JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AI-Driven Phishing Detection:A Machine Learning Approach Using NLP.

Prof.Sheetal Shimpikar
Department of Computer Engineering
Pillai College of Engineering
Navi Mumbai, India
sheetalshimpikar@mes.ac.in

Suraj Alam
Department of Information Technology
Pillai College of Engineering
Navi Mumbai, India
suraja22it@student.mes.ac.in

Aman Jankar
Department of Information Technology
Pillai College of Engineering
Navi Mumbai, India
ajankar22it@student.mes.ac.in

Prathamesh Golhar
Department of Information Technology
Pillai College of Engineering
Navi Mumbai, India
pgolhar22@student.mes.ac.in

Hemant Koli
Department of Information Technology
Pillai College of Engineering
Navi Mumbai
hkoli22it@student.mes.ac.in

***Abstract*—The AI-powered phishing detection system leverages Natural Language Processing (NLP) and Machine Learning (ML) to identify and mitigate phishing threats in emails and websites. The system takes raw email content or website URLs as input and processes them using advanced techniques such as tokenization, stopword removal, and TF-IDF vectorization. A Random Forest Classifier and LSTM-based deep learning model are employed to classify inputs as phishing or legitimate. The input data is transformed into numerical features, which are fed into the trained model for real-time prediction. The output is a binary decision (phishing or non-phishing),delivered through a Flask-based REST API. The system is implemented as a browser extension and email plugin, enabling seamless integration into user workflows. The model is trained on a diverse dataset of phishing and legitimate samples, achieving high accuracy and robustness. This project highlights the application of AI in cybersecurity, offering an effective solution to combat phishing attacks and enhance user safety.**

Keywords*— *Artificial Intelligence, Machine Learning, Phishing Detection, Cybersecurity, Natural Language Processing, Email Classification, Feature Extraction, Data Preprocessing, Supervised Learning, Anomaly Detection, Neural Networks, Model Training, URL Analysis, Spam Filtering, Threat Intelligence, Pattern Recognition, Real-time Detection.

I. INTRODUCTION

Phishing remains one of the most prevalent and damaging threats in the cybersecurity landscape, exploiting human vulnerability through deceptive communication tactics. As attackers continue to evolve their methods, traditional detection mechanisms struggle to keep pace. In response, researchers have increasingly turned to artificial intelligence (AI), particularly machine learning (ML) and natural language processing (NLP), to enhance phishing detection capabilities. This survey paper presents a comprehensive overview of AI-driven approaches to phishing detection, with a focus on NLP-based techniques that analyze textual features in emails, websites, and messages. By examining the state-of-the-art in machine learning models, feature extraction methods, and real-world applications, this paper highlights current advancements, identifies existing challenges, and outlines future directions in the field. The goal is to provide a structured understanding of how AI and NLP can be effectively harnessed to build adaptive, intelligent systems capable of mitigating phishing threats in dynamic and complex environments.

A. Fundamentals

Phishing is a type of cyberattack in which attackers attempt to trick individuals into providing sensitive information such as usernames, passwords, banking details, or personal identification. These attacks are commonly carried out via deceptive emails, websites, or messages that appear to come from trusted sources. With the increase in digital communication, phishing has become one of the most dangerous and widespread threats to cybersecurity today. Phishing emails and websites often mimic legitimate entities, making them hard to detect, especially for non-technical users. Traditional rule-based filtering systems (e.g., blacklist

URLs, keyword matching) often fail to detect newer or more sophisticated phishing attempts. Similarly, signature-based approaches cannot effectively identify previously unseen (zero-day) phishing threats. Due to the evolving nature of phishing techniques, a dynamic, intelligent, and adaptive detection system is required. This project proposes an AI-driven phishing detection system that uses Natural Language Processing (NLP) and Machine Learning (ML) techniques to identify and flag phishing content from emails or websites.

