International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

Intelligent Phishing Detection and Spam Classification System Using Rule-Based Text Analysis

CHINTAPALLI ROHITH KUMAR NARASIMHA

PG Scholar. Department of MCA, DNR College, Bhimavaram, Andhra Pradesh

B. Suryanarayana Murthy

(Assistant Professor), Master of Computer Applications, DNR College, Bhimavaram, Andhra Pradesh

ABSTRACT

With the rapid growth of digital communication, email and web-based interactions have become essential components of daily life. However, this growth has also led to a significant rise in cyber threats, particularly phishing attacks, which aim to deceive users into revealing sensitive information such as passwords, banking details, and personal credentials. This project presents an intelligent phishing detection and spam classification system developed using a rule-based text analysis approach within a Django web framework.The system is designed to identify and categorize phishing attacks and spam messages based on predefined keyword patterns and heuristic rules. It consists of multiple modules, including user authentication, phishing detection during login, email composition and classification, malicious URL checking, and user feedback collection. When a user attempts to log in, the system analyzes the entered email string using regular expressions and keyword matching techniques to detect potential phishing indicators such as suspicious domains, numeric patterns, and misleading structures. Based on this analysis, the system classifies the input into categories like spear phishing, whaling phishing, pharming attacks, or search engine phishing.

In addition to login analysis, the system includes an email classification module that processes message content and categorizes it into different domains such as financial, social networking, cloud storage, and others. It also determines whether an email should be marked as spam or delivered to the inbox based on the presence of malicious or suspicious keywords. This classification helps users identify harmful content before interacting with it.Another key feature is the phishing thread detection module, which analyzes user-provided text or URLs to identify hidden malicious patterns related to sensitive information such as API keys, bank IDs, and private identifiers. If such patterns are detected, the system flags the content as phishing; otherwise, it is considered legitimate.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

The proposed system emphasizes simplicity and efficiency by avoiding complex machine learning models and instead relying on a transparent and interpretable rule-based approach. This makes it easier to implement, maintain, and understand while still providing effective protection against common phishing techniques.Overall, this project contributes to enhancing cybersecurity awareness and protection by providing users with a proactive tool to detect and prevent phishing attacks and spam messages in real time. It can be further extended by integrating machine learning algorithms for improved accuracy and adaptability to evolving cyber threats.

Keywords: Phishing Detection, Spam Classification, Cybersecurity, Email Filtering, Rule-Based System, Text Mining, Django Web Application, Threat Analysis, Malware Detection, Secure Communication

I. INTRODUCTION

In the modern digital era, communication systems such as email, messaging platforms, and online services play a vital role in both personal and professional environments. While these technologies provide convenience and efficiency, they also expose users to various cybersecurity threats. Among these threats, phishing attacks have emerged as one of the most prevalent and dangerous forms of cybercrime.Phishing is a fraudulent attempt to obtain sensitive information by disguising as a trustworthy entity in electronic communication. Attackers often use deceptive emails, fake websites, or malicious links to trick users into revealing confidential data such as login credentials, credit card numbers, and personal identification details. These attacks can lead to severe consequences, including financial loss, identity theft, and data breaches.

Traditional methods of phishing detection rely heavily on blacklists, signature-based detection, or user awareness. However, these approaches have limitations. Blacklists require continuous updates, signature-based systems struggle to detect new or evolving threats, and user awareness alone is often insufficient due to the increasing sophistication of phishing techniques.To address these challenges, this project proposes a rule-based phishing detection and spam classification system implemented using the Django web framework. The system focuses on analyzing textual patterns in user inputs, emails, and URLs to identify potential threats. By leveraging regular expressions and predefined keyword sets, the system can detect suspicious patterns indicative of phishing attempts.

The application includes several functional modules. The user authentication module ensures secure login and registration. During login, the system evaluates input data to identify phishing characteristics. The email classification module processes user-

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

generated messages and categorizes them based on content analysis, helping users distinguish between legitimate and malicious communications. Additionally, the phishing detection module examines URLs and text inputs to identify hidden threats associated with sensitive data exposure.One of the key advantages of this system is its simplicity and interpretability. Unlike complex machine learning models, the rule-based approach provides clear reasoning behind each classification decision, making it easier for developers and users to understand the system's behavior. This transparency is particularly useful in educational and research contexts.Furthermore, the system is designed to be scalable and extensible. New rules and keywords can be easily added to improve detection accuracy and adapt to emerging phishing techniques. The integration of user feedback also allows for continuous improvement and refinement of the system.

In conclusion, this project aims to provide an effective and user-friendly solution for detecting phishing attacks and classifying spam messages. By combining rule-based analysis with a web-based interface, it offers a practical tool for enhancing cybersecurity and protecting users from online threats.

