



# Cybersecurity Threat Detection on Social Media Using NLP and URL Intelligence.

Ms. Simran Shinde

Prashant Agrahari , Dhonde Krish

Pillai College of Arts Commerce & Science(Empowered Autonomous)

From Department of Computer Science

[simranshinde@mes.ac.in](mailto:simranshinde@mes.ac.in)

## Abstract:

Social media has become deeply embedded in everyday life, supporting communication, education, business activities, and large-scale information exchange. While this openness allows users to connect easily and share content instantly, it also exposes social media platforms to a wide range of cyber threats. Phishing attempts, online scams, impersonation, and the circulation of malicious links are increasingly common. Rather than attacking systems directly, cybercriminals often focus on manipulating human behavior by creating a sense of urgency, trust, or curiosity. Existing protection methods such as manual content review and static blacklist-based filtering struggle to keep pace with the constantly evolving and informal nature of social media communication.

This project proposes an intelligent cybersecurity threat detection framework that integrates Natural Language Processing (NLP) with URL intelligence to identify phishing and scam content on social media platforms. Textual data is analyzed through preprocessing and TF-IDF-based feature extraction, followed by supervised machine learning techniques to recognize language patterns frequently used in social engineering attacks. At the same time, embedded URLs are evaluated using structural and behavioral characteristics such as domain structure, suspicious tokens, and redirection behavior to calculate a link-level risk score. These textual and URL-based insights are then combined using a weighted risk fusion strategy to produce a unified threat score for each message.

Experimental evaluation demonstrates that the hybrid approach outperforms text-only detection methods by achieving higher accuracy, precision, and recall, especially in cases involving shortened or disguised links. The proposed framework provides a practical and scalable solution for real-time monitoring and early threat detection, enhancing cybersecurity protection across educational institutions, organizations, and online communities.

Keywords: Cybersecurity, Social Media, Phishing Detection, NLP, Malicious URLs, Machine Learning.

## 1. Introduction

The rapid growth of social media platforms has significantly changed the way people communicate, learn, and interact in the digital world. Platforms such as messaging applications, microblogging services, and social networking sites are now used daily for academic collaboration, professional networking, online marketing, and information sharing. Their ease of use, real-time interaction, and wide reach have made them essential digital spaces for individuals and organizations alike. However, this same accessibility has also created opportunities for cybercriminals to exploit users on a large scale.

Unlike traditional cyber attacks that focus on exploiting software or network vulnerabilities, many social media-based attacks are centered on social engineering. Attackers manipulate human emotions such as trust, urgency, fear, or

curiosity to convince users to perform unsafe actions. Common threats include phishing messages that steal login credentials, scam campaigns that promise rewards or financial benefits, impersonation of trusted individuals or organizations, and the distribution of malicious URLs. Once an account is compromised, it is often used to spread further attacks, increasing the overall impact of the threat.

Detecting such malicious activity on social media is particularly challenging due to the informal and dynamic nature of online communication. Messages are often short, use slang or emojis, and may mix multiple languages, making traditional keyword-based detection methods unreliable. Additionally, attackers frequently hide malicious links using URL shorteners, misleading domain names, or redirection chains, allowing them to bypass static blacklists and manual moderation systems. As a result, many harmful messages remain undetected until users are already affected.

To address these challenges, there is a growing need for intelligent and adaptive security solutions that can analyze both what a message says and how embedded links behave. Natural Language Processing (NLP) techniques make it possible to examine textual patterns and identify persuasive or suspicious language commonly used in scams and phishing attempts. At the same time, URL intelligence provides valuable technical indicators by analyzing link structure, complexity, and obfuscation techniques used by attackers.

This project proposes a hybrid cybersecurity threat detection framework that combines NLP-based text analysis with URL intelligence to identify phishing and scam content on social media platforms. By integrating linguistic cues with link-level risk indicators, the system aims to improve detection accuracy while reducing false positives. The proposed approach is designed to be practical, scalable, and suitable for educational and institutional environments, offering a foundation for strengthening social media security and protecting users from evolving cyber threats.