B. Objectives

The project aims to achieve several crucial objectives to improve the effectiveness and scalability of phishing detection systems through advanced AI and machine learning techniques. Firstly, it focuses on analyzing and comparing existing phishing detection methods to identify their limitations, particularly the challenges of traditional rule-based systems in detecting evolving patterns. Secondly, the project aims to design a scalable and intelligent phishing detection system by leveraging natural language processing (NLP) and supervised machine learning approaches for improved accuracy. In addition, it seeks to extract meaningful features from emails and URLs using advanced techniques such as TF-IDF, n-grams, metadata analysis, and BERT embeddings, thereby enhancing the model's contextual understanding. Another key goal is to train and evaluate a variety of machine learning models, including Random Forest, Logistic Regression, and LSTM, to determine the most effective solution for phishing classification. Furthermore, the project intends to deploy the optimized model in a real-time environment using web APIs like Flask or FastAPI, enabling practical and seamless integration. It also aims to provide timely alerts and implement automatic blocking mechanisms to prevent phishing attempts proactively. Lastly, the performance of the model will be assessed using standard evaluation metrics such as accuracy, precision, recall, and F1-score, ensuring a robust and reliable phishing detection framework. Through these objectives, the project strives to build a comprehensive, adaptive, and high-performance solution to combat phishing threats effectively.

C. Scope

The scope of this project encompasses the development of an intelligent, machine learning-based phishing detection system designed to operate effectively in real-time environments. This solution aims to enhance cybersecurity by identifying phishing attempts through detailed analysis of emails and websites, leveraging both natural language processing (NLP) and advanced pattern recognition techniques. Core features of the system include the detection of suspicious content in email bodies, headers, and embedded URLs, as well as structural and content-based analysis of websites to uncover malicious intent. The platform is built to be scalable, capable of handling large volumes of data in real time, and adaptable to multiple languages and varied text formats to address a diverse range of phishing scenarios. Designed with integration in mind, the system can be embedded into email clients, web browsers, or deployed as standalone applications through APIs, making it flexible and accessible for various use cases. Emphasis is placed on security and privacy, incorporating data protection techniques such as masking and secure logging to ensure user confidentiality. Furthermore, the scope of the project includes continuous model retraining and updates, enabling the system to evolve alongside emerging phishing strategies and maintain its effectiveness. Ultimately, this phishing detection framework aims to provide a robust, adaptive, and user-friendly defense mechanism, empowering users and organizations to proactively combat cyber threats.

II. LITERATURE SURVEY

The literature survey reviews a variety of studies focused on AI-based phishing detection methods, covering both email and webpage threats. These works employ supervised learning models such as Random Forests, SVMs, CNNs, RNNs, and LSTMs, as well as hybrid deep learning approaches. Researchers like Bandahala and Suhaili (2021), Bauskar and Madhavaram (2022), and Kapoor (2023) highlight the effectiveness of deep neural networks and natural language processing (NLP) in identifying suspicious content, URLs, and metadata. CNNs and LSTMs are particularly noted for their ability to capture contextual patterns in sequential data, enhancing detection accuracy for phishing emails and obfuscated links. In parallel, webpage-based detection models incorporate multi-phase architectures, leveraging HTML structure, redirection analysis, WHOIS data, and blacklist verification, as discussed in studies by Moira and Dine. These models emphasize real-time classification through trained machine learning algorithms, with CNN-based frameworks demonstrating robust performance on complex, noisy inputs. Despite these advances, challenges remain, including high computational demands, the need for large labeled datasets, frequent updates to blacklists, and complex integration. Addressing these limitations through scalable, modular, and adaptive AI architectures is crucial for improving real-time phishing detection and cybersecurity resilience.

A. Literature Review

The literature survey explores a range of studies focusing on the application of machine learning, deep learning, and NLP-based approaches for phishing detection through emails and websites. These works collectively aim to enhance real-time cybersecurity by leveraging intelligent systems capable of adapting to evolving phishing techniques. One such study, *"The Role of Artificial Intelligence in Detecting and Preventing Phishing Emails"* by Tasneem A. Bandahala and Nur-Sheba S. Suhaili [1], empirically evaluates various deep learning models, including CNNs, RNNs, and LSTMs, for phishing email detection. Their approach focuses on using metadata, textual content, and sender attributes to train hybrid models with improved precision and recall.