II. LITERATURE SURVEY (WITH EXISTING METHODS)

Phishing detection has been an active area of research in cybersecurity due to the increasing frequency and sophistication of cyberattacks. Various techniques have been proposed in the literature to identify and mitigate phishing threats, ranging from traditional rule-based systems to advanced machine learning and deep learning approaches.One of the earliest approaches to phishing detection involves blacklist-based methods. These systems maintain a database of known malicious URLs and domains. When a user attempts to access a website or open an email, the system checks it against the blacklist. While this method is simple and efficient, it suffers from a major limitation: it cannot detect new or previously unknown phishing sites, making it ineffective against zero-day attacks.

Heuristic-based approaches improve upon blacklist methods by analyzing features of URLs, emails, and web pages. These features include URL length, presence of special characters, domain age, and suspicious keywords. Rule-based systems fall under this category and are widely used due to their simplicity and interpretability. However, they require continuous updates and may produce false positives if not carefully designed.Machine learning techniques have gained popularity for phishing detection due to their ability to learn patterns from data. Algorithms such as Decision Trees, Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression are commonly used. These models are trained on labeled datasets containing phishing and legitimate samples,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

enabling them to classify new inputs with high accuracy. However, machine learning models require large datasets, feature engineering, and computational resources.

Deep learning approaches, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have also been explored for phishing detection. These models can automatically extract complex features from raw data, such as text and URLs, without manual feature engineering. While they offer improved accuracy, they are computationally expensive and less interpretable compared to traditional methods.Another emerging approach is the use of Natural Language Processing (NLP) techniques to analyze email content and detect phishing attempts. NLP methods focus on identifying linguistic patterns, sentiment, and semantic anomalies in text. These techniques can effectively detect social engineering tactics used in phishing emails.

Hybrid systems that combine multiple approaches have also been proposed to improve detection performance. For example, combining rule-based filtering with machine learning models can provide both interpretability and adaptability.

Despite the advancements in phishing detection techniques, challenges remain. Attackers continuously evolve their strategies, making it difficult for static systems to keep up. Additionally, achieving a balance between detection accuracy and false positives is a critical concern.The proposed system in this project adopts a rule-based approach due to its simplicity, ease of implementation, and suitability for real-time applications. While it may not match the accuracy of advanced machine learning models, it provides a strong foundation for detecting common phishing patterns and can be extended with additional techniques in the future.

III. EXISTING SYSTEM

Existing phishing detection systems primarily rely on traditional methods such as blacklist filtering, signature-based detection, and basic heuristic analysis. Blacklist-based systems maintain a repository of known malicious websites and email addresses. When a user interacts with a link or email, it is checked against this database. Although effective for known threats, these systems fail to detect newly created phishing sites, making them vulnerable to zero-day attacks.Signature-based detection methods identify phishing attempts by matching patterns or signatures of known attacks. These systems are widely used in antivirus and email filtering solutions. However, they require frequent updates and are not effective against novel or modified phishing techniques.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

Heuristic-based systems analyze specific features of emails and URLs, such as suspicious keywords, unusual formatting, and domain inconsistencies. While these systems can detect previously unknown threats, they often generate false positives and require careful tuning of rules.More advanced systems use machine learning algorithms to classify phishing and legitimate content. These systems offer higher accuracy but require large datasets, extensive training, and computational resources. Additionally, they may lack transparency, making it difficult to understand the reasoning behind their decisions.

Overall, existing systems face challenges in terms of adaptability, accuracy, and interpretability. The proposed system addresses these limitations by using a rule-based approach that is simple, transparent, and easy to update, providing an effective solution for detecting common phishing attacks and spam messages.

IV. PROPOSED METHOD

The proposed system is an intelligent phishing detection and spam classification platform developed using a rule-based approach integrated within a Django web application. Unlike traditional systems that rely heavily on blacklists or computationally expensive machine learning models, this system focuses on real-time text analysis using predefined patterns and keyword matching techniques.The system consists of multiple modules, including user authentication, phishing detection during login, email classification, phishing URL checking, and feedback management. During login, the system analyzes user inputs such as email IDs using regular expressions and rule-based logic to identify suspicious patterns like numeric anomalies, domain inconsistencies, and misleading structures. Based on these features, it classifies attacks into categories such as spear phishing, whaling phishing, pharming, and search engine phishing.The email classification module processes message content and categorizes emails into domains such as financial, cloud services, social networking, and others. It also determines whether the message should be marked as spam or inbox based on the presence of predefined malicious keywords. This ensures early detection of harmful content before user interaction.