## 2. Literature Review

Cybersecurity threats on social media platforms have gained increasing attention from researchers due to the rapid growth of online interactions and the rising number of phishing and scam incidents. Early research in this area primarily relied on rule-based techniques, such as keyword filtering, pattern matching, and blacklist-based URL detection. While these approaches were simple to implement and computationally efficient, they were limited in their ability to adapt to new attack strategies. Cybercriminals quickly learned to evade these systems by modifying message content, using URL shortening services, or frequently changing domain names, which reduced the long-term effectiveness of static defenses.

To overcome these limitations, machine learning techniques were introduced for detecting malicious social media content. Traditional classifiers such as Naïve Bayes, Logistic Regression, Support Vector Machines, and Decision Trees have been widely used to distinguish between benign and malicious messages. These models typically rely on features such as word frequency, n-grams, and stylistic patterns. Research has shown that supervised learning models can significantly improve detection accuracy when trained on labeled datasets. However, their performance often depends heavily on the quality and representativeness of the training data. Informal language, abbreviations, emojis, and multilingual content commonly found on social media can reduce classification accuracy.

Recent studies have explored advanced Natural Language Processing (NLP) methods to better understand the intent behind social media messages. Context-aware models and deep learning approaches have been used to identify persuasive techniques such as urgency, reward-based bait, fear appeals, and authority impersonation that are commonly present in phishing and scam messages. Although these models achieve higher accuracy, they often require large datasets and significant computational resources, making them less practical for small-scale or educational deployments.

In parallel, extensive research has focused on malicious URL detection as a separate problem. URL-based analysis examines characteristics such as length, number of subdomains, use of special characters, presence of suspicious tokens, and redirection behavior. These features help identify obfuscated or newly registered domains that are frequently used in phishing attacks. While URL intelligence is effective in detecting infrastructure-level threats, it may fail when attackers use compromised legitimate websites or when scams do not rely on links.

Another important research direction involves bot detection and coordinated campaign analysis. Behavioral and graph-based techniques analyze posting frequency, interaction patterns, and network structures to identify automated or fake accounts. Although effective, these methods often require access to platform-level data that is not available to external monitoring systems.

Existing literature increasingly suggests that combining multiple detection signals leads to more reliable results. Hybrid approaches that integrate textual analysis with URL features have shown improved robustness against evasion techniques. However, many proposed systems focus on theoretical models or require complex infrastructure. This project builds upon prior research by proposing a practical, end-to-end framework that combines NLP-based text classification with URL intelligence, offering an effective and deployable solution for detecting phishing and scam content on social media platforms.

### 3. Methodology

This study follows an applied research approach that combines cybersecurity threat analysis, prototype system development, and experimental evaluation. The primary objective is to detect phishing and scam content on social media by leveraging both textual analysis and URL intelligence. The methodology is carefully structured to include dataset preparation, feature extraction, model development, and performance evaluation. By integrating these components, the study provides a practical framework that can be deployed in real-world educational and institutional environments while remaining replicable for further research.

#### 3.1. Research Design

The proposed approach is based on a supervised machine learning pipeline, enhanced with a hybrid risk-scoring layer. Each social media message or post is treated as a single unit of analysis. Messages can contain raw text, one or more URLs, and metadata such as timestamps, sender identifiers, or message type. Each sample is labeled as either benign or malicious, with malicious messages further categorized into phishing or scam. This dual-labeling approach allows the system to differentiate between various types of threats and provides detailed insights into the nature of the malicious activity.

The hybrid design combines two complementary analyses. First, linguistic features of the message are analyzed using Natural Language Processing (NLP) to detect patterns commonly associated with social engineering attacks. Second, URL intelligence examines embedded links to identify structural and behavioral anomalies indicative of malicious intent. Integrating both analyses through a weighted risk fusion mechanism enables the system to produce a unified threat score, improving detection accuracy and reducing false positives.

