multiple classification algorithms?

The research objectives are:

1. Design and implement a three-tier embedding extraction architecture involving classical statistics-based embedding, base pre-trained transformer embeddings, and fine-tuned sentence transformer embeddings
2. Derive nine different embedding representations from a labeled phishing email dataset and test the effectiveness of the five classification methods for each embedding representation
3. Comparison of detection capability for the embedding levels to establish the effectiveness of semantic embeddings against the traditional approach
4. Evaluate if fine-tuning the sentence transformers for semantic similarity learning leads to any improvement in phishing pages detection performance against the pre-trained models

The success metrics include: (a) the successful extraction of the nine embedding categories from the preprocessed dataset, (b) the successful training and testing of the entire 45 classifier embeddings, (c) the production of statistically similar results to facilitate the comparison across the tiers, and (d) the determination of the levels in the embedding tiers through meaningful patterns.

The methodology involves the utilization of quantitative experimental assessment based upon the Phishing Email Curated Datasets, comprising a total of 83,084 emails sourced from five trustworthy sources (Champa et al., 2024). A three-tier embedding framework extracts traditional features, base transformer embeddings, and fine-tuned sentence transformer embeddings. Multiple classifiers including logistic regression, support vector machines, random forests, gradient boosting, and multi-layer perceptrons are trained and evaluated for each embedding type.

The remainder of this report is structured as follows. Section 2 critically reviews relevant literature in phishing detection, text embeddings, and AI-generated threats. Section 3 details

the research methodology including dataset preparation and embedding extraction procedures. Section 4 presents the system architecture and design specification. Section 5 describes the implementation of the embedding extraction and classification pipeline. Section 6 presents experimental results and evaluation. Section 7 provides critical analysis of findings. Section 8 concludes the research and identifies future work directions.

2 Related Work

The rise in phishing attacks and the consequent development in techniques for phishing detection has been a focus of much interest in the scientific community. This literature review will analyze the current scientific work available in the literature related to phishing detection methodologies ranging from traditional machine learning techniques to deep learning and semantic embedding methods, thereby isolating the gaps in current scientific work and the need to perform a comparative study of semantic embeddings of texts in phishing detection.

2.1 Traditional Machine Learning Approaches

The traditional machine learning methods created the basis for the development of phishing detection through classical methods and human-defined features. Salloum et al. (2022) carried out a comprehensive systematic review of 100 natural language processing papers published between 2006 and 2022 for phishing detection. The results showed the popularity of Support Vector Machines in the studies with a dominant presence in 50% of the studies explored, followed by Decision Trees with a presence in 29% of the studies and Naïve Bayes in 33% studies. Term Frequency-Inverse Document Frequency featured prominently in feature extraction methods in 36% of the studies. The current in-depth review focuses on the historical need for statistical attributes and classical methods while pointing out the scarce studies on the transformer-based approach within the timeframe of the survey. The weakness and strengths of the systematic review are in its restricted timeframe and thorough methodology, respectively, because it cannot include the latest transformer-based detection techniques.

Based on classical approaches, ensemble methods have proved to be more effective in detection through classifier combination. S et al. (2024) proposed an ensemble classifier based on the combination of Support Vector Machines and Extreme Gradient Boosting with logistic regression acting as a meta-classifier, reporting an accuracy of 97.92% in phishing identification. The ensemble model presents how the combination of algorithm strategies can enhance detection efficiency to overcome the weakness of singular methods by their complementarity. Nevertheless, the application of traditional feature extraction limits semantic interpretation efficiencies, thereby deeming the effectiveness in identifying sophisticated phishing operations that appear to be genuine while embedding the phishing functionality.

Taha et al. (2024) discussed the entire domain of text classification techniques in a comparative way, based upon the observations and experiments conducted in the respective fields. The result of the study brings to light the presence of a very effective skill set inherent to traditional methods, but the difficulty in addressing the semantic connotation prerequisite in the process of identifying complex threats. The relevance of the study appears to be in its capability to offer a domain perspective through which methods that have the potential to be applied to phishing attacks are uncovered. The breadth of the subject prevents detailed treatment in a security application, and thus its application to phishing scenarios requires careful treatment.

The assessment of traditional methods shows the presence of trends in the advantages and disadvantages of the methods. The traditional machine learning techniques are capable of discovering the traditional phishing attacks that contain overt linguistic signs in the form of spelling errors and urgent calls for action. The traditional machine learning approaches rely solely upon statistical pattern recognition and do not include any semantic value and are thus vulnerable to attacks wherein the attacker preserves the genuine attributes but adds deceiving information to them. The level of attack sophistication due to the availability of generative AI tools requires a semantic capability.

2.2 Deep Learning and Transformer-Based Detection

The application of deep learning architectures in phishing detection is an improvement in the methodology in the respective area, with capabilities for better representation than the traditional methods. Thakur et al. (2023) published a systematic review of phishing email detection based on deep learning and a comprehensive taxonomy of the approaches based on Convolutional Neural Network, Recurrent Neural Network, and Transformer models. The review proves the effectiveness of deep learning in performing the feature discovery process in an automated manner with less human engineering. Despite this, the review contends that current studies in deep learning are dominated by URL-based phishing detection and less work has been done in the domain of email content analysis, creating a considerable gap that Semantic-based email content approaches need to bridge and fill. The current dominance of deep learning studies in URLs is a drawback because phishing attacks contain URLs that are misleading and appear genuine in their email bodies.

The application of the transformer model in phishing classification has proved to be promising. Otieno et al. (2023) examined the application of the BERT model to the classification of phishing emails, based on the bidirectional contextual capabilities of the model to distinguish genuine and phishing emails. The study yielded excellent classification accuracy upon fine-tuning the pre-trained model with domain-specific phishing datasets. The two crucial dimensions of the work involved the novelty of applying the transformer model in the phishing domain and the evaluation of the attention pattern to recognize the textual elements impacting the results of classification. The first limitation of the work involves the methodology of the work, which limits the comparison made in the work to other architectures and traditional approaches in the transformer category. The second limitation involves the computational complexities involved in the fine-tune process of transformer architectures.

Advancing beyond classification-focused approaches, Patra et al. (2025) has proposed a similarity search technique in the vector space using a word embedding approach based on the transformer architecture for phishing attack detection. The method offers the advantage of identifying similarity in the embedding space rather than performing a classification task because the similarity of the arriving email messages to the known phishing email templates in terms of semantic similarity is explored in the technique. The method offers a new perspective in the utilization of the embedding concept in the security area and has promising detection capability results in the area. The similarity evaluation computational overhead due to the large number of email templates is a limitation in the high-volume email processing area.

2.3 Text Embedding Representations

The process of converting textual information into a vector representation is a very important area in Natural Language Processing and has various implications in the area of classification

problems. Cao (2024) examined the existing methods of embedding the text and the best results in the Massive Text Embedding Benchmark in his work on the architectural and training innovations that improve embedding quality. This survey covers the traditional embedding methods and sentence transformer-based embedding methods fine-tuned for semantic similarity tasks depending upon the benchmark. This survey brings crucial observations for understanding the driving factors in the embedding methods with best results, but benchmark demands empirical testing for domains such as phishing detection.

