

AI-Based Phishing Attack Detection Using Natural Language Processing (NLP)

Suraj Kalbhor

Department of Computer Science

Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri ,Pune, Maharashtra,
India

kalbhorsuraj18@gmail.com

sakshi Kadam

Department of Computer Science

Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri, Pune,
Maharashtra, India

sakshikadam5970@gmail.com

ABSTRACT

In today's digital world, phishing attacks have become one of the most serious cybersecurity threats. Cybercriminals use fake emails, messages, and websites to trick users into sharing sensitive information such as passwords, bank details, and personal data. Traditional phishing detection methods are often unable to detect advanced and evolving attacks effectively.

Artificial Intelligence (AI) combined with Natural Language Processing (NLP) provides a powerful solution to detect phishing attacks. NLP helps in analyzing the textual content of emails and messages to identify suspicious patterns, keywords, and malicious intent. AI models such as machine learning and deep learning improve detection accuracy by learning from large datasets.

This research paper focuses on the use of AI and NLP techniques for phishing detection. It explains the working process, methodologies, advantages, and challenges of implementing such systems. The study also highlights how these technologies can improve cybersecurity and protect users from online fraud.

KEYWORDS

Phishing Detection, Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning, Cybersecurity, Email Filtering, Text Classification

OBJECTIVE OF THE STUDY

The main objective of this study is to develop an intelligent system that can detect phishing attacks by analyzing the language used in digital communication such as emails, messages, and websites. Since phishing attacks often rely on manipulating human emotions through text, this study focuses on applying Natural Language Processing (NLP) techniques to identify suspicious patterns in language.

Another important objective is to compare different machine learning and deep learning models to determine which approach is most effective in detecting phishing content. The study also aims to improve detection accuracy while minimizing false positives, so that legitimate messages are not incorrectly flagged as threats.

Additionally, the research seeks to understand common linguistic features used in phishing attempts, such as urgency, fear, or impersonation, and use these insights to strengthen detection models. Finally, the study aims to contribute toward building safer digital communication systems that can protect users from online fraud.

INTRODUCTION

With the rapid growth of internet usage, cyberattacks have also increased significantly. One of the most common types of cyberattacks is phishing, where attackers pretend to be trusted entities to steal sensitive information. Phishing emails often look genuine, making it difficult for users to identify them.

Traditional detection systems rely on blacklists or rule-based approaches, which are not effective against new and unknown phishing attacks. Therefore, advanced technologies like Artificial Intelligence and Natural Language Processing are used to improve detection systems.

NLP allows machines to understand and analyse human language. By using NLP techniques, systems can examine email content, detect suspicious patterns, and classify messages as phishing or legitimate. AI models further enhance this process by learning from past data and improving over time.

LITERATURE REVIEW

Many researchers have worked on phishing detection using AI and NLP techniques.

- **Hosseinzadeh et al. (2025)** proposed a BERT-based deep learning model for phishing email detection. Their study showed improved accuracy compared to traditional machine learning models.
- **Elsadig et al. (2022)** developed an intelligent system using NLP feature extraction for phishing URL and email detection. The system achieved high precision and recall.
- **Sahingoz et al. (2019)** used machine learning algorithms like Random Forest and Support Vector Machine (SVM) to detect phishing websites.
- **Verma and Hossain (2017)** focused on detecting phishing emails using NLP-based feature extraction methods.
- **Abdelhamid et al. (2014)** introduced a phishing detection model using text classification techniques.

These studies indicate that AI and NLP significantly improve phishing detection accuracy. However, challenges like data imbalance and evolving attack patterns still exist.

METHODOLOGY

Data Collection

Data is collected from various sources such as:

- Phishing email datasets (e.g., Enron dataset)
- Spam email repositories
- Real-world email samples
- Phishing websites and URLs

Data Preprocessing

Before analysis, the data is cleaned and prepared:

- Removal of stop words
- Tokenization (splitting text into words)
- Stemming and lemmatization
- Removal of special characters

Feature Extraction using NLP

Important features are extracted using NLP techniques:

- Term Frequency-Inverse Document Frequency (TF-IDF)
- Bag of Words (BoW)
- Word embeddings (Word2Vec, GloVe)
- Detection of suspicious keywords (e.g., “urgent”, “verify account”)

Model Implementation

Various AI models are used:

- Machine Learning Models:
 - Naive Bayes
 - Support Vector Machine (SVM)
 - Random Forest
- Deep Learning Models:
 - Recurrent Neural Networks (RNN)

- Long Short-Term Memory (LSTM)
- BERT (Bidirectional Encoder Representations from Transformers)

System Workflow

1. Input email/message
2. Preprocessing using NLP
3. Feature extraction
4. Model classification
5. Output: Phishing or Legitimate

RESULTS AND DISCUSSION

Accuracy Improvement

AI-based models show higher accuracy compared to traditional methods. Deep learning models like BERT achieve accuracy above 95%.

Real-Time Detection

The system can detect phishing emails in real-time, helping users avoid scams.

Reduced False Positives

Advanced NLP techniques reduce incorrect classification of legitimate emails.

Limitations and Challenges

- High computational cost
- Need for large datasets
- Evolving phishing techniques
- Data privacy concerns

Future Scope

Future systems can integrate:

- Real-time browser protection
- Multi-language phishing detection
- Integration with cloud-based security systems
- Use of advanced AI models like GPT and transformers

CONCLUSION

AI-based phishing detection using NLP is an effective solution for modern cybersecurity challenges. It helps in identifying malicious emails and protecting users from fraud. NLP techniques enable better understanding of textual content, while AI models improve accuracy and adaptability.

Although there are challenges such as high cost and evolving attack methods, continuous advancements in AI and machine learning will enhance the effectiveness of phishing detection systems. These technologies will play a crucial role in securing digital communication in the future.

REFERENCES

- [1] Hosseinzadeh, M., Akbari, A., & Daryaei, P. (2025). *Improving Phishing Email Detection Performance Through a BERT-Based Deep Framework*.
- [2] Elsadig, M., Osman, H., & Abdalla, F. (2022). *Intelligent Deep Learning Approach for Phishing URL and Email Detection*.
- [3] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). *Machine learning based phishing detection*.
- [4] Verma, R., & Hossain, N. (2017). *Semantic feature selection for phishing email detection*.
- [5] Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). *Phishing detection using association rule mining*.