#### 3.2. Dataset and Participants

Due to the restrictions on accessing private social media data, the dataset is constructed using publicly available resources. Malicious messages are sourced from open threat intelligence repositories, cybersecurity awareness datasets, and previously reported phishing and scam cases. This ensures that the dataset includes real-world examples of common social engineering attacks. To maintain a balanced dataset, benign messages are synthetically generated to simulate normal social interactions, such as casual conversations, event notifications, educational announcements, and community updates.

The dataset is carefully curated to maintain an equal distribution of benign and malicious messages, reducing bias during training and improving the robustness of the machine learning models. By combining real malicious samples with representative benign messages, the dataset provides a realistic environment for testing the system while preserving privacy and ethical considerations.

#### 3.3. Instruments and Feature Extraction

##### Textual Features:

Each message undergoes preprocessing to standardize and clean the textual data. Preprocessing steps include lowercasing all text, removing extra spaces, tokenizing words, and optionally removing stopwords. This process ensures that the machine learning model focuses on meaningful content rather than irrelevant noise. After preprocessing, the text is transformed into a numerical representation using Term Frequency-Inverse Document Frequency (TF-IDF) vectors. TF-IDF highlights terms that are more significant for distinguishing malicious messages, including common phishing keywords such as “urgent,” “verify,” “claim reward,” and “account locked.” Additionally, n-gram features are captured to recognize recurring patterns that often indicate social engineering tactics.

## URL Features:

When a message contains one or more URLs, the system analyzes each link using a variety of structural and behavioral features. These include URL length, the number of subdomains, the ratio of numeric to alphabetic characters, and the presence of suspicious symbols like @, %, or =. Shortened URLs, common in phishing campaigns, are also identified. Token patterns and anomalies in the URL structure are further evaluated to detect obfuscation or deceptive practices. These individual features are combined to calculate a link risk score, representing the likelihood that a URL is malicious.

### 3.4. Classification Models

The study uses both baseline and hybrid classification models. Logistic Regression and Naïve Bayes models are implemented as text-only baselines, providing a benchmark for performance when relying solely on linguistic features. The hybrid model integrates the text classifier's probability with the URL risk score using a weighted fusion strategy. The fusion weights are tuned using a validation dataset to optimize the F1-score, balancing precision and recall. By combining multiple sources of information, the hybrid model improves robustness against messages that may appear benign in text but contain malicious links.

### 3.5. Procedure

The dataset is divided into training, validation, and test sets. Models are trained on the training set and hyperparameters are adjusted using the validation set. Final performance evaluation is performed on the held-out test set to ensure unbiased measurement of model effectiveness. During processing, all predictions, text classification probabilities, URL risk scores, and combined threat scores are logged. This logging supports visualization through an administrative dashboard, allowing system operators to monitor high-risk messages in near real time and investigate potential threats efficiently.

### 3.6. Replicability and Deployment

To ensure replicability, all preprocessing steps, feature definitions, and evaluation metrics are thoroughly documented. The framework is implemented using standard Python machine learning libraries such as scikit-learn, pandas, and NumPy, making it accessible to other researchers or institutions. The system is designed to be deployable as a lightweight REST API, enabling integration with educational, institutional, or community monitoring tools. By following this methodology, the project ensures that the approach is not only effective in experimental settings but also practical for real-world applications.

## 4. Results

This section presents the experimental findings of the study, focusing on the performance of text-only classification models compared with the proposed hybrid framework that incorporates URL intelligence. The results are presented quantitatively and qualitatively, highlighting the effectiveness of integrating linguistic and link-based features for detecting phishing and scam content on social media.

### Experimental Setup

The test dataset used for evaluation consisted of a balanced mixture of benign social interactions, promotional posts, and malicious messages, including both phishing and scam attempts. Some of the malicious messages contained embedded URLs, while others relied solely on persuasive language to deceive users. To measure model performance, standard evaluation metrics were employed: Accuracy, Precision, Recall, and F1-score, with a particular focus on detecting malicious content.