Similarly, Sanjay Ramdas Bauskar and Chandrakanth Rao Madhavaram [2] introduce a dual-stage AI framework in *"AI-Driven Phishing Email Detection: Leveraging Big Data Analytics for Enhanced Cybersecurity"*. The model combines traditional ML with deep learning and incorporates big data processing to analyze large-scale email datasets. Feature extraction is used to process structural and content-based indicators, highlighting the effectiveness of hybrid techniques in large, dynamic datasets.

Maxwell Moira [3], in *"AI-Driven Phishing Detection: Enhancing Email Security Through Deep Learning"*, proposes a real-time phishing detection system using CNNs, RNNs, and XGBoost, complemented by NLP-based analysis of metadata, embedded URLs, and suspicious phrases. The study emphasizes automated and continuous threat monitoring for real-time detection and response.

A comprehensive adaptive phishing detection approach is detailed by Adrian-Viorel Andriu [4], in *"Adaptive Phishing Detection: Harnessing the Power of Artificial Intelligence for Enhanced Email Security,"* where AI techniques, including machine learning algorithms and natural language processing, are integrated to create a real-time system that dynamically learns from diverse phishing and legitimate email datasets. This system addresses the challenges of evolving phishing tactics by identifying social engineering patterns and adapting to new threats with high accuracy and low false positives..

Meera Kapoor [5] , in *"Comparative Analysis of AI Algorithms for Enhancing Phishing Detection in Real-Time Email Security,"* benchmarks multiple AI algorithms such as SVM, Random Forest, CNNs, and RNNs using standard metrics like accuracy, precision, recall, and ROC-AUC. This study provides valuable guidance on algorithm selection and performance tuning, enabling more effective real-time phishing detection

III. EXPLORATORY ANALYSIS OF LOCATIONAL DATA

A. Overview

The literature reveals a diverse set of supervised learning methods for phishing detection, each with distinct strengths and limitations. LSTM networks effectively capture sequential patterns and contextual cues in phishing emails, but they demand extensive training data and computational resources. Random Forest models, by contrast, are faster and more interpretable, performing well on structured features like sender metadata, though they struggle with obfuscated text. CNN-based approaches are powerful for analyzing raw webpage content and URLs, extracting hierarchical features from unstructured inputs, but they risk overfitting and have longer training times. Webpage-based phishing detection architectures integrate multiple techniques—including HTML form filtering, redirection checks, blacklist lookups, and WHOIS data analysis—to deliver real-time, robust protection. However, these systems require significant computational power and complex integration. Ultimately, while traditional machine learning models offer ease of deployment, deep learning methods provide higher accuracy, and hybrid systems that combine machine learning, NLP, and real-time intelligence strike an optimal balance for scalable, adaptive phishing detection solutions.

B. Existing Architecture:

Author (Year)	**Advantages**	**Limitations**
Tasneem A. Bandahala & Nur-Sheba S. Suhaili (2025)	Deep learning (CNN, RNN, LSTM) models achieved high accuracy in phishing detection using rich email datasets.	Requires large labeled datasets and significant computational resources for model training.
Sanjay Ramdas Bauskar & Chandrakanth Rao Madhavaram (2024)	Dual-stage AI and Big Data analytics improved large-scale phishing email detection.	Complex architecture; performance depends on data quality and preprocessing.
Maxwell Moira (2024)	CNNs, RNNs, and XGBoost enable real-time phishing detection with NLP analysis of content and metadata	Model explainability is limited; false positives possible with unseen patterns.
Iqra Naseer (2024)	Combines ML and NLP for fast and automated phishing detection with human-AI collaboration.	Overreliance on automation may overlook nuanced human-targeted attacks.
Adrian-Viorel Andriu (2024)	Adaptive AI with NLP detects linguistic manipulation and improves over time.	NLP models may struggle with multilingual or obfuscated phishing content.
Oladimeji Azeez Lamina & Waliu Adebayo Ayuba (2024)	Achieved 98.7% accuracy using adaptive reinforcement learning; strong against evolving threats.	High training complexity and slower response during real-time adaptation.
Meera Kapoor (2024)	Ensemble ML models (Random Forest, Gradient Boosting) achieved 95%+ accuracy; robust and interpretable.	Deep learning alternatives showed slower training and higher resource needs.
Faizal Dine (2024)	Combined ML, AI, and user education improved phishing resilience; ensemble models achieved high precision.tactics	Dependent on user awareness; lacks coverage of zero-day phishing .
Sajid Ali (2024)	Detects fear and urgency cues with better explainability.	Limited data and adaptability to new attacks.

