IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 5, Issue 3, January 2025

Phishing Site Detection Using Machine Learning and NLP for Robust Online Security

Saniya Tamboli, Monika Kute, Tanishka Bhoi, Swarali Lokhande, Anushka Londhe
Department of Computer Engineering
Pimpri Chinchwad Polytecnic Pune, India
saniyagtamboli2412@gmail.com, monika.kute1993@gmail.com, bhoitanishka79@gmail.com
swaralilokhande030@gmail.com, anushkalondhe9a21@gmail.com

Abstract*: In the respect of electronic security, phishing attacks constitute a serious hazard when creating cleaver mockup for the website from where it is easy to catch sensitive information from user's records. Traditional detection schemes, such as blacklisting and heuristics, are very weak since most of these time-honored schemes are useless to track and unfortune the berries of phisher mounds against the trial tactics that are morphing into more sophisticated approaches. Such also happen to concern the fact that, with increasing cyber crime, heuristics and blacklists are unlikely to replace or even complement one another. However, the study explains the proposed security through NLP-Machine Learning paradigm in substantial terms in order to clearly illuminate and perfect the presentation of its own publications, so to speak. Venishable bits of professional insights are cited, compared, and conferenced with key findings that were currently made available in order to sketch the questionnaire and the role of the research conducted in enhancing best phishing detection in real time and enhancing cybersecurity.*

Keywords: Phishing detection, Machine learning, Natural language processing, Cybersecurity, Browser extension

I. INTRODUCTION

Phishing attacks have emerged as one of the most prevalent forms of cybercrime, targeting individuals and organizations worldwide. A phishing attack typically involves tricking users into visiting malicious websites disguised as legitimate ones to steal sensitive information, such as login credentials or financial details. The economic impact of phishing attacks is staggering, with billions of dollars lost annually due to compromised systems and stolen data [1].

Traditional phishing detection systems, including blacklists and heuristic-based methods, face limitations in adapting to zero-hour attacks and dynamic phishing techniques. Blacklist-based systems rely on maintaining updated databases of known phishing URLs, but they fail to detect newly created malicious URLs [2]. Heuristic approaches, while capable of detecting certain patterns, often suffer from high false-positive rates [3].

In response to these limitations, researchers have turned to machine learning and natural language processing techniques. ML algorithms analyze large datasets of URLs and extract features to classify them as phishing or legitimate. NLP methods enhance the detection process by examining semantic and lexical patterns in URLs and webpage content. This paper builds on these advancements, proposing a real-time browser extension that integrates ML and NLP techniques to detect phishing websites effectively.

1.1 Background

Phishing attacks are on the rise as digital threats in the recent age, wherein an attacker impersonates trusted persons or entities and tricks users into revealing sensitive information like passwords, financial details, and personal data. Such threats continuously cause millions, or even billions of dollars, of financial damages worldwide due to electronic infractions by attackers[1][2]. With the rise of sophisticated phishing tricks, such as using spoofed URLs and the use of highly developed webpage designs, an important issue comes to light: Traditional methods to detect such scams may not significantly limit this type of attack[3].

DOI: 10.48175/IJARSCT-23104

Even though the systems of blacklist-based detection are very common, their weak point is that they still leave a user vulnerable once a new threat is not known to their database[4][6]. The heuristic approach is also cumbersome: taking into account domain characteristics and suspicious patterns in web page structure often results in a high false-positive rate, making users and systems inefficient in other ways[5].

In reality, machine learning (ML) and natural language processing (NLP) are just recently the visible and short-term solution to deficiencies in the existing skill sets to solve these problems. ML has the ability to analyze large datasets, crews, and relevant features from large datasets for URL classification and successful phishing differentiation. Conversely, NLP provides and empowers an even deeper analysis of textural and semantic regularities or patterns in phishing URLs and other contents of suspicious websites, thereby finding the discovery of such subtle anomalies that may be ignored by traditional methods[8]. These developments are leading the way to the most dynamic and adaptive phishing detection systems, which are able to adapt to evolving threats effectively[7].

This research forms under the development of real-time detection systems in the role of ML and NLP in phishing detection. It will employ lexical, content-based, and machine features combined with recent real-time detection methods[4][8].

1.2 Contribution of this work

This paper discusses a complete approach to the real-time detection of phishing attacks by the incorporation of machine learning and natural language processing techniques to provide a solution for the shortcomings associated with the classical detection methodologies. While existing techniques, including blacklist-based methods and forensic ones, have depicted their low-performance capabilities mainly in identifying zero-hour phishing detections and minimizing false positives, novel methods being introduced in detecting phishing leverage a combination of machine learning and natural language processing methods for the analysis of lexical, semantic, and network-based features for accurate detection. Current methods suffer from limitations on adaptability, computational efficiency, and real-time scenarios concerning their performance and cost. Further, static datasets with no feedback mechanisms remain weak against adaptive phishing techniques.