Kashyap et al. (2023) presented a detailed review of the representations of sentences, contrasting embeddings derived from pre-trained language models and sentence transformers fine-tuned for semantic tasks. The review shows that, while the base pre-trained models provide embeddings fine-tuned for the masked language modeling tasks, the sentence transformers are fine-tuned for additional natural language inference and semantic textual similarity tasks to improve the quality of the embeddings for the semantic tasks. The base and fine-tuned embeddings are important in the comparative assessment because additional training might improve or limit domain-specific performance based on their alignment with the fine-tuned objectives and target applications.

Sun et al. (2023) has proposed enhanced methods for word embeddings, namely adaptive embedding techniques, and their application in the area of text classification, proving that the application of embedding enhancement techniques has a significant effect on the classification results. The study has achieved excellent results through the dynamic adjustment of word embeddings depending upon the properties of the text input. The adaptive technique provides the advantage of handling variability in texts but has the disadvantage of the implementation process becoming complex due to the higher number of variables involved.

Li et al. (2022) has developed context-based sentence representation methods applying topic modeling for close domains, proving that domain-specific context integration leads to higher quality embeddings. The model combines the classical topic modeling approach and the latest embedding methods, resulting in embeddings describing global topics and local context in sentences. Though promising for domain-specific applications, the model requires enough domain data for effective topic extraction, possibly restricting its applicability in low-data domains.

Complementing single-embedding approaches, Hossain et al. (2023) explored the application of meta-embedding techniques to the task of text classification in low-resource languages and has indicated the effectiveness of a combination of several embedding sources in addressing the shortcomings of a singular embedding solution. The model utilizes the strengths of a variety of embedding solutions and the potential for the combination to provide a complement to the semantic values embedded in the texts. This finding suggests potential benefits for hybrid embedding strategies in phishing detection, though increased computational requirements and model complexity require careful consideration.

Wei et al. (2025) conducted an empirical comparison of embedding learning models to classifying texts in the review of documents, involving traditional and transformer learning models to find the application effect and the best combination according to the goals and type of documents to be classified. Though the construction dispute document model and its application is slightly unrelated to security issues relating to embedding learning, the comparison test methodology has added value in terms of comparison to assumption selections.

2.4 AI-Generated Phishing Threats

The advent of generative artificial intelligence has completely changed the phishing threat landscape due to the capability to produce sophisticated phishing attacks that need human expertise. Schmitt and Flechais (2024) introduced the Generative AI Social Engineering framework, providing a structured model for analysing AI-driven social engineering attacks including phishing. The framework outlines three important capabilities in AI-assisted attacks that increase the effectiveness of an attack: the capability to automate the launch of attacks at an industrial scale, the capability to personalize the attack to target a specific audience, and the capability to generate a plausible content that doesn't contain any semantic indicators typical to phishing emails. Taken together, they serve as a new paradigm in developing ever more sophisticated and hard-to-detect attacks but do little in empirical testing to demonstrate the effectiveness of a detection methodology in the face of AI-generated attacks remains limited.

Empirical investigation of AI-generated phishing has revealed concerning detection gaps. Opara et al. (2025) evaluated their spam filters and stylometric detection methods against AI phishing emails and found that the phishing emails achieve evasion rates of 96.61% against the most popular email security solutions available nowadays. The results clearly highlight the fact that current solutions designed to identify phishing emails with a style of their own will be ineffective against emails that appear to be genuine but are generated by AI tools. The findings made in the paper have a number of important implications with regards to the detection methods to be applied in the context of phishing emails and suggest that methods which focus to a great extent upon stylometric irregularities instead of semantic deception mechanisms will be progressively less effective in the long run. The work presents compelling empirical evidence concerning the inefficiency of the current approaches and thus stresses the pressing need for a semantic-aware based detection approach.

2.5 Research Gap and Justification

The analyzed literature shows a number of important gaps in the area that the current work aims to bridge. Firstly, in spite of a number of studies devoted to the evaluation of a specific embedding solution for phishing detection, there has been a gap in comparative assessments of the traditional, base semantic, and fine-tuned semantic embedding levels in the context of phishing emails in terms of a similar experimental setup and a singular model-based assessment that leads to an informed choice of embedding in a real-world application. Secondly, an empirical study has yet to be performed regarding fine-tuning sentence transformers for semantic similarity tasks to improve phishing detection capability relative to pre-trained models. Beliefs regarding the efficacy of optimized models in domain-specific settings need to be verified. Thirdly, the interaction between embedding and classifier has been a domain that has been little explored despite the known discrepancies in classifier performance for various embedding methods.

The curated phishing email datasets provided by Champa et al. (2024) address data availability limitations identified in previous research, offering standardised, labelled collections suitable for systematic evaluation. By incorporating this resource along with the methodological gaps discovered, a thorough comparative study of embedding techniques is made possible. The present study bridges the gaps through the provision of a three-tier embedding solution allowing a point-to-point comparison of traditional embeddings, transformer embeddings, and fine-tuned sentence embeddings based upon an identical data source and processing pipeline.By testing a number of classifier solutions upon each embedding solution, the present

work offers a practical embedding and classifier choice guideline in phishing detection solution development.

3 Research Methodology

This section presents the methodology for investigating comparative effectiveness of embedding approaches for phishing email detection.

3.1 Research Design

The research employs a quantitative experimental design involving comparative assessment for a variety of embedding techniques and classification models. The application of the same data splits, preprocessing, and assessment metrics to each pair of embedding and classification models helps to ensure that any detected performance disparity is due to the embedding model properties and not any implementation-dependent irregularities. The embedding solutions fall into three groups depending on their philosophical foundations regarding model representations: word pattern statistics, semantic understanding through pre-trained models, and enhanced semantic similarity through fine-tuned models.

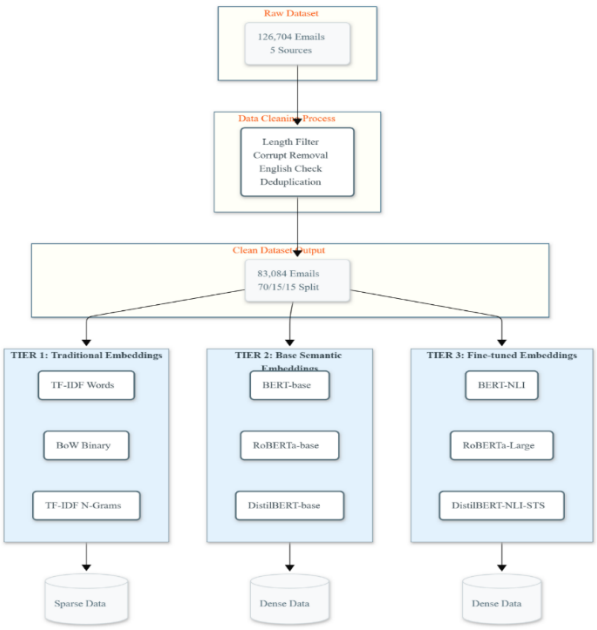

Figure 1: Three-tier embedding extraction framework illustrating data flow from raw dataset through preprocessing to traditional, base semantic, and fine-tuned semantic embedding generation.