Table I. Comparison of AI based Learning Papers

The existing system for "News Headlines Driven Stock Sentiment Analysis" involves several crucial steps and components. Firstly, news data undergoes cleaning and preprocessing to remove irrelevant information, correct errors, handle missing data, and standardize text formats. Subsequently, the preprocessed data is tokenized, stop words

are removed, and stemming or lemmatization may be applied for further analysis. VADER, a sentiment analysis tool, assigns polarity scores to each news headline based on the presence of positive or negative sentiment words. These sentiment scores are then calculated for each headline. The dataset is split into training and testing subsets, with deep learning models like RoBERTa, BERT, or other architectures trained using the training data. Performance evaluation involves validation using a separate dataset to tune hyperparameters and testing to assess generalization capabilities. Metrics like accuracy, precision, recall, and F1-score are utilized to evaluate and compare model performance. Conclusions drawn from these evaluations provide insights into the impact of news sentiment on stock market behavior, aiding investors and analysts in making informed decisions. Overall, the system employs advanced NLP techniques and deep learning models to predict stock market sentiment, offering valuable insights to stakeholders.

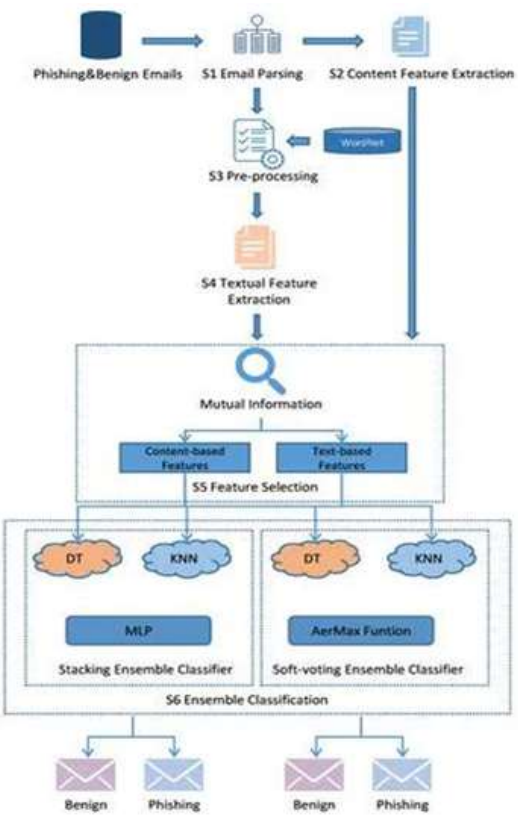

Fig. 3.1.1 Existing system architecture

Proposed System Architecture:

The proposed system architecture for news headlines driven stock sentiment analysis presents a comprehensive framework leveraging advanced natural language processing (NLP) techniques and deep learning models. The architecture involves data collection from financial data providers and news sources, followed by preprocessing, entity recognition, and sentiment analysis. Integration of datasets, feature engineering, and model training with techniques like recurrent neural networks (RNNs) or convolutional neural networks (CNNs) are employed for predicting stock price movements. The system aims to provide valuable insights for investors and analysts by combining news sentiment with historical stock data, ultimately enhancing decision-making in the stock market.