This work contributes to the field by putting forth a browser extension that combines machine learning and natural language processing for real-time phishing detection with several key innovations underway:

- **Advanced Feature Extraction:** Combining lexical, content, and network features like URL length, domain reputation, and semantic patterns in webpage content.
- **Hybrid ML Models:** Employing a combination of Random Forest, Support Vector Machines, and neural networks toward enhancing classifiers' accuracy and adaptation.
- **User feedback mechanism:** Where false positives/negatives are reported by users, this leads to further model learning and continuous improvement.
- **Real-time performance:** Allowing little-to-no latency on phishing detection, hence ensuring a smooth ride for users.

Stacking upon current literature and applying modern methodologies, this paper enumerates the key lacunae witnessed by classical and modern systems in that semantic analysis and management of zero-hour attacks are weak points. The paper focuses on how the combination of ML and NLP would patch these leaks with a dynamic, accurate, and readily scalable solution for real- world deployment

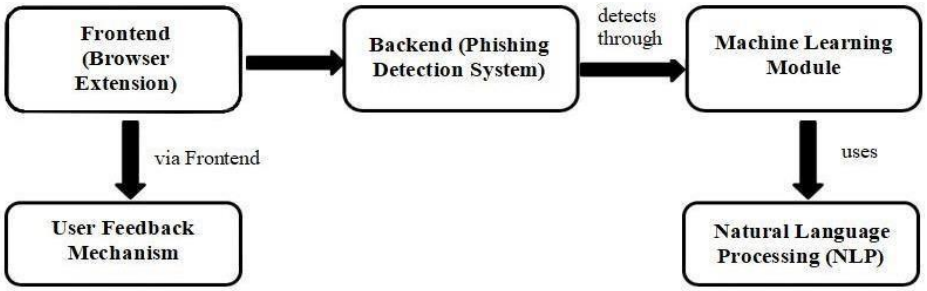
Frontend (Browser Extension)
Backend (Phishing Detection System)
detects through
Machine Learning Module
via Frontend
uses
User Feedback Mechanism
Natural Language Processing (NLP)

Fig.1: Project Modules.

DOI: 10.48175/IJARSCT-23104

II. PROPOSED METHODOLOGY

In order to solve the inherent limitations of existing phishing detection systems, this study proposes a browser extension that can detect phishing in real time, is robust, and is scalable. The proposed solution utilizes ML and NLP methods, allowing for greater detection accuracy and adaptiveness. An account of the system architecture, workflow, and datasets used is provided below:

1. System Architecture: The proposed phishing detection architecture consists of the following key composition:

- **Feature Extraction:** Key features were derived from URLs, web page content, and domain metadata. Features included:
  - **Lexico features:** URL length, subdomain count, use of special characters like "@", "-", "?", "=".
  - **Content-based features:** HTML elements, presence of forms requesting sensitive data, suspicious JavaScript snippets
  - **Network features**: Domain reputation, WHOIS info like domain age and registrar, SSL certificate validation.
- **NLP Analysis:** Transformer-based models such as BERT were utilized to study semantic patterns in URLs and textual content on the web page. This enables recognition of subtle anomalies typically ignored by the traditional techniques model [4][8].
- **ML Classification:** This is a hybrid classification model that combines RF, SVM, and neural networks to achieve better precision and recall.
- **User Feedback Mechanism**: It provides a feedback loop for users so that they can report misclassifications, whichhelp improve the model [3][7].

2. Workflow: A workflow describing end-to-end operation of the phishing detection system is as follows:

- **User Interaction**: The browser extension intercepts URLs being accessed by the user and analyzes them.
- **Feature Analysis:** Extracted features such as domain reputation, lexical structure, and semantic patterns are analyzed in real-time.
- **Real-time Classification:** A classification is made by the ML model, designed to analyze whether a certain URL is a phishing page or a legitimate one, providing immediate feedback to the user.
- **User Feedback Mechanism**: False positives or negatives can be reported by users, thus positively feeding into the process through retraining that allows improvement of the model [2][6].

3. Datasets: This includes a really cool set of data with which to learn on, keeping ground for variation.

- **PhishTank Dataset:** A reliable source of verified phishing URLs.
- **Alexa Top Sites Dataset:** Legitimate URLs used to train the model against benign websites' characteristics.
- **