Additionally, the system includes a phishing thread detection module that analyzes user-provided text or URLs for sensitive keywords such as API keys, bank IDs, and private identifiers. If such patterns are detected, the system flags the content as phishing.This rule-based approach ensures transparency, ease of implementation, and quick decision-making. Although advanced machine learning systems achieve high accuracy, rule-based systems remain effective for detecting common phishing patterns and are easier to maintain and extend. Studies show that combining feature-based detection with pattern analysis improves real-time phishing detection efficiency

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

V. IMPLEMENTATION

The implementation of the proposed phishing detection and spam classification system is carried out using the Django web framework, which follows the Model-View-Template (MVT) architecture. The system is divided into multiple modules, each responsible for specific functionalities.

The user authentication module enables users to register and log in securely. During login, the system captures user credentials and applies rule-based phishing detection using Python's regular expressions (re module). The email input is analyzed by splitting it into tokens and checking for suspicious patterns such as the presence of "http", domain suffixes like ".com" or ".info", and numeric characters. Based on these features, the system classifies the input into phishing categories and stores the results in the database.

The email classification module is implemented in the user dashboard. When a user composes an email, the message content is tokenized and compared against predefined keyword lists representing different threat categories such as violence, financial fraud, threats, illegal activities, and social engineering tactics. These keywords are stored in arrays and matched using conditional logic. The system counts occurrences of each category and assigns the email to the category with the highest frequency.Spam detection is performed by checking whether any suspicious keywords are present in the message. If detected, the email is marked as "spam"; otherwise, it is classified as "inbox." The classified email is then stored in the database using Django models.

The phishing URL detection module allows users to input text or URLs. The system compares extracted tokens against a list of sensitive keywords such as "bankid", "apikey", and "privateid". If any match is found, the system flags the input as phishing; otherwise, it is considered legitimate. The results are stored for future reference and analysis.

The system also includes modules for viewing inbox and spam emails, deleting messages, and collecting user feedback. All data interactions are handled using Django's ORM, ensuring efficient database operations.Overall, the implementation is simple, modular, and scalable. While modern systems use deep learning for phishing detection, rule-based systems remain useful due to their low computational cost and interpretability .

VI. ALGORITHMS

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

The system uses rule-based algorithms combined with text processing techniques for phishing detection and spam classification. These algorithms rely on pattern matching, keyword analysis, and frequency comparison.

1. Login Phishing Detection Algorithm

- Input: Email/username string
- Process:
    1. Tokenize input using regular expressions
    2. Check for patterns such as “http”, domain extensions (.com, .in, etc.), and numeric characters
    3. Apply conditional rules:
        - Numeric presence → Spear phishing
        - Domain keywords → Pharming
        - “http” presence → Search engine phishing
        - Default → Whaling phishing
- Output: Attack type classification

2. Email Classification Algorithm

- Input: Email message content
- Process:

    1. Tokenize text into words
    2. Compare words with predefined keyword lists (violence, financial, threat, etc.)
    3. Count frequency of matches in each category
    4. Select category with highest frequency

- Output: Email category

3. Spam Detection Algorithm

- Input: Email tokens
- Process:

    1. Check if any keyword matches predefined suspicious lists
    2. If match found → mark as spam

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

3. Else → mark as inbox

- Output: Spam/InBox classification

4. Phishing URL Detection Algorithm

- Input: URL/text
- Process:

1. Tokenize input
2. Compare with sensitive keyword list
3. If match found → phishing detected

- Output: Legitimate or phishing

These rule-based algorithms are efficient and interpretable, though modern approaches use machine learning and neural networks for higher accuracy .

VII. SYSTEM DESIGN

The system design follows a modular and layered architecture based on the Django MVT (Model-View-Template) framework. It ensures separation of concerns, scalability, and maintainability.

1. Architecture Overview

The system consists of three main layers:

- **Presentation Layer (Templates):** Handles user interface components such as login pages, dashboards, email forms, and result displays.
- **Application Layer (Views):** Contains business logic, including phishing detection, spam classification, and request handling.
- **Data Layer (Models):** Manages database interactions using Django ORM for storing user data, emails, phishing results, and feedback.

2. Module Design

a) User Module

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

Handles registration, login, and session management. It ensures secure access to the system and stores user credentials in the database.

b) Phishing Detection Module

Analyzes login inputs and URLs using rule-based logic and keyword matching. It identifies phishing patterns and stores detected attack types.

c) Email Classification Module

Processes email content and categorizes it into predefined domains using frequency-based keyword analysis.

d) Spam Filtering Module

Determines whether an email is spam or inbox based on keyword presence.

e) Data Management Module

Handles storage and retrieval of emails, phishing results, and user feedback.

