GLOBAL YOUTH SCIENCE JOURNAL

AI-BASED PHISHING AND SCAM MESSAGE DETECTOR

A Proposed Machine Learning and Natural Language Processing Framework for Identifying Fraudulent Digital Messages

AUTHORS: [Enter all authors' full names]

AFFILIATION: Department of Computer Science and Engineering, AAA College of Engineering and Technology, Sivakasi, Tamil Nadu, India

PROJECT GUIDE: [Enter guide's name and designation]

OVERVIEW

This manuscript proposes an artificial intelligence-based framework for identifying phishing and scam messages across SMS, email and chat applications. It outlines a planned natural language processing and supervised machine learning workflow, together with message classification and user-warning stages; implementation and experimental evaluation remain future work.

KEYWORDS: Phishing detection; Scam classification; Machine learning; Natural language processing; Cybersecurity

ABSTRACT

Phishing and scam messages distributed through SMS, email and chat applications are used to deceive recipients, obtain sensitive information and direct users to potentially harmful websites. Manual identification can be difficult because fraudulent content may imitate trusted organisations and frequently changes its wording. This paper proposes an artificial intelligence-based framework for classifying digital messages as legitimate, phishing or scam by combining natural language processing (NLP) with supervised machine learning. The proposed workflow includes collecting labelled message data, preprocessing text, extracting linguistic and URL-related features, training classification models and presenting a warning when suspicious content is detected. Python-based tools and machine learning libraries are considered for the planned development. The framework aims to support safer digital communication and improve users' awareness of online fraud. As the work remains at the proposal stage, implementation, testing and quantitative performance evaluation are future activities.

INTRODUCTION

Digital communication through SMS, email and messaging applications has become part of everyday life. However, these channels are also used to distribute phishing and scam messages. Such messages may imitate banks, delivery services, online platforms or familiar contacts to persuade recipients to reveal passwords, one-time passwords, financial details or other personal information. Some messages also contain links to deceptive or malicious websites.

Manual inspection is time-consuming and may not be reliable for users who receive many messages. A learning-based approach can analyse patterns in message content and help identify suspicious communication. This project proposes a framework that applies NLP and machine learning to classify messages and provide understandable alerts to users.

Fraudulent messages can be difficult to distinguish from genuine communication, particularly when attackers change message wording, use shortened URLs or imitate trusted senders. The project

Global Youth Science Journal Preprints - Open Access (CC BY 4.0)

therefore addresses the need for a system that can analyse message text and relevant URL characteristics and classify messages into legitimate, phishing or scam categories.

The project aims to design a framework for SMS, email and chat messages; apply NLP preprocessing; investigate supervised machine learning classifiers; examine linguistic, structural and URL-related indicators; and propose a user alert for messages classified as potentially fraudulent. The project background identifies research directions in phishing detection based on feature engineering, supervised machine learning and learning-based analysis of phishing emails. Complete bibliographic details for the studies referred to in the original project material were not supplied and must be verified before submission.

RESULTS

No experimental results are available at this stage because the proposed system has not yet been implemented or tested. Accordingly, this manuscript reports no measured accuracy, precision, recall, F1-score or other performance values. Following implementation, the team plans to evaluate the classifier using a labelled test dataset and report suitable quantitative measures.

DISCUSSION

The proposed framework is intended to support the identification of suspicious digital messages by combining message-text analysis with selected URL-related features. Its planned output has three categories: legitimate, phishing and scam. A cautionary alert is intended when the predicted category is phishing or scam. The proposed system should be treated as a support tool rather than a guarantee that a message or link is safe.

The initial scope is limited to text-based SMS, email and chat messages. Detection performance is expected to depend on dataset quality, class balance and how closely the training data represents current scam patterns. Regional languages, newly emerging attack methods, shortened URLs and highly personalised messages may require additional analysis. The project is primarily aligned with Sustainable Development Goal 16 through its intended contribution to safer digital communication and awareness of online fraud, and is also related to Goal 9 through the proposed use of AI and machine learning for digital security.

MATERIALS AND METHODS

The proposed framework is organised into six connected stages. First, labelled examples of legitimate and fraudulent SMS, email and chat messages will be identified and obtained. Second, the message text will be cleaned, its case standardised, unnecessary characters removed and content tokenised. Third, text will be represented for analysis and linguistic, structural and URL-related features will be examined. Fourth, suitable supervised machine learning classifiers will be trained using labelled examples. Fifth, an input message will be classified as legitimate, phishing or scam. Finally, a cautionary warning will be displayed when the predicted category is phishing or scam.

Stage	Planned activity
1. Data collection	Identify and obtain labelled examples of legitimate and fraudulent SMS, email and chat messages.
2. Preprocessing	Clean message text, standardise case, remove unnecessary characters and tokenise the content.
3. Feature extraction	Represent text and examine linguistic, structural and URL-related features.
4. Model training	Train suitable supervised machine learning classifiers using labelled examples.
5. Classification	Classify an input message as legitimate, phishing or scam.

Global Youth Science Journal Preprints - Open Access (CC BY 4.0)

Stage	Planned activity
6. Alert generation	Display a cautionary warning when the predicted category is phishing or scam.

Table 1. Planned stages of the proposed message-detection framework.

The proposed system architecture is: Message Input → Text Preprocessing → Feature Extraction → Machine Learning Classifier → Category Prediction → User Alert. A message supplied by the user is passed to the preprocessing module. The processed text and selected features are provided to a trained classifier, and the predicted category is displayed. If the message is classified as phishing or scam, the interface is intended to show a cautionary alert. This architecture remains to be implemented.

The proposed software environment includes Python 3.x, Jupyter Notebook or Visual Studio Code, Pandas and NumPy for data handling, NLTK for text preprocessing, Scikit-learn for machine learning and Flask for a possible web interface. These are candidate tools and should be updated to match the tools actually selected by the project team. Indicative hardware is a laptop or desktop with an Intel i5-equivalent processor, 8 GB RAM, adequate storage and internet access; these are suggested baseline requirements rather than measured system requirements.

The planned work comprises literature review and dataset identification, data collection and preprocessing, feature extraction and model development, testing and evaluation, and preparation of the final documentation. The sequence and duration may be adjusted according to the team’s academic schedule.

ACKNOWLEDGMENTS

No acknowledgments were provided in the source manuscript. This section may be completed by the authors if applicable.

REFERENCES

The source manuscript did not contain complete bibliographic references. It stated that the exact studies mentioned in the original project poster should be copied from the original papers and verified before journal submission. The authors should add the verified sources in the journal’s required Chicago style, including full URLs or DOIs beginning with “https://”. This placeholder is not a completed reference list.

FIGURES AND FIGURE TITLES/CAPTIONS

Figure 1. Proposed message-detection workflow. Message input is processed and represented through text preprocessing and feature extraction before classification. The predicted category is then presented to the user, with a cautionary alert intended for phishing or scam predictions. This is a conceptual architecture and has not yet been implemented.

TABLES WITH TITLES/CAPTIONS

Table 1 presents the six planned stages of the framework, from labelled data collection through alert generation. No experimental data table is available at the proposal stage.

APPENDIX

No appendix was supplied. The authors may include supplementary material only where it is essential to understanding the manuscript.

Global Youth Science Journal Preprints - Open Access (CC BY 4.0)