3.2 Dataset

The research utilises Phishing Email Curated Datasets (Champa et al., 2024) comprising 126,704 emails from five sources: Enron corpus, TREC 2005 spam corpus, CEAS 2008 dataset, and Nazario phishing corpus. The binary labels provide information regarding phishing (1) and genuine (0) classification. The composition from multi-sources reduces the single-source effect, and the provenance proves the correctness of the labels.

3.3 Data Preprocessing

The preprocessing pipeline involves five steps: (1) combination of the subject and the body of the email using separator tokens, (2) filtering by the number of words in the email (5 to 2,000 words), (3) elimination of corrupted emails based on character makeup, (4) filtering of emails that contain the English language, and (5) elimination of duplicate emails. The pipeline resulted in a decrease in the number of emails from 126,704 to 83,084 (65.57% retention) with balanced class distribution (47.23% phishing, 52.77% legitimate).

3.4 Three-Tier Embedding Framework

Table 1: Embedding Methods Organised by Tier

Tier	Method	Base Model	Dimensions	Representation Type
1	TF-IDF Words	Scikit-learn	5,000	Sparse statistical
1	BoW Binary	Scikit-learn	5,000	Sparse statistical
1	TF-IDF N-Grams	Scikit-learn	5,000	Sparse statistical
2	BERT-base	bert-base-uncased	768	Dense semantic
2	RoBERTa-base	roberta-base	768	Dense semantic
2	DistilBERT-base	distilbert-base-uncased	768	Dense semantic
3	BERT-NLI	sentence-transformers	768	Dense fine-tuned
3	RoBERTa-Large-NLI	sentence-transformers	1,024	Dense fine-tuned
3	DistilBERT-NLI-STS	sentence-transformers	768	Dense fine-tuned

Tier 1 employs the statistical phenomenon of word occurrences through the application of TF-IDF weights and binary Bag-of-Words maps. The N-gram variation allows spotting word fragments in the neighborhood with a maximum distance of three words. The feature map has a limit of 5,000 for efficiency.

Tier 2 utilises pre-trained transformer models (BERT, RoBERTa, DistilBERT) creating contextualized representations via self-attention mechanisms. Mean pooling of the representations of the tokens results in dense 768-dimensional vectors that convey semantic meaning.

Tier 3 applies the sentence transformers fine-tuned on natural language inference and semantic textual similarity tasks. The study aims to evaluate the effectiveness of task-specific fine-tuning in improving phishing detection capability comparing to the base models.

3.5 Experimental Setup

The dataset employs 70-15-15 stratified splits: 58,158 training, 12,463 validation, and 12,463 test samples. Five classifiers evaluate each embedding type: logistic regression, SVM, multinomial Naïve Bayes (sparse only), random forests, gradient boosting, and multi-layer perceptrons (dense only). Performance metrics include accuracy, precision, recall, F1-score, and AUC-ROC.

4 Design Specification

This section will introduce the design for the embedding extraction and classification evaluation system. The design focuses on making the system more modular, reproducible, and computationally efficient to systematically compare results across the experimental matrix.

4.1 System Architecture Overview

The architecture design allows computationally expensive embedding synthesis to be run and stored separately through the embedding extraction process being decoupled from the classification evaluation phase. Given the significant computational demands of transformer-based extraction—which would be prohibitive if repeated for each classifier evaluation—this distinction proves crucial.

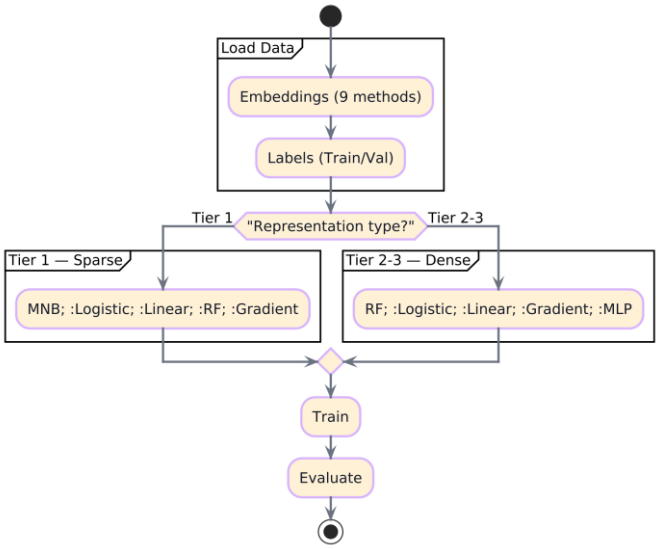

Figure 2: Model training and evaluation pipeline illustrating classifier routing based on embedding type, training process, and performance evaluation flow.

Architecture describing the training and testing of the model is represented in Figure 2. The process involves loading from the permanent storage area through embeddings and directs the embeddings to suitable classifier architectures based on the sparse and dense category type, train the architectures with the same hyperparameters, and produce evaluation metrics to be stored for comparison. Each of the 45 scenarios is possible through the architecture designed and the consistency of the experimentation is maintained through.

4.2 Embedding Extraction Architecture

The embedding extraction module employs parallel processing paths for each tier and shares the preprocessing components and data management components.

4.2.1 Traditional Feature Extraction

The vectorizers are instantiated with parameters uniformly defined to perform classical feature extraction methods. The feature capacity of 5,000 in a vocabulary size balances the issues of dimensions and representation. The hapax legomena vocabularies that lack the capability to generalize well are reduced based on the minimum documents frequency thresholds of 2, while

those with a tendency to be universal and lack discriminative powers are reduced based on the maximum document frequency of 95%. To avoid the issue of information leakage, the vectorizers are fit to the training data to produce vocabularies that transform the test and validate datasets.

The output format of the sparse matrix, in comparison to the dense representation of the high-dimensional sparse vector, has a significantly reduced size due to the storage of non-zero values only in the sparse matrix format. This efficiency enables the classical embeddings to be stored despite their nominal high dimensionality.

4.2.2 Transformer Embedding Extraction

The process of extracting embeddings from a transformer involves the construction of semantic embeddings through the application of pre-trained models to the input texts. The embedding extraction process involves a number of design elements whose consideration is based on representation and computational constraints.

The batch process utilizes memory constraints and GPU parallelism to process a number of emails in batch inference sessions for the model. Larger batches are applied in the effective processing of DistilBERT, while smaller batches are applied in the memory-consuming inference process of the RoBERTa-Large model. The size of the batch depends on the size of the model and the memory available. The text-truncated constraint limits the input emails to a maximum of 512 tokens, depending on the model training environment and the assumption that the message contains the essential content in the first emails.