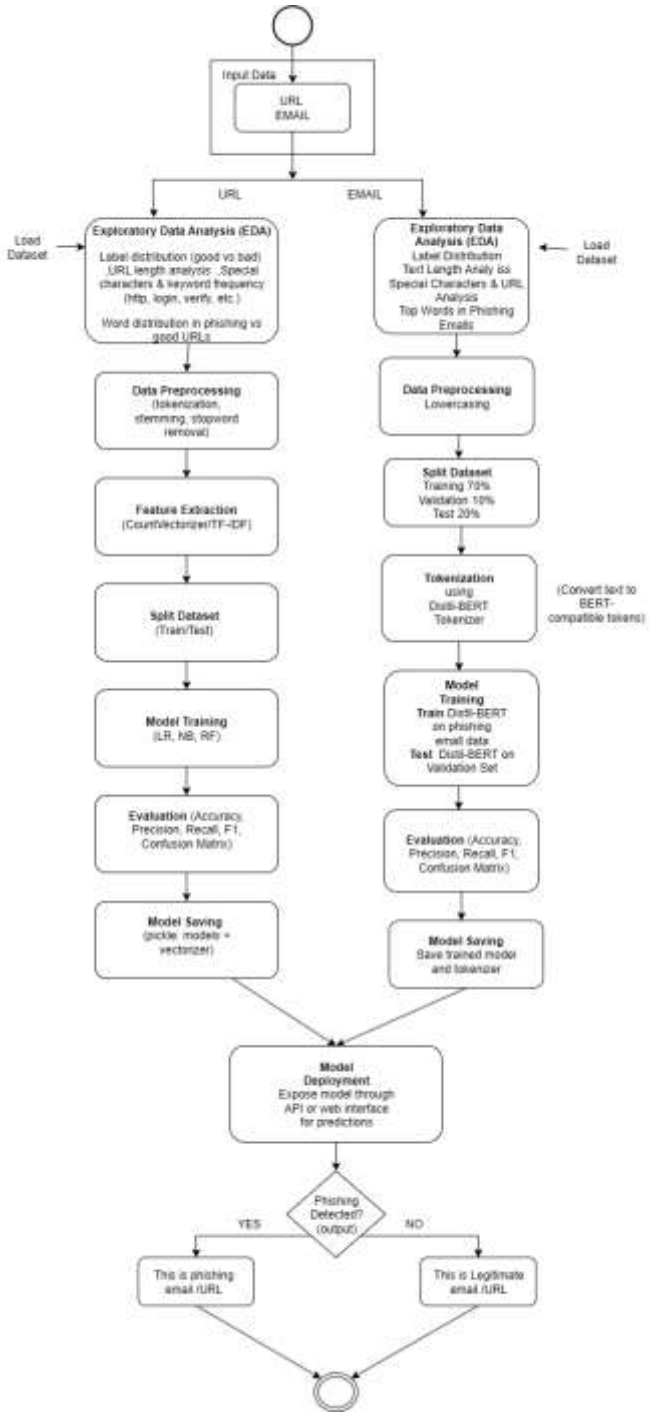

Fig. 3.1.2 Proposed system architecture

IV. IMPLEMENTATION DETAILS

1. Data Collection:

This module serves as the initial phase, where raw data comprising emails, messages, or communication logs is collected. This data can be sourced from open phishing datasets, email archives, or API-based real-time data capture systems. The purpose is to acquire a diverse corpus of both benign and phishing samples for model training and evaluation.

2. Data Preprocessing:

Once the raw data is collected, it undergoes rigorous preprocessing. The key steps involved are: Tokenization: Breaking text into individual words or tokens. Named Entity

Recognition (NER): Identifying entities like names, URLs, and domains. Sentiment Analysis: Understanding the emotional tone, which can often be suspicious or urgent in phishing. Noise Removal: Stripping unwanted characters, punctuation, and irrelevant patterns. This phase ensures that the data is normalized, cleaned, and in a suitable format for further processing.

3. Feature Extraction:

Preprocessed data is transformed into structured input features using two key strategies: 16 TF-IDF (Term Frequency-Inverse Document Frequency): Highlights important words across the corpus. LSTM (Long Short-Term Memory): Captures long-range dependencies and sequential context from textual input, critical for understanding subtle phishing cues. These features are essential for building models that understand both surface-level patterns and deeper contextual cues in text.

