INDIRA GANDHI NAITONAL OPEN UNIVERSITY

MSEP - 028

PhishGuard: AI and Deep Learning-Based Phishing Email Detection System

By

Student's Full Name: Lalita Dev

Enrolment No.: 2400171188

Submitted to the School of Vocational Education and Training, IGNOU in partial fulfilment of the requirements for the award of the degree

Master of Science (Information Security)

(MSCIS)

Year of Submission – 2025

Indira Gandhi National Open University

Maidan Garhi

New Delhi – 110068

SCHOOL OF VOCATIONAL EDUCATION AND TRAINING, IGNOU, MAIDAN GARHI, NEW DELHI – 110 068

II.PROFORMA FOR THE APPROVAL OF MSCIS/PGDIS PROJECT PROPOSAL

(Note: All entries of the proforma of approval should be filled up with appropriate and complete information. Incomplete proforma of approval in any respect will be summarily rejected.)

Project Proposal No :……..
(for office use only)

Course Code:…MSEP028……..
Enrolment No.: ……2400171188……
Study Centre: .38046 - RAJDHANI COLLEGE
Regional Centre:DELHI **RC Code:**3….
E-mail: ..devlata1998@gmail.com……...
Mobile/Tel No.: 7827243179……..

1. Name and Address of the Student — Lalita Dev
162/1 , Qutub vihar phase 2, hanuman chowk new delhi 110071

2. Title of the Project — PhishGuard: AI and Deep Learning-Based Phishing Email Detection System

lalita dev
Signature of the Student
Date: .08-.09-.2025….

For Office Use Only

☐ Approved ☐ Not Approved

……
Signature, Designation, Stamp of the Project Proposal Evaluator
Date: ……

Suggestions for reformulating the Project:

Ensure that you include the following while submitting the Project Proposal:

1. **Proforma for Approval of Project Proposal duly filled and signed by the student with date.**
2. **Project proposal (12-15 pages).**

Note:

i. *Violation of the project guidelines will lead to the rejection of the project at any stage.*

A photocopy of the complete Project Proposal (along with Project Proforma, Project Proposal) submitted, should be retained by the student for future reference.

Table of Contents

PhishGuard: An AI and Deep Learning-Based Phishing Email Detection System

1.0 Title

PhishGuard uses a BiLSTM and a fine-tuned DistilBERT transformer to classify emails as phishing or legitimate, balancing contextual accuracy with efficient inference and robust operational scalability considerations for real-world use.

2.0 Introduction and Objectives

2.1 Introduction

Email is the primary medium for personal and professional correspondence and, correspondingly, a major vector for phishing attacks. Such deceptive messages compel recipients to disclose sensitive information, causing financial loss, identity theft, and large-scale breaches. Conventional rule-based and signature-based filters are increasingly inadequate against sophisticated tactics—content obfuscation, look-alike domains, and targeted social-engineering—that generalize poorly beyond known patterns.

Leveraging advances in artificial intelligence and deep learning can markedly improve detection by capturing semantic and contextual cues that legacy systems miss. The PhishGuard project will employ a bidirectional LSTM with pre-trained word embeddings and a fine-tuned DistilBERT transformer to classify emails as phishing or legitimate, with evaluation emphasizing both predictive performance and operational considerations (inference latency and minimization of false negatives) to ensure practical deployability.

2.2 Problem Statement

Phishing remains a pervasive cyber threat that exploits human trust to illicitly obtain sensitive information — including credentials, financial data, and personally identifiable information. Traditional signature- and rule-based filters are increasingly inadequate because adversaries continually adapt their tactics (e.g., content obfuscation, look-alike domains, and sophisticated social-engineering). These evolving techniques increase the likelihood of financial loss, identity theft, and organizational compromise. Consequently, there is a need for an automated, content-aware detection solution that leverages modern AI and deep learning methods to identify phishing emails with improved accuracy and robustness.

2.3 Objectives

The objectives of this project are:

1. To design and implement an AI-based phishing email detection system using BiLSTM and DistilBERT models.
2. To preprocess and balance phishing and legitimate email datasets for reliable training and evaluation.
3. To establish baseline models and compare them with BiLSTM and DistilBERT using metrics such as Accuracy, Precision, Recall, F1-score, ROC-AUC, and PR-AUC.
4. To optimize decision thresholds to reduce false negatives while maintaining an acceptable false positive rate.
5. To develop a user-friendly web-based prototype that demonstrates real-time phishing detection.
6. To ensure reproducibility through documented code, model artifacts, and evaluation results.
7. To explore the feasibility of integrating the system into enterprise email security frameworks and outline future enhancements.