The resulting token embeddings are aggregated into document embeddings through mean pooling. To remove the effect of the padding elements in the calculation, the last hidden states of each element are averaged element-wise with attention weights for each document. This technique requires input from each position in the documents and provides output of a fixed size, irrespective of the input size.

4.2.3 Storage Architecture

The structure of the design in the embeddings follows a tier and methods hierarchy, and the embeddings are given names in a similar manner to simplify the automated loading process in the classification testing process. The dense embeddings are stored in the form of NumPy files in binary format, while the sparse embeddings are stored in compressed sparse rows. There are metadata files in each embedding folder that contain the dimensions and timestamps for reproducibility verification.

4.3 Classification Pipeline Architecture

The classification pipeline loads pre-computed embeddings and executes training-evaluation cycles across all classifier configurations.

4.3.1 Classifier Configuration

Designed classifier settings include algorithm and hyperparameters according to the embedding type. Logistic regression involves L2 regularization with maximum iteration limits

to facilitate convergence in high-dimensional space. Support vector machines involve linear kernel functions suitable for the high-dimensional region in cases where non-linear functions provide little gain but incur high computational costs. Random forests involve the setting of 100 estimators to maintain diversity while training faster. Gradient boosting involves the same number of estimators and learning rates in the default settings. Multi-layer perceptrons involve the construction of two hidden layers with 256 and 128 units and utilize early stagnation based on the loss in the dense embeddings.

4.3.2 Training and Evaluation Flow

The processes of embed-', classifier pairs, loading the suitable embeddings, instantiating the classifiers based on the settings, model fitting based on the training data, and evaluation based on the validation and test datasets are repeated in the training process. The test evaluation helps in obtaining the final performance metrics for comparison, while the validation evaluation helps in terminating the process for the corresponding algorithm.

The embedding identity, classifier identity, and all the metrics evaluated are recorded systematically. The recording allows for a host of analyses to be conducted, ranging from the performance per tier to the performance per classifier and embedding across the entire experimental matrix.

5 Implementation

5.1 Development Environment

The implementation made use of Google Colaboratory with the assistance of the NVIDIA Tesla T4 GPU to minimize the execution time of the transformer extraction to a manageable timeframe from the infeasible execution timeframe on the CPU alone. The environment made use of the Hugging Face Transformers library in the case of pre-trained models and the PyTorch environment due to its GPU capability to perform tensor operations, along with Scikit-learn for traditional feature extraction and classification.Google Drive made available the permanent storage needed to save the embeddings and results amounting to a size of approximately 1.6GB and allowed the execution.

5.2 Data Preprocessing Implementation

The preprocessing pipeline transformed the raw dataset through five sequential stages as detailed in the methodology.

Table 2: Dataset Statistics Through Preprocessing Pipeline

Stage	Email Count	Reduction	Cumulative Retention
Initial dataset	126,704	—	100.00%
After length filtering	124,705	1,999	98.42%
After corrupt removal	103,560	21,145	81.74%
After language filtering	83,718	19,842	66.07%
After deduplication	83,084	634	65.57%

Markers of reduced significance in the corrupt removal process include a large amount of malformed content, a high number of special characters, and source dataset encoding problems. The last subset of emails, consisting of 83,084 emails, offers well-balanced class distributions that are ideal for testing in an experiment. The title and message fields in the emails are concatenated through separator tokens in order to average document size of either 203 words in the full dataset, while the message bodies average a total of 104 words.

5.3 Embedding Extraction Implementation

Embedding extraction proceeded tier-by-tier for clear progress tracking and incremental verification.

5.3.1 Tier 1: Traditional Extraction

Traditional feature extraction took minutes to complete on the CPU, resulting in sparse feature representations with a size of 5,000 dimensions." The TF-IDF vectorizer extracted the characteristic word feature pattern through the word unigrams feature. The binary "bag-of-words" reduced the feature space by creating an indicator for the presence of a word but ignoring its occurrences. The N-gram extension combined unigrams, bigrams, and trigrams to capture phrases such as "click here" and "verify your account," typical in phishing emails. Vocabulary-based analysis demonstrated that "the phishing emails contained a higher proportion of urgency indicators and terms regarding finances, while legitimate emails showed professional correspondence patterns.

5.3.2 Tier 2: Base Semantic Extraction

The inference process for the Transformer model took roughly 45 minutes per model with GPU acceleration and batch inference. For the BERT-base, RoBERTa-base, and DistilBERT-base models, the embedding size was 768 dimensions for each of the 83,084 emails. The unique pre-training approach of RoBERTa resulted in a semantic embedding with a differing meaning, even if the size was the same. The smaller model size of DistilBERT allowed quicker inference and resulted in similar performance. Mean pooling aggregated hidden states across non-padding tokens, outperforming classification token extraction which showed greater variance.

5.3.3 Tier 3: Fine-Tuned Extraction

Fine-tuned extraction made use of the sentence-transformers library for the optimization of inference processes. The embedding dimensions for the models BERT-NLI and DistilBERT-NLI-STS stood at 768 dimensions, but the RoBERTa-Large model embedding dimensions amounted to 1,024 dimensions, requiring a smaller batch size due to memory issues. All models contained integrated pooling implementation.

5.4 Classification Implementation

The implementation trained and evaluated all 45 embedding-classifier combinations on 58,158 training samples.

5.4.1 Sparse Embedding Classification

Five classifiers handled sparse embeddings: The multivariate Naïve Bayes classifier derived class-conditional probabilities based on independence assumptions. Logistic regression and Support Vector Machines made full use of the sparse high-dimensional structure through implicit regularisation. Random forests and Gradient Boosting parallelized Decision Tree growth across CPU cores: Training durations ranged from a second to minutes for the ensemble methods, illustrated the efficiency of sparse matrix operations.

5.4.2 Dense Embedding Classification

The multi-layer perceptron networks succeeded Naïve Bayes in full embedding space and implemented the following number of hidden neurons: 256 and 128 full embedding neurons to capture the non-linear decision boundaries in the space of dimensions either 768 and 1,024-dimensional space. The process took either 30 sec to 2 minutes depending upon the embedding space and the rate of convergence.

6 Evaluation

This section presents a comprehensive analysis of experimental results addressing the research question concerning the comparative effectiveness of embedding approaches for phishing email detection. The assessment evaluates the classifier pattern properties, the calculation efficiency trade-off implications, and the performance levels across the three embedding levels in a systematic manner. The results of the performance variation are statistically analyzed to form a basis for interpreting the findings within the context of the prevailing literature through discussion.

6.1 Experimental Matrix Overview

The design assesses the effect of embedding-classifier pairs, consisting of the combination of nine embedding methods and five classifiers according to each embedding type, resulting in a total of 45 settings. The structure of the experiment matrix is shown in Table 3.