4. Analyze Data using NLP Models:

This module utilizes advanced Natural Language Processing (NLP) models to analyze the extracted features and learn textual patterns associated with phishing attacks. These models are capable of: Recognizing urgency, threats, or impersonation attempts. Detecting language styles commonly used in phishing (e.g., misspellings, emotional triggers).

5. Detect Phishing:

Using the insights from NLP analysis, the system classifies each message or email as either phishing or benign. It leverages machine learning algorithms trained on labeled datasets for binary classification. Models like Logistic Regression, Random Forest, or Deep Learning (LSTM/MLP) can be used here.

6. Validate Result:

Post-classification, the output is validated using ground truth labels or human-reviewed examples. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess the model's performance..

7. Provide Result:

The final decision (phishing or benign) is shared with the user or forwarded to downstream systems (e.g., email filters, alert systems). A UI/console interface may also display flagged messages, prediction confidence, and recommended actions.

8. Generate Reports & Deploy Updated Model:

Based on validation results, performance reports are generated. This feedback loop allows for continuous model improvement. New data can be added for retraining, and the system can be redeployed with improved performance and threat detection accuracy.

9. Evaluate Performance:

This ongoing process assesses the efficiency of the system, focusing on: Detection speed False positive and false negative rates Robustness against new phishing styles Adaptability across platforms Regular performance evaluations ensure that the phishing detection model remains reliable, secure, and effective against evolving threats. This proposed architecture bridges the gap between static phishing filters and modern adaptive solutions.

By combining NLP, machine learning, and deep learning, it ensures robust detection across diverse phishing formats and maintains a proactive defense mechanism suitable for real-time applications in cybersecurity, email protection, and fraud prevention

Algorithm:

A. LSTM (Long Short-Term Memory):

1. Introduction to LSTM:

Brief overview of LSTM as a type of recurrent neural network (RNN) architecture. Explanation of its ability to retain long-term dependencies and handle sequential data.

2. LSTM Architecture:

Description of LSTM architecture, including input, output, forget, and memory cells. Explanation of how LSTM units process sequential input data and update their internal state.

3. Preprocessing for LSTM:

Data preparation steps such as tokenization, padding, and sequence length normalization. Discussion on preparing sequential input data suitable for LSTM model training.

5. LSTM Model Building:

Steps involved in constructing an LSTM model using frameworks like TensorFlow or PyTorch. Configuration of LSTM layers, including the number of units, activation functions, and dropout regularization.

6. LSTM Model Training:

Process of training the LSTM model on prepared sequential input data. Explanation of optimization algorithms, loss functions, and training parameters.

7. LSTM Model Evaluation:

Assessment of LSTM model performance using metrics like accuracy, precision, recall, and F1-score. Validation techniques such as cross-validation to ensure robustness of the model.

B. BERT (Bidirectional Encoder Representations from Transformers):

1. Introduction to BERT:

Overview of BERT as a transformer-based deep learning model for natural language understanding tasks. Description of its pre-training on large text corpora and fine-tuning for specific downstream tasks.

2. BERT Architecture:

Explanation of BERT architecture, including encoder layers, attention mechanisms, and self-attention. Discussion on how BERT captures contextual information bidirectionally from input text.

3. Preprocessing for BERT:

Data preprocessing steps specific to BERT, such as tokenization, padding, and special token additions. Explanation of BERT's tokenization techniques, including WordPiece tokenization.

4. BERT Model Fine-tuning:
Process of fine-tuning pre-trained BERT models on task-specific datasets. Techniques for adapting BERT to downstream tasks such as sentiment analysis or text classification.

5. BERT Model Evaluation:
Evaluation of fine-tuned BERT models on test datasets using appropriate evaluation metrics. Comparison with baseline models to assess the effectiveness of BERT for the given task.