3. Data Flow

1. User inputs data (login/email/URL)
2. System processes input using rule-based algorithms
3. Classification results are generated
4. Results are stored in the database
5. Output is displayed to the user

4. Design Advantages

- Modular structure improves maintainability
- Rule-based logic ensures transparency
- Low computational cost enables real-time processing
- Easily extendable with machine learning integration

Modern research emphasizes hybrid architectures combining rule-based and machine learning approaches for improved accuracy and adaptability .

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

SYSTEM DESIGN IMAGES

1. User Login Page

Description:
Displays login form where users enter email and password. If phishing patterns are detected in the email, the system shows the attack type.

Screenshot should include:

- Email & Password fields
- Login button
- Output message like: *"Spear Phishing Email"*

2. User Registration Page

Description:
Allows new users to register.

Screenshot should include:

- First Name, Last Name
- Email, User ID
- Password
- Gender
- Register button

3. User Dashboard (User Page)

Description:
Main interface where users compose emails and view classification results.

Screenshot should include:

- "To", "Subject", "Message" fields
- Submit button
- Output showing:
    - Category (e.g., Financial / Social Networking)

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

- Spam or Inbox status

4. Email Classification Result

Description:
Shows classification after sending message.

Screenshot should include:

- Message content
- Detected Category (e.g., *Financial*)
- Spam status (*spam/inbox*)

5. Inbox Page

Description:
Displays all non-spam emails.

Screenshot should include:

- Table/List of emails
- Sender, Subject, Message
- Delete option

6. Spam Page

Description:
Displays detected spam emails.

Screenshot should include:

- Emails marked as spam
- Delete option

7. Phishing URL Detection Page

Description:
Checks if entered text/URL is phishing.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

Screenshot should include:

- Input field for URL/text
- Submit button
- Output:
    - *"Phishing is Detected"* OR *"Legitimate"*

8. Phishing Detection History

Description:
Shows previous phishing checks.

Screenshot should include:

- List of checked URLs/text
- Result (Phishing / Legitimate)

9. User Details Page

Description:
Displays logged-in user information.

Screenshot should include:

- Name, Email, Gender
- User ID

10. Feedback Page

Description:
Allows users to submit feedback.

Screenshot should include:

- Feedback text area
- Submit button

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

VIII. CONCLUSION

The proposed phishing detection and spam classification system provides an effective and lightweight solution for identifying cyber threats in digital communication. By utilizing a rule-based approach, the system achieves real-time detection of phishing attacks and spam messages without the need for complex machine learning models.

The system successfully analyzes login inputs, email content, and URLs to identify suspicious patterns and classify them into relevant categories. Its modular design ensures scalability and ease of integration with additional features such as machine learning or NLP-based enhancements. The use of keyword matching and pattern recognition allows for transparent decision-making, making it suitable for both practical applications and educational purposes.

Although advanced machine learning and deep learning techniques have demonstrated higher accuracy in phishing detection, they require large datasets and computational resources. In contrast, the proposed system offers a simple and interpretable alternative that performs well for common phishing scenarios. Research indicates that combining rule-based methods with advanced techniques can further enhance detection accuracy and reduce false positives .

Future work can focus on integrating machine learning algorithms, real-time threat intelligence feeds, and NLP techniques to improve detection capabilities. Additionally, incorporating user behavior analysis and adaptive learning mechanisms can help the system evolve with emerging cyber threats.

In conclusion, this project contributes to improving cybersecurity by providing a practical and efficient tool for detecting phishing attacks and managing spam. It enhances user awareness and protection, making it a valuable solution in today’s digital landscape.

REFERENCES

1. Thakur et al., “Deep-Learning-Based Phishing Email Detection,” Electronics, 2023
2. Ghalechyan et al., “Phishing URL Detection with Neural Networks,” Scientific Reports, 2024
3. Bezerra et al., “Phishing Detection using Machine Learning,” Springer, 2024
4. Tamal et al., “Phishing URL Dataset,” Frontiers in Computer Science, 2024

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

5. Tamal et al., “Feature Vectorization for Phishing Detection,” Frontiers, 2024
6. Komosny, “Phishing Detection in Non-English Webpages,” Scientific Reports, 2025
7. Kritika, “Deep Learning for Phishing URL Detection,” Journal of Cyber Security Technology, 2024
8. Hajara, “Review on Phishing Website Detection,” IJSDR, 2023
9. Sensors Journal, “Comparative Study of Phishing Detection Models,” 2024
10. ScienceDirect, “Systematic Literature Review on Phishing Detection,” 2023
11. Chen et al., “Transaction Phishing Detection on Ethereum,” arXiv, 2024
12. Ovi et al., “PhishGuard Ensemble Model,” arXiv, 2024
13. Ige et al., “Survey on ML & DL Phishing Detection,” arXiv, 2024
14. Newaz et al., “Framework for Phishing Website Detection,” arXiv, 2024
15. Additional dataset-based phishing detection research, Frontiers, 2024