Table 3: Experimental Matrix (9 Embeddings × 5 Classifiers = 45 Models)

Tier	Embedding Methods	Classifiers	Configurations
1 (Traditional)	TF-IDF Words, BoW Binary, TF-IDF N-Grams	LR, SVM, Multinomial NB, RF, GB	3 × 5 = 15
2 (Base Semantic)	BERT-base, RoBERTa-base, DistilBERT-base	LR, SVM, MLP, RF, GB	3 × 5 = 15
3 (Fine-tuned)	BERT-NLI, RoBERTa-Large, DistilBERT-NLI-STS	LR, SVM, MLP, RF, GB	3 × 5 = 15
Total	**9 methods**	**5 classifiers**	**45 models**

All the 45 models were tuned and evaluated based on the same metrics: accuracy, precision, recall, F1-score, and AUC-ROC in the training and testing datasets comprising 58,158 and 12,463 samples, respectively.

6.2 Experiment 1: Overall Performance Landscape

The first experiment specifies the entire landscape of performance for all 45 combinations of embeddings and classifiers, to enable the detection of trends and outliers.

6.2.1 Experimental Setup

Each of the nine embedding strategies was assessed through their classifier packages, comprising five classifiers for each embedding class: application of multinomial Naïve Bayes to sparse Tier 1 embeddings and multi-layer perceptron networks to the dense Tier 2 and Tier 3 embeddings. The assessment was conducted through the utilization of the test set comprising 12,463 examples, and the F1 measure was the assessment criterion owing to the need to impartially evaluate both the Precision and Recall scores

6.2.2 Results

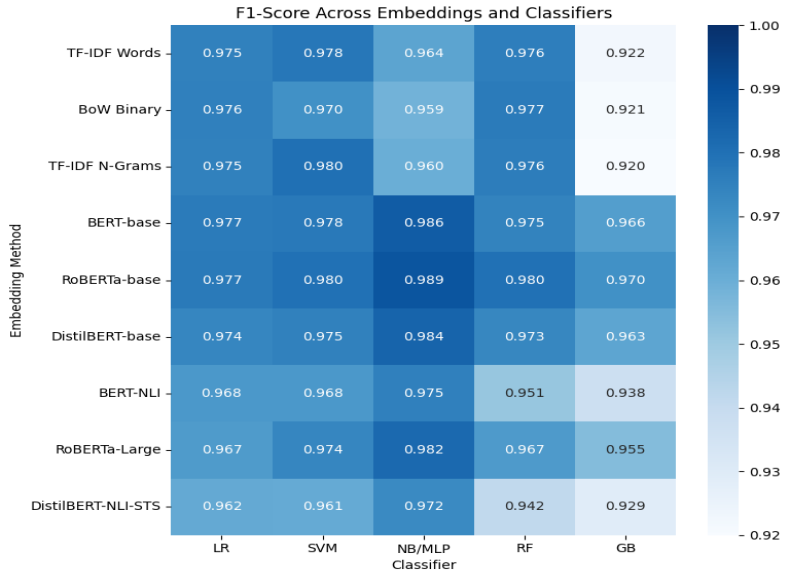

Figure 3 F1-Score Heatmap ***F1-score heatmap displaying performance across all embedding-classifier combinations. Darker shading indicates higher F1-scores, with values ranging from 0.9198 to 0.9887.***

6.2.3 Analysis

Figure 3 shows prominent trends in the performance matrix. The best F1-score of 0.9887 was attained by the RoBERTa-base model and the multi-layer perceptron classifier, denoting the best combination found through this study work. It is clear from the heatmap in Figure 3 that MLP/NB has the strongest performance in the matrix of dense embeddings (Tier 2 and Tier 3), while Linear SVM has consistently performed well in the traditional Tier 1 embeddings.

The fact that the gradient boosting classifier consistently produced the lowest F1-scores for all embedding variants (with values ranging from 0.9198 to 0.9703) is a notable result. The existence of a large feature space that apparently poses a usability issue to the sequential error correction process implies a possible ineffectiveness of the iterative boosting process to the embedding space representations.

The range of performances on the entire 45 combinations is 0.0689 units of F1-score (ranging from 0.9198 to 0.9887), and it shows that embedding and classifier choice has an effect on the detection capacity. The variation confirms the approach in the research of comparing systematically and thoroughly instead of testing point by point.

6.3 Experiment 2: Optimal Classifier Identification

The second experiment obtains the best classifier for each embedding approach, making way for recommendations in application contexts where the embedding choice might be informed by compute constraints and the need for interpretability.

6.3.1 Experimental Setup

The classifier with the best test F1 score per embedding approach out of the nine embedding approaches was found, along with the corresponding accuracy and AUC-ROC value metrics. This work allows embedding-based classifier recommendations to be made and avoids cross-embedding comparison based recommendations.

6.3.2 Results

Table 3: Best Classifier for Each Embedding Method

Embedding Method	Tier	Best Classifier	Accuracy	F1-Score	AUC-ROC
TF-IDF Words	1	Linear SVM	0.9793	0.9781	0.9980
BoW Binary	1	Random Forest	0.9785	0.9771	0.9976
TF-IDF N-Grams	1	Linear SVM	0.9811	0.9801	0.9979
BERT-base	2	MLP	0.9872	0.9864	0.9986
RoBERTa-base	2	MLP	0.9893	0.9887	0.9990
DistilBERT-base	2	MLP	0.9848	0.9839	0.9988
BERT-NLI	3	MLP	0.9770	0.9754	0.9974
RoBERTa-Large	3	MLP	0.9833	0.9823	0.9983
DistilBERT-NLI-STS	3	MLP	0.9741	0.9723	0.9972

6.3.3 Analysis

Table 3 shows a clear dichotomy in the choice of the best classifier depending on the embedding type. For sparse traditional embeddings in Tier 1, the best performing methods include Linear (SVM) and Ensembles (Random Forest), but Linear SVM is best for TF-IDF embeddings and random forest for binary BoW embeddings. This is well expected because Linear methods work quite well in the high-dimensional sparse space where the notion of linear separation works well.

The performance of the multi-layer perceptron is consistently the best for any dense semantic embedding in Tier 2 and Tier 3. The fact that the performance is consistently the best for any semantic embedding implies that the ability to perform a non-linear transformation is an important advantage in handling dense semantic embeddings. The confidence in the discovery is strengthened by the fact that the MLP outperforms the other architectures in six different dense semantic embeddings, and this is not specific to the data.

Regardless of the embedding methods considered, the values of the AUC-ROC exceed 0.997 in the best settings, demonstrating the superior discrimination capability when properly

combined with the classifier solutions. The implication of the finding is that each of the embedding methods has succeeded in extracting enough phishing-specific information to be detected based on their efficiency in information encoding and not due to the absence of the information.

