Email Phishing Detection Report

Omar Al-Najjar, **Hamzah Drawsheh**, **Mohammad Othman**

December 13, 2025

Abstract

This report presents the development of a hybrid multi-model system for detecting phishing emails. Using the Enron email dataset, extensive preprocessing, feature engineering, TF-IDF vectorization, dimensionality reduction, clustering, graph-based analysis, and neural network classification were applied. The system addresses class imbalance and leverages advanced techniques for feature extraction and model optimization. Experimental results demonstrate high accuracy and strong effectiveness in distinguishing phishing from legitimate emails.

1 Introduction

The objective of this project is to develop a robust hybrid system for phishing email detection by combining textual analysis, engineered metadata features, unsupervised clustering, graph-based structural analysis, and supervised neural network classification.

2 Dataset Description

2.1 Data Source

The Enron email dataset (`https://www.cs.cmu.edu/enron/`) contains approximately 520,000 emails collected from employees of the Enron organization during the early 2000s.

2.2 Dataset Features

The following features were extracted from each email:

- Subject
- Body
- Date
- File path
- To
- From
- CC
- BCC
- URLs

3 Data Parsing and Preprocessing

3.1 Parsing

All raw email files were parsed to extract structured header fields and textual content.

3.2 Preprocessing

- No duplicated records were found in the dataset.
- Missing values in the `subject` and `to` fields were replaced with empty strings.

4 Feature Engineering

4.1 Textual and Metadata Features

- Cyclical encoding of temporal features (hour, day, weekday, month).
- Extraction of structural metadata such as CC/BCC presence and URL counts.
- Rule-based heuristics for preliminary phishing risk labeling.

5 TF-IDF Vectorization

TF-IDF vectorization was applied to both subject lines and email bodies:

- **Subject:** minimum document frequency = 100, maximum document frequency = 50%, 100 features.
- **Body:** same thresholds, 300 features.

Total TF-IDF features: 400

6 Dimensionality Reduction

6.1 Truncated SVD

Truncated Singular Value Decomposition (SVD) reduced the TF-IDF feature space from 400 dimensions to 100 components while preserving semantic information.

7 Feature Scaling

7.1 Standard Scaler

Standard scaling was applied to all numerical features to ensure consistent magnitudes for clustering and neural network training.

8 Clustering

8.1 K-Means Clustering

MiniBatch K-Means clustering was applied to the reduced feature space. The optimal number of clusters was selected using the elbow method and silhouette analysis.

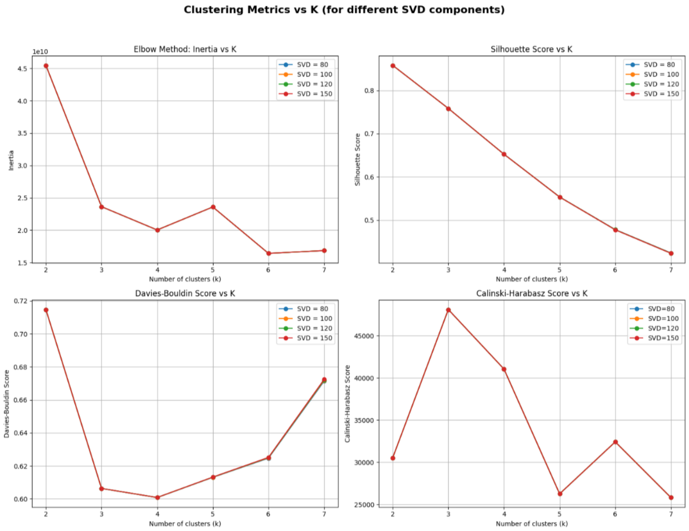

Figure 1: K-Means clustering evaluation using elbow and silhouette methods.

8.2 Risk Level Assignment

Clusters were analyzed to determine phishing risk levels.

Table 1: Cluster-Based Phishing Risk Levels

Cluster	Total Emails	Phishing Emails	Phishing Rate
0	501,021	13,411	0.0268
1	1,231	981	0.7969
2	15,097	7,023	0.4652

9 Sender and Domain Network Analysis

9.1 Top 15 Senders and Domains

A network-based analysis was conducted using the top 15 external senders and domains most frequently associated with phishing emails.