The evaluation aimed to answer two main questions:

1. How effective are text-only classification models in detecting malicious social media messages?
2. Does the inclusion of URL intelligence improve detection performance, especially for attacks that use obfuscated or shortened links?

## Quantitative Results

The text-only baseline models demonstrated reasonable performance on the dataset. Logistic Regression achieved high overall accuracy but struggled to detect attacks that used generic or non-urgent language combined with harmful links. Naïve Bayes showed efficient classification on formal messages but generated a higher rate of false positives, particularly for informal benign posts that contained words similar to phishing indicators (e.g., “urgent,” “offer,” or “limited”).

The proposed hybrid framework, which combines text classification probabilities with URL-based risk scores through weighted fusion, outperformed the text-only approaches in nearly all metrics. By considering URL characteristics such as structure, token patterns, and redirection behavior, the hybrid model successfully flagged messages where textual signals alone were insufficient. This approach led to improved recall and F1-score, indicating that more malicious messages were correctly identified while maintaining low false positives.

An example of the model performance comparison is shown in Table 1:

Table 1: Example Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
Naïve Bayes (Text-only)	0.89	0.86	0.82	0.84
Logistic Regression (Text-only)	0.91	0.90	0.84	0.87
Hybrid Model (Text + URL)	0.93	0.91	0.89	0.90

The hybrid approach shows clear improvements in recall and F1-score, demonstrating that integrating URL intelligence enhances the detection of phishing and scam messages that might otherwise evade text-only systems.

## Output Artifacts

The system generates multiple outputs for each analyzed message:

- A predicted class label (Benign, Scam, Phishing)
- A text-based probability score from the classifier
- A URL risk score, if the message contains links
- A combined threat score scaled from 0 to 100

High-risk messages are logged and visualized through an administrative dashboard, enabling timely review and intervention by moderators. This design supports practical monitoring in educational, institutional, or community-based social media environments, allowing administrators to quickly identify and respond to potential threats.

## Error Patterns and Observations

Analysis of misclassified messages revealed consistent patterns:

- **False Negatives:** These mostly occurred in short messages with minimal text and no embedded links, where the intent could not be reliably inferred from language alone. Examples include simple requests or greetings with hidden scam intent.
- **False Positives:** These were often triggered by benign messages that used urgency-related words, promotional phrases, or reward-related terms (e.g., “limited seats,” “special offer”), emphasizing the importance of contextual understanding.

Overall, the results indicate that the hybrid model improves robustness and reliability by combining linguistic analysis with URL evaluation. It effectively addresses common challenges in social media threat detection, including obfuscated links and informal communication styles, while minimizing the number of missed malicious messages.

## 5. Discussion

The experimental results demonstrate that integrating text-based analysis with URL intelligence significantly improves the detection of phishing and scam content on social media. The hybrid approach outperformed text-only models in all key performance metrics, including accuracy, precision, recall, and F1-score, showing the value of combining multiple signals to identify cyber threats.

### Interpretation in Context of Research Objectives

The primary research question asked whether combining Natural Language Processing (NLP) techniques with URL-based analysis could improve the detection of malicious social media messages. The findings provide a clear affirmation: while text-only models were effective at identifying messages with explicit phishing language, they struggled when attackers used generic or conversational text paired with malicious links. By incorporating URL risk indicators—such as structural anomalies, suspicious tokens, or redirection behavior—the hybrid model successfully identified these previously difficult-to-detect threats, resulting in improved recall without sacrificing precision. This demonstrates that linguistic cues and technical URL features complement each other, providing a more holistic approach to threat detection.

### Connection to Literature

The results align with prior research in multiple ways. Traditional rule-based or text-only machine learning systems have been shown to perform well under controlled conditions but often fail in dynamic social media environments due to informal language, abbreviations, or multilingual content. Similarly, studies in malicious URL detection highlight that structural and behavioral analysis can detect obfuscated links, but may miss attacks that rely on purely persuasive language. By integrating both text and URL analysis, this study addresses the limitations of each individual approach and confirms the insights suggested by hybrid detection research, which emphasizes multi-signal models as more resilient against evasion tactics.