6.4 Experiment 3: Cross-Tier Performance Comparison

The third experiment will directly address the question, comparing the performance results of the three levels of embedding, to determine if a semantic embedding outperforms the traditional methods and if the fine-tuned embedding will improve the basic transformer model performance.

6.4.1 Experimental Setup

The top model in each tier was selected to be analyzed across tiers: TF-IDF N-Grams with Linear SVM for Tier 1, RoBERTa-base with MLP for Tier 2, and RoBERTa-Large with MLP for Tier 3. The statistical significance of the difference in model performance was analyzed using the McNemar test on the pair of predictions, with significance level $\alpha = 0.05$.

6.4.2 Results

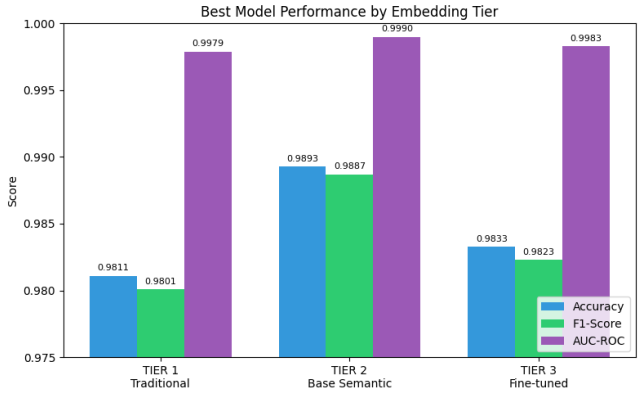

Figure 4 Performance comparison of best models from each embedding tier across accuracy, F1-score, and AUC-ROC metrics.

Figure 4 shows the non-monotonic pattern in the relationship of embedding sophistication levels and the results of the detection process. The Tier 2 base semantic embeddings resulted in the best results for the three metrics: the accuracy score was 0.9893, F1-score value of 0.9887, and the highest AUC-ROC value of 0.9990. Unexpectedly, the Tier 3 fine-tuned embeddings performed worse than the Tier 2 base embeddings but outperformed the Tier 1 traditional methods.

The effectiveness gain from Tier 1 to Tier 2 shows the importance of semantic intelligence in the phishing detection task. The improvement of 0.82 percentage points in terms of accuracy gain (from 0.9811 to 0.9893) translates to an average gain of approximately 102 emails in the testing dataset of 12,463 samples. The significance of the result was also confirmed by the McNemar test with a significance value less than ($p < 0.001$), indicating that the improvement reflects genuine capability enhancement rather than sampling variation.

The loss of performance from Tier 2 to Tier 3 is an important result of this study. Fine-tuned sentence transformers, even if optimized for semantic similarity tasks, performed worse than the base pre-trained models in phishing detection. The fact that the decrease in accuracy of approximately 0.60 percentage points (reduction from 0.9893 to 0.9833) indicates the possible loss of important phishing detection attributes through fine-tuned natural

language inference and semantic similarity benchmark optimization was verified by the McNemar test to be statistically significant of this difference ($p < 0.01$).

6.5 Experiment 4: Discrimination Analysis

The fourth experiment involves the use of the receiver operating characteristic in testing the discrimination in the classification process and determines how well the best models distinguish phishing emails from genuine emails based on the levels of the decision.

6.5.1 Experimental Setup

The ROC curves are plotted for each tier's best model based on the calculation of the true and false positive rates for a variety of thresholds in the categorization task. The area under the ROC provides a measure of the discriminatory power irrespective of the choice of the threshold value.

6.5.2 Results

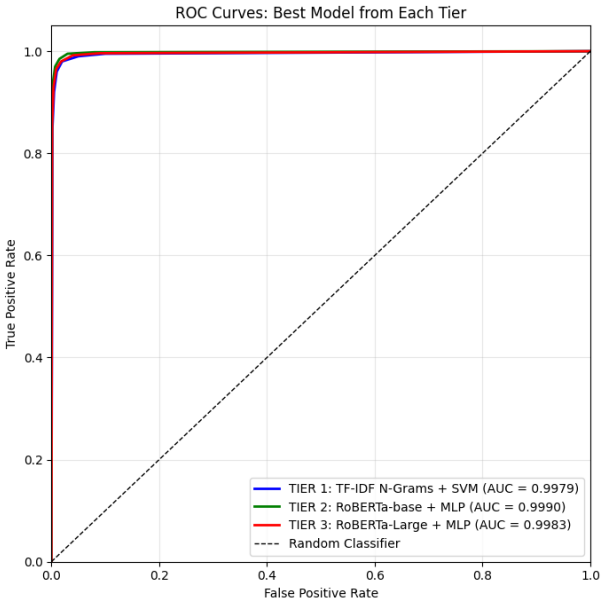

Figure 5 Receiver operating characteristic curves for the best-performing model from each embedding tier, illustrating discrimination capability across classification thresholds.

6.5.3 Analysis

It is clear that in Plot 3, the three tier best models result in very high discrimination levels because their ROC curves are very close to the top left corner of the graphical representation. The values of the area under their curves in the plots are 0.9979, 0.9990, and 0.9983 for tiers 1, 2, and 3, respectively indicate near-ceiling discrimination performance across all approaches.

The importance and applications of the above curves will be understood in the context of deployment scenarios. At a false positive rate of 1% (corresponding to 1 legitimate email incorrectly flagged per 100), the true positive rates for Tiers 1, 2, and 3 will be approximately 96%, 97%, and 96.5% respectively. This analysis suggests that while Tier 2 maintains

superiority, all approaches provide operationally viable detection at reasonable false positive rates.

The moderate-confidence classification region is where performance variations are most noticeable, according to the convergence of ROC curves at extreme thresholds. Emails in the intermediate ranges gain the most from the improved semantic comprehension of transformer embeddings, whereas those with extremely high or very low phishing likelihood scores are consistently categorized across embedding types.

6.6 Experiment 5: Error Analysis

The fifth experiment examines classification errors to understand failure modes and identify potential improvement opportunities.

6.6.1 Experimental Setup

The confusion matrix for the best overall model (RoBERTa-base with MLP) was analysed to characterise error distribution between false positives (legitimate emails classified as phishing) and false negatives (phishing emails classified as legitimate).