3.0 Project Category

This project is classified as an Application Development Project in Cybersecurity and Information Security. It involves developing *PhishGuard*, an AI-based system using NLP and Deep Learning (BiLSTM and DistilBERT) for phishing email detection. The classification is justified as the work not only applies advanced models but also delivers a practical, demonstrable prototype with real-world security relevance.

4.0 Research Methodology

4.1 Research Approach

The research follows an experimental and comparative approach. Two deep learning models will be developed, trained, and evaluated under controlled conditions:

- BiLSTM with Word Embeddings (Word2Vec/GloVe)
- Transformer-based DistilBERT

The models will be compared using standard evaluation metrics, with emphasis on both detection accuracy and practical applicability through a prototype web-based application.

4.2 Data Collection and Preprocessing

Data Source: The project will utilize the *Curated Phishing Email Dataset* available on Figshare:

🔗 https://figshare.com/articles/dataset/Curated_Dataset_-_Phishing_Email/24899952

Preprocessing Steps

- Cleaning: Removal of HTML tags, signatures, headers, and duplicate entries.
- Normalization: Lowercasing, stop-word removal, and tokenization.
- Replacement: Masking sensitive tokens such as URLs, email addresses, and numbers with placeholders.
- Handling Class Imbalance: Application of SMOTE (Synthetic Minority Oversampling Technique) or class weighting.
- Splitting: Stratified division into 70% training, 15% validation, and 15% testing sets.

4.3 Model Development

BiLSTM Model

- Input: Pre-trained Word2Vec/GloVe embeddings.
- Architecture: Bidirectional LSTM layers with dropout regularization.
- Output: Sigmoid classifier for binary classification (phishing vs. legitimate).

DistilBERT Model

- Input: Tokenized email text sequences (up to a maximum length).
- Fine-tuning: Adaptation of pre-trained DistilBERT weights for phishing classification.
- Advantages: Contextualized word representation with reduced size and faster inference compared to full BERT.

4.4 Training Strategy

- Optimizer: Adam with tuned learning rates.
- Regularization: Dropout, batch normalization, and early stopping.
- Threshold Calibration: Dynamic thresholding based on Precision–Recall curve analysis to minimize false negatives.
- **Evaluation Metrics**
  - Accuracy
  - Precision, Recall, F1-score
  - ROC-AUC, PR-AUC
  - Confusion Matrix

4.5 Tools, Platforms, and Technologies

Category	Libraries / Frameworks Used
Programming Language	Python 3.11
Web Frameworks	Flask , flask-cors , FastAPI
Deep Learning Frameworks	PyTorch (2.5.1+cu121), Torchvision, Torchaudio
Transformer Models	Hugging Face Transformers , Tokenizers, accelerate
NLP Tools	NLTK (3.9.1), Gensim (4.3.2)
ML Libraries	Scikit-learn, SciPy, Pandas, NumPy
Visualization	Matplotlib, Seaborn
Database	MySQL (mysql-connector-python 8.3.0)
Development Tools	Jupyter Notebook, VS Code, Git/GitHub

4.6 System Requirements

Component	Specification
Processor	Intel i5 or higher (multi-core recommended)
RAM	Minimum 8 GB (16 GB recommended)
Storage	20 GB free disk space (datasets + models)
GPU (Optional)	NVIDIA CUDA-enabled GPU (CUDA 12.1 supported by Torch 2.5.1)

4.7 Software Requirements

Software	Details
Operating System	Windows 10/11, Ubuntu Linux, or macOS
Python Version	3.8 – 3.11 (current: 3.11)
Database	MySQL Server & Workbench
Browser	Latest Chrome / Firefox / Edge

4.8 Validation and Testing

- Cross-validation: K-fold cross-validation to ensure model stability.
- Error Analysis: Examination of misclassified emails, including obfuscated URLs and multilingual content.
- Reproducibility: Fixed random seeds, configuration files, and experiment logs to ensure consistent results.

4.9 Expected Outcomes

The methodology is expected to yield:

1. Two trained phishing detection models (BiLSTM and DistilBERT).
2. A comparative analysis highlighting their respective strengths and limitations.
3. A prototype system capable of real-time phishing email detection and classification.