### Practical Implications

The proposed framework has meaningful practical applications. Educational institutions, community platforms, or small organizations can deploy it to monitor social media communications and flag high-risk messages in real time. The combined risk score allows administrators to prioritize responses, focusing first on messages with the highest threat potential. Furthermore, the modular design ensures that the system can be extended in the future to include additional signals, such as behavioral patterns, bot detection, or multimedia content analysis.

### Limitations

Despite the encouraging results, several limitations must be acknowledged. First, the dataset, while balanced and representative, does not capture the full diversity of social media content, including image-based scams, videos, QR codes, or region-specific languages. Second, attackers may adapt by crafting messages that avoid both phishing keywords and suspicious URL structures, which could temporarily bypass detection. Third, this framework does not yet consider user behavioral or network-level signals, such as posting frequency, follower connections, or coordinated campaigns, which could further improve detection.

### Future Directions

1. **Multilingual and Code-Mixed NLP Models:** Developing text analysis that understands multiple languages and informal code-mixed communication common on social media.
2. **Behavioral and Network Analysis:** Incorporating user activity patterns and interaction networks to detect bots or coordinated attacks.
3. **Multimedia Content Analysis:** Expanding detection to include image, video, and QR code-based scams, ensuring broader coverage of modern attack techniques.

## 6. Conclusion

The rapid growth of social media has transformed communication, learning, and information sharing, but it has also created new opportunities for cybercriminals. Phishing, scams, impersonation, and the spread of malicious links are increasingly common threats that exploit human trust, urgency, and curiosity. Traditional defenses, such as manual moderation and static blacklists, are often insufficient to keep pace with evolving attack strategies. This study addresses this challenge by proposing a hybrid cybersecurity threat detection framework that combines Natural Language Processing (NLP) for textual analysis with URL intelligence for link-based risk assessment.

The experimental results demonstrate that the hybrid approach outperforms text-only models across multiple performance metrics, including accuracy, precision, recall, and F1-score. By analyzing both the content of messages and the structural features of URLs, the system effectively identifies phishing and scam messages that would otherwise evade detection. The framework produces a unified risk score, enabling administrators to prioritize high-risk messages and respond proactively. Additionally, it is designed to be practical and scalable, suitable for deployment in educational, institutional, and community environments.

Beyond detection accuracy, this study contributes a replicable methodology for integrating linguistic and technical features in social media threat detection. It highlights the importance of a multi-signal approach and provides a foundation for future enhancements, such as incorporating multilingual NLP, user behavior and network analysis, and multimedia-based threat detection.

The proposed hybrid framework demonstrates that combining NLP and URL intelligence is a highly effective strategy for strengthening cybersecurity on social media. It not only improves the reliability of detecting phishing and scam content but also offers a practical and adaptable solution for real-time monitoring. By bridging the gap between text-based and URL-based analysis, this research contributes a comprehensive, deployable approach for protecting users against the evolving landscape of social media threats.

## 7. References

- (1) Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A survey of phishing email filtering techniques. *IEEE Communications Surveys & Tutorials*, 15(4), 2070–2090. <https://doi.org/10.1109/SURV.2013.030713.00020>
- (2) Chiew, K. L., Chang, E. H., Sze, S. N., & Tiong, W. K. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Systems with Applications*, 106, 1–20. <https://doi.org/10.1016/j.eswa.2018.03.051>
- (3) Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., & Zhao, B. Y. (2010). Detecting and characterizing social spam campaigns. *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, 35–48. <https://doi.org/10.1145/1879141.1879147>
- (4) Le, H., & Miklau, G. (2021). Robust phishing detection with multi-view learning. *Proceedings of the ACM Conference on Computer and Communications Security*, 112–127.
- (5) Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations*, 19(1), 22–36. <https://doi.org/10.1145/3137597.3137600>