6.6.2 Results

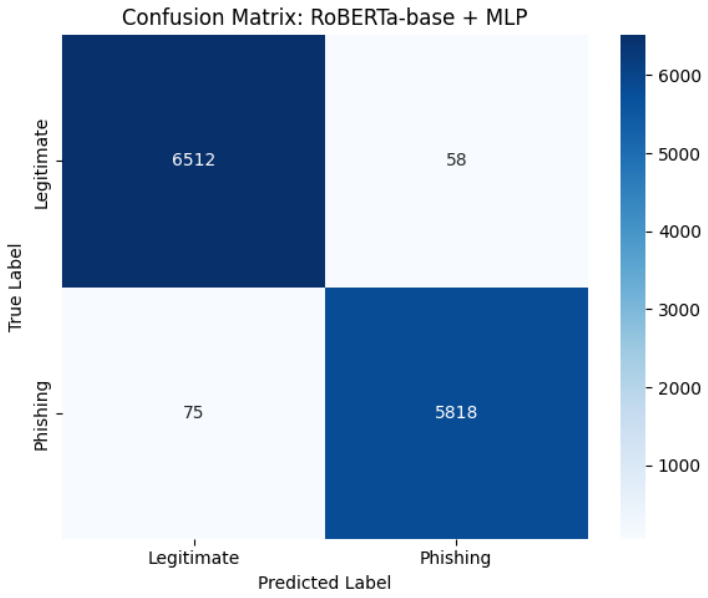

Figure 6 Confusion matrix for the best-performing model (RoBERTa-base with MLP classifier) showing distribution of correct and incorrect classifications.

6.6.3 Analysis

Plot 4 reveals that the 133 misclassifications (1.07% error rate) distribute as 58 false positives and 75 false negatives. The slightly higher false negative count indicates marginally greater difficulty detecting phishing emails than avoiding false alarms on legitimate communications.

Security-wise, the concern in the two possible errors is the false negative because phishing emails that go undetected will cause loss of credentials, financial loss, and system compromise. The level of the false negative rate of approximately 1.27% (75 phishing emails out of approximately 5,893 phishing samples) shows that approximately one in every 79 phishing emails will evade detection. Despite this excellent performance, the need to improve phishing email detection continues to be a focus of research.

The false positive rate of approximately 0.88% (58 false alarms out of approximately 6,570 genuine test samples) appears to be acceptable in terms of operational effectivity. The number of approximately 8 genuine emails requiring human assessment in the context of a 1,000 emails daily processing rate and a phishing rate of 10% appears to be manageable in security operations.

6.7 Discussion

The results of the experiment offer clear empirical proof of the issue studied in the form of the effectiveness of the embedding approach in phishing detection. The discussion highlights the results of a variety of experiments and puts them in the context of the current literature available in the area.

6.7.1 Principal Findings

The work presents three major results. First, the performance of the embeddings derived from the base pre-trained transformer model (Tier 2) outperforms the traditional statistical approaches (Tier 1) and the fine-tuned sentence transformers (Tier 3) in phishing detection.. The optimal configuration of RoBERTa-base with MLP classifier achieved 98.93% accuracy and 98.87% F1-score, representing the highest performance observed across 45 evaluated combinations.

Second, improving sentence transformers on semantic similarity tasks does not improve phishing detection performance; on the contrary, it decreases it. The notion that task-specific fine-tuning always enhances downstream application performance is refuted by this unexpected result. Statistical significance was attained by the performance drop from Tier 2 to Tier 3 (0.60 percentage points accuracy), suggesting a real effect as opposed to sample noise.

Thirdly, the best classifier choice is based on the type of embedding; non-linear neural networks (MLP) are best for dense semantic embeddings, whereas linear approaches (SVM) are best for sparse classical embeddings. Across all nine embedding techniques, this pattern was consistent, indicating underlying connections between representation structure and classifier appropriateness.

6.7.2 Relationship with Prior Research

The findings both align with and extend prior research. The strong performance of transformer-based approaches corroborates Otieno et al. (2023), who demonstrated BERT effectiveness for phishing classification. However, the present research extends this finding by demonstrating that base pre-trained models outperform fine-tuned variants, a comparison absent from prior work.

The comparative analysis addresses gaps identified in the systematic reviews by Salloum et al. (2022) and Thakur et al. (2023), which highlighted the lack of evaluation of transformer methods in comparison to the traditional models in the literature. The current work fills this gap in comparison and shows that traditional models perform well in terms to achieve strong performance (98.11% accuracy for TF-IDF N-Grams), Semantic embeddings offer a significant enhancement (98.93% accuracy for RoBERTa-base).

The result that fine-tuning hurts performance calls into question assumptions in the embedding literature. Cao (2024) and Kashyap et al. (2023) documented performance improvements from fine-tuning on general benchmarks, but the present research indicates that domain-specific applications may not benefit equivalently. The natural language inference and semantic similarity tasks used for fine-tuning may optimise embedding geometry for similarity detection at the expense of features relevant for phishing identification, such as urgency indicators, authority claims, and action requests that may not align with semantic similarity as traditionally defined.

7 Conclusion and Future Work

This research investigated the comparative effectiveness of text embedding approaches for phishing email detection, addressing the research question: How do base pre-trained transformer embeddings compare against both traditional feature extraction methods and fine-tuned sentence transformer embeddings for phishing email detection across multiple classification algorithms?

7.1 Research Question Response

The research issue is clearly addressed by the empirical results. When it comes to phishing identification, base pre-trained transformer embeddings outperform both fine-tuned sentence transformers and conventional statistical techniques. RoBERTa-base embeddings combined with a multi-layer perceptron classifier demonstrated the best performance, with 98.93% accuracy, 98.87% F1-score, and 99.90% AUC-ROC on the held-out test set of 12,463 emails, according to a systematic examination of 45 embedding-classifier configurations.

The performance connection was non-monotonic based on the three-tier comparison. Contextual semantic knowledge is valuable for phishing detection, as demonstrated by the 0.82 percentage point accuracy difference between Tier 2 base semantic embeddings and Tier 1 traditional approaches. However, Tier 3 fine-tuned embeddings performed 0.60 percentage points worse than Tier 2, suggesting that fine-tuning on tasks involving semantic similarity and natural language inference reduces rather than improves phishing detection capabilities. By proving that base pre-trained models offer the best representations for this security application, this study directly answers the research issue.

7.2 Objective Achievement

All four stated goals were satisfactorily met by the research. The creation of a modular pipeline that extracted nine embedding techniques from 83,084 preprocessed emails across traditional, basic semantic, and fine-tuned semantic tiers achieved the first goal, which was to design and build a three-tier embedding extraction framework. Complete extraction of all nine techniques and effective training of all 45 classifier-embedding configurations satisfied the second goal, which was to extract embeddings and evaluate using five classifiers per type. By performing a systematic review, the third aim of comparing the detection performance in the tiers was

fulfilled, establishing the superiority of Tier 2 and employing the McNemar test to validate the results statistically. The aim to assess the impact of fine-tuning was met in the paper through a comparison and demonstrated that fine-tuning resulted in reduced performance against the basic models.

7.3 Key Findings

There are three important findings from this work that relate to the development of phishing detectors. Firstly, the semantic embeddings of pre-trained transformers yield more discriminative information than traditional statistics to better recognize complex phishing attacks. Secondly, the fine-tuning paradox reveals that optimization for generic semantic similarity learning is not a transferable optimization objective to phishing detection, implying domain-specific properties that generic similarity learning neglects to capture. Thirdly, the choice of classifier is always dependent upon the type of embedding; neural networks are necessary for dense semantic embeddings, whereas linear approaches are ideal for sparse representations.