5.0 Scope of the Solution

PhishGuard is conceptualized as an AI-driven phishing email detection system with both academic significance and practical applicability. Its scope is defined as follows:

5.1 Functional Scope

- Phishing Email Detection: Automatic classification of emails as phishing or legitimate using deep learning models.
- Comparative Model Analysis: Empirical evaluation and performance comparison of BiLSTM and DistilBERT architectures.
- Prototype Deployment: Development of a functional, web-based demonstration system accessible to end-users.

5.2 Practical Applications

- Enterprise Security: Potential integration into organizational email gateways to strengthen defenses against phishing.
- Personal Security: Usability for individual users as a lightweight email filtering tool.
- Education and Training: A reference framework for demonstrating AI and NLP applications in cybersecurity.

5.3 Benefits

- Improved security against evolving phishing threats.
- Lower false negative rates compared to conventional rule-based filters.
- Scalable architecture adaptable to real-world deployment scenarios.

6.0 Analysis

- **Limitations of Existing Solutions**
  - Rule-based filters and blacklists are reactive and ineffective against zero-day attacks.
  - Traditional ML models (e.g., SVM, Naïve Bayes) fail to capture semantic and contextual meaning in email text.
- **Proposed Approach**
  - BiLSTM with Word2Vec/GloVe embeddings to capture sequential and structural patterns.
  - Fine-tuned DistilBERT transformer to extract deep contextual semantics with efficiency.
  - Comparative evaluation using Accuracy, Precision, Recall, F1-score, ROC-AUC, and PR-AUC.
- **Expected Insights**
  - BiLSTM: Strong at modeling sequential structures; lightweight but less effective in ambiguous contexts.
  - DistilBERT: Superior contextual understanding and real-world robustness; slightly higher computational cost.
  - Comparative Study: Highlights trade-offs between efficiency (BiLSTM) and accuracy (DistilBERT), offering deployment guidance for diverse environments.
- **Relevance**
  - Demonstrates how deep learning models can overcome limitations of legacy systems.
  - Provides a framework adaptable to enterprise security, personal protection, & educational use.

7.0 Future Scope and Enhancements

While PhishGuard establishes a strong foundation for phishing detection, several enhancements can extend its capabilities and real-world applicability:

- **Multi-lingual Support:** Extend detection beyond English to include regional and international languages, addressing phishing as a global challenge.
- **Image-based Phishing Detection:** Integrate computer vision methods to analyze email attachments, fake brand logos, and embedded phishing content.
- **Advanced Model Architectures:** Explore ensemble learning (e.g., BiLSTM + DistilBERT + CNN) to achieve higher accuracy and robustness.
- **Browser and Email Client Integration:** Develop lightweight extensions for browsers (Chrome, Firefox) or integrate with email clients (Outlook, Thunderbird, Gmail API).
- **Real-time Threat Intelligence:** Incorporate external threat feeds, WHOIS data, and dynamic blacklists to detect zero-day phishing campaigns.
- **Adaptive Learning:** Enable incremental retraining with newly observed phishing samples, ensuring resilience against evolving attacker tactics.

8.0 conclusion of project

PhishGuard aims to enhance email security by leveraging BiLSTM and DistilBERT models for intelligent phishing detection. The project integrates comparative analysis, prototype deployment, and practical evaluation, providing a scalable, robust, and adaptable framework for both academic research and real-world cybersecurity applications.

9.0 Bibliography and Literature Survey

9.1 Literature Survey

- **Traditional Filters** – Rule-based systems and blacklists (e.g., SpamAssassin) provided limited protection but were ineffective against novel or obfuscated phishing attacks.
- **Machine Learning** – Models such as Naïve Bayes, Random Forest, and SVM improved over rules but lacked contextual understanding of email content.
- **Deep Learning** – BiLSTM with word embeddings captured sequential patterns better than classical ML but struggled with nuanced semantics.
- **Transformers** – BERT-based models, especially DistilBERT, advanced phishing detection by providing strong contextual representation while being computationally efficient for real-time use.

9.2 References

- Vaswani, A. et al. (2017). *Attention Is All You Need*. NeurIPS.
- Devlin, J. et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL.
- Sanh, V. et al. (2019). *DistilBERT: Smaller, Faster, Cheaper and Lighter*. arXiv.
- Rao, J., & Ali, S. (2021). *Phishing Email Detection Using NLP and Deep Learning*. IEEE.
- Curated Phishing Email Dataset. Figshare (2024). Link
- OpenAI Documentation.
- Official Documentation of PyTorch, Scikit-learn, MySQL, Flask.
- IGNOU MCA Project Guidelines.
- ChatGPT-assisted Research Notes.