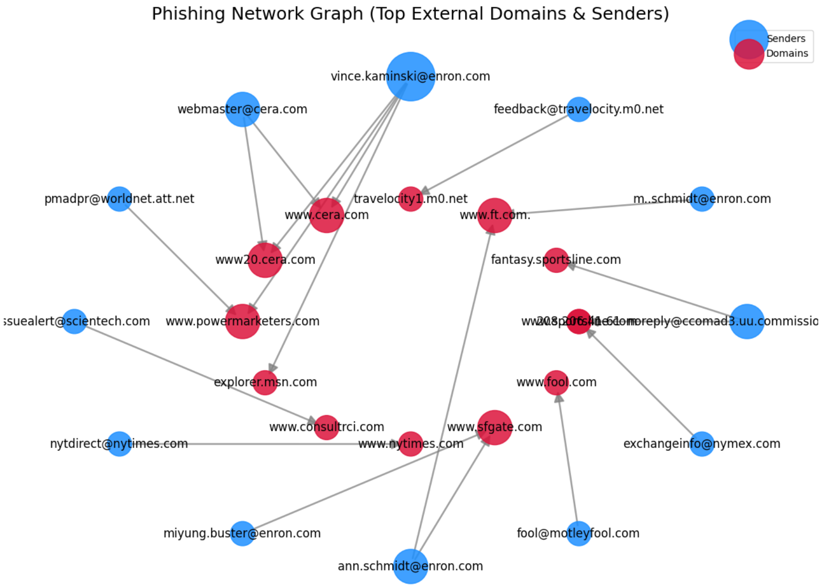

Figure 2: Network graph of top senders and domains associated with phishing emails.

10 Neural Network Model

10.1 Architecture

The neural network receives a combined feature vector consisting of reduced TF-IDF features, engineered metadata, cluster identifiers, and network-based indicators.

Table 2: Neural Network Architecture

Layer	Configuration
Input	Dimension = `input_dim`
Hidden Layer 1	256 neurons, BatchNorm, ReLU, Dropout(0.4)
Hidden Layer 2	128 neurons, BatchNorm, ReLU, Dropout(0.3)
Hidden Layer 3	64 neurons, BatchNorm, ReLU, Dropout(0.3)
Hidden Layer 4	32 neurons, BatchNorm, ReLU, Dropout(0.2)
Output	1 neuron, Sigmoid activation

10.2 Class Imbalance Handling

Random undersampling reduced class imbalance from approximately 98.5:1.5 to nearly 2:1.

10.3 Training Process

- Binary Cross-Entropy loss
- Adam optimizer with L2 regularization
- 30 epochs with early stopping

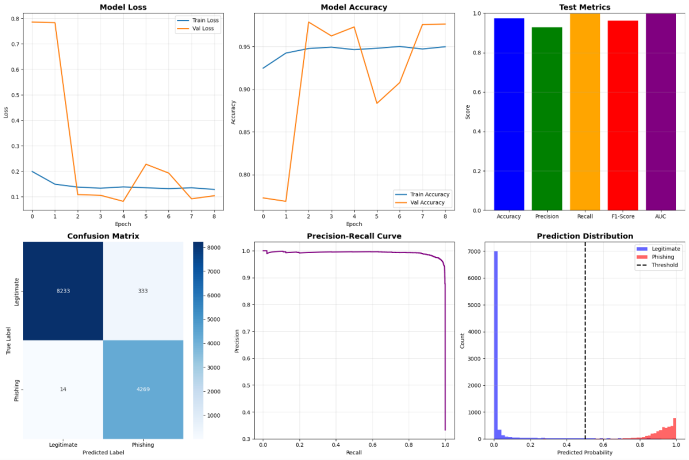

Figure 3: Neural network training performance.

10.4 Evaluation Metrics

Table 3: Test Set Performance Metrics

Metric	Value
Accuracy	0.9730
Precision	0.9276
Recall	0.9967
F1-Score	0.9609
AUC-ROC	0.9973

Table 4: Confusion Matrix

	Predicted Legitimate	Predicted Phishing
Actual Legitimate	8233	333
Actual Phishing	14	4269

11 Conclusion

The proposed hybrid phishing detection system achieves high accuracy, with estimated real-world performance ranging between 90% and 95%, demonstrating its practical applicability in operational email environments. The integration of unsupervised clustering and sender–domain network analysis provides an additional layer of contextual understanding that complements traditional text-based classification. Clustering enables the identification of latent behavioral patterns and risk groupings among emails, while network analysis highlights structural relationships between malicious senders and frequently abused domains.

This multi-perspective approach improves the robustness of the system against evolving phishing strategies, as it does not rely solely on lexical features that can easily be obfuscated by attackers. Furthermore, the inclusion of interpretable components—such as cluster-based risk levels and network connectivity indicators—enhances transparency and supports more informed decision-making by security analysts. Overall, the hybrid architecture balances predictive performance with interpretability, making it well-suited for real-world deployment and future extension to adaptive and continuously learning security systems.