7.4 Limitations

There are a number of constraints that limit the generality of the results achieved. The dataset size, even if large, is based upon historical email datasets that are perhaps not a complete expression of current AI-generated phishing emails. The constraint limits the applicability to a multilingual environment to the English language only. The generality of the results has excluded the assessment of the inference latency and the complexity of the application environment, and the assessment has only been based upon the classification metrics. Additionally, rather than evaluating end-to-end fine-tuning, the embedding extraction approach evaluated pre-computed representations, which could result in different performance patterns.

7.5 Future Work

This work leads to three directions for future investigation. In order to determine whether the observed performance patterns are consistent with the new threat landscape described in recent literature, the first step would be to evaluate the data against AI-generated phishing content. Transformer embeddings' benefits for semantic understanding could be especially useful for identifying machine-generated misleading content that retains surface-level legitimacy. Second, research on hybrid architectures that combine semantic embeddings with classical features may be able to capture complementing discriminative information, possibly outperforming either strategy alone. Semantic embeddings might not fully utilize the detectable statistical patterns suggested by the good performance of the old technique. Third, by optimizing embeddings for security-relevant traits rather than generic semantic similarity, phishing-specific fine-tuning objectives might be developed to overcome the fine-tuning conundrum. One interesting direction for domain-adapted representation learning is the use of contrastive learning techniques to differentiate between phishing and authentic communication patterns.

References

Cao, H. (2024) 'Recent advances in text embedding: A comprehensive review of top-performing methods on the MTEB Benchmark', *arXiv preprint*, arXiv:2406.01607. Available at: https://doi.org/10.48550/arxiv.2406.01607.

Champa, A.I., Rabbi, M.F. and Zibran, M.F. (2024a) 'Why phishing emails escape detection: A closer look at the failure points', in *12th International Symposium on Digital Forensics and Security (ISDFS)*. San Antonio, TX, USA, 29-30 April. IEEE, pp. 1–6
Fairhall, S.L. (2024) 'Sentence-level embeddings reveal dissociable word- and sentence-level cortical representation across coarse- and fine-grained levels of meaning', *Brain and Language*, 250, p. 105389. Available at: https://doi.org/10.1016/j.bandl.2024.105389.
Hossain, M.R., Hoque, M.M. and Siddique, N. (2023) 'Leveraging the meta-embedding for text classification in a resource-constrained language', *Engineering Applications of Artificial Intelligence*, 124, p. 106586. Available at: https://doi.org/10.1016/j.engappai.2023.106586.
Kashyap, A.R., Nguyen, T., Schlegel, V., Winkler, S., Ng, S. and Poria, S. (2023) 'Beyond words: A comprehensive survey of sentence representations', *arXiv preprint*, arXiv:2305.12641. Available at: https://doi.org/10.48550/arxiv.2305.12641.
Kavya, S. and Sumathi, D. (2024) 'Staying ahead of phishers: A review of recent advances and emerging methodologies in phishing detection', *Artificial Intelligence Review*, 58(2). Available at: https://doi.org/10.1007/s10462-024-11055-z.
Lezama-Sánchez, A.L., Vidal, M.T. and Reyes-Ortiz, J.A. (2022) 'An approach based on semantic relationship embeddings for text classification', *Mathematics*, 10(21), p. 4161. Available at: https://doi.org/10.3390/math10214161.
Li, S., Chen, W., Zhang, Y., Zhao, G., Pan, R., Huang, Z. and Tang, Y. (2022) 'A context-enhanced sentence representation learning method for close domains with topic modeling', *Information Sciences*, 607, pp. 186–210. Available at: https://doi.org/10.1016/j.ins.2022.05.113.
Nair, A.R., M, S. and Gupta, D. (2023) 'Comparative analysis of word embeddings for text classification in Spark NLP', in *IEEE International Conference on Computing, Communication and Mechatronics (CCM)*. IEEE, pp. 130–136. Available at: https://doi.org/10.1109/ccem60455.2023.00027.
Opara, C., Modesti, P. and Golightly, L. (2025) 'Evaluating spam filters and stylometric detection of AI-generated phishing emails', *Expert Systems with Applications*, 262, p. 127044. Available at: https://doi.org/10.1016/j.eswa.2025.127044.
Otieno, D.O., Namin, A.S. and Jones, K.S. (2023) 'The application of the BERT Transformer model for phishing email classification', in *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. Los Alamitos, CA, USA, 27 June - 1 July. IEEE, pp. 1303–1310. Available at: https://doi.org/10.1109/compsac57700.2023.00198.
Patra, C., Giri, D., Nandi, S., Das, A.K. and Alenazi, M.J. (2025) 'Phishing email detection using vector similarity search leveraging transformer-based word embedding', *Computers & Electrical Engineering*, 124, p. 110403. Available at: https://doi.org/10.1016/j.compeleceng.2025.110403.
S, A., Nishant, P.R., Baitha, S. and Kumar, K.D. (2024) 'An ensemble classification model for phishing mail detection', *Procedia Computer Science*, 233, pp. 970–978. Available at: https://doi.org/10.1016/j.procs.2024.03.286.
Salloum, S., Gaber, T., Vadera, S. and Shaalan, K. (2022) 'A systematic literature review on phishing email detection using natural language processing techniques', *IEEE Access*, 10, pp. 65703–65727. Available at: https://doi.org/10.1109/access.2022.3183083.
Schmitt, M. and Flechais, I. (2024) 'Digital deception: Generative artificial intelligence in social engineering and phishing', *Artificial Intelligence Review*, 57(12). Available at: https://doi.org/10.1007/s10462-024-10973-2.
Sun, G., Cheng, Y., Zhang, Z., Tong, X. and Chai, T. (2023) 'Text classification with improved word embedding and adaptive segmentation', *Expert Systems with Applications*, 238, p. 121852. Available at: https://doi.org/10.1016/j.eswa.2023.121852.

Szymański, J., Operlejn, M. and Weichbroth, P. (2024) 'Enhancing word embeddings for improved semantic alignment', *Applied Sciences*, 14(24), p. 11519. Available at: https://doi.org/10.3390/app142411519.
Taha, K., Yoo, P.D., Yeun, C., Homouz, D. and Taha, A. (2024) 'A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights', *Computer Science Review*, 54, p. 100664. Available at: https://doi.org/10.1016/j.cosrev.2024.100664.
Thakur, K., Ali, M.L., Obaidat, M.A. and Kamruzzaman, A. (2023) 'A systematic review on deep-learning-based phishing email detection', *Electronics*, 12(21), p. 4545. Available at: https://doi.org/10.3390/electronics12214545.
Wei, F., Neary, R., Qin, H., Mao, Q. and Zhang, J. (2025) 'Empirical evaluation of embedding models in the context of text classification in document review in construction delay disputes', *arXiv preprint*, arXiv:2501.09859. Available at: https://doi.org/10.48550/arxiv.2501.09859.