C. TF-IDF (Term Frequency-Inverse Document Frequency) with NLTK:

TF-IDF (Term Frequency-Inverse Document Frequency) is a fundamental technique in natural language processing (NLP) for assessing the significance of words in documents relative to a larger corpus. Initially, TF-IDF involves calculating scores for terms within documents, considering both their frequency within the document and their rarity across the corpus. This process requires preprocessing steps like tokenization, stop word removal, and stemming to prepare the text data adequately. Implementation of TF-IDF calculations can be achieved using NLTK's built-in functions or custom code, with discussions on parameter tuning for optimizing scores. Once computed, TF-IDF scores find application in various text analysis tasks, including document similarity measurement and keyword extraction. Moreover, TF-IDF integration into machine learning pipelines facilitates tasks like text classification and clustering, offering insights into the relevance and importance of terms within textual data.

1) Hardware and Software Specifications

For our project the required specifications are given in Table 2 and Table 3 respectively.

Table 2 . Hardware details

Components	Specifications
Processor	Intel i5/i7 or ARM 64-bit (cloud-hosted if required)
RAM	Minimum 8 GB (recommended 16 GB)
Storage	SSD, 256 GB or higher
Network	High-speed internet for API communication and real time updates
GPU	Integrated GPU (e.g., Intel UHD or Iris Graphics)

Table 3. Software Details

Operating System	Windows
Programming Language	Python (for backend & ML), React (for UI)
Database	NoSQL (MongoDB)

IV. Conclusion

The proposed AI-driven phishing detection system demonstrates the potential of integrating Machine Learning (ML) and Natural Language Processing (NLP) to effectively combat evolving phishing threats in digital communication. By leveraging advanced models such as Random Forest, LSTM, and BERT, the system accurately identifies malicious intent in emails and websites through linguistic pattern recognition, feature extraction, and contextual analysis. Unlike traditional rule-based systems that rely on static signatures or keyword matching, this intelligent framework adapts dynamically to new phishing strategies, ensuring higher accuracy and robustness.

Through comprehensive preprocessing, feature engineering, and real-time classification, the system provides a scalable, automated, and adaptive solution capable of safeguarding users from deceptive cyberattacks. Its integration as a Flask-based REST API further enables seamless deployment in browsers, email clients, and enterprise environments.

Experimental evaluations highlight the model's strong performance in terms of accuracy, precision, recall, and F1-score, validating its reliability for practical cybersecurity applications. Moreover, the system's modular architecture supports continuous learning, ensuring resilience against emerging phishing variants and zero-day attacks.

In conclusion, the fusion of AI and NLP provides a powerful defense mechanism against phishing threats, significantly enhancing the efficiency and intelligence of cybersecurity systems. Future work may involve incorporating multilingual support, transformer-based architectures, and reinforcement learning to further improve adaptability and real-time detection in large-scale environments.

References

[1] Tasneem A. Bandahala, Nur-Sheba S. Suhaili (JAN 2025) The Role of Artificial Intelligence in Detecting and Preventing Phishing Emails.

[2]Sanjay Ramdas Bauskar,Chandrakanth Rao Madhavaram(2024) AI-Driven Phishing Email Detection: Leveraging Big Data Analytics for Enhanced Cybersecurity.

[3]Maxwell Moira (2024) AI-Driven Phishing Detection: Enhancing Email Security Through Deep Learning.

[4] Adrian-Viorel ANDRIU(2024) Adaptive Phishing Detection: Harnessing the Power of Artificial Intelligence for Enhanced Email Security.

[5] Meera Kapoor(2024) Comparative Analysis of AI Algorithms for Enhancing Phishing Detection in Real-Time Email Security.

6] Faizal Dine (2024) Enhancing Phishing Threat Detection and Resilience: Leveraging Machine Learning, AI, and User Education in Cybersecurity.

[7] Sajid Ali(2024)The Role of AI in Social Engineering Attack Prevention: NLP-Based Solutions for Phishing and Scams.

[8] Shreyas Kumar, Anisha Menezes (2024) What The Phish! Effects of AI on Phishing Attacks and Defense.

[9]Oladimeji Azeez Lamina , Waliu Adebayo Ayuba (2024) Ai-Powered Phishing Detection And Prevention.

[10]Iqra Naseer (2024) The role of artificial intelligence in detecting and preventing cyber and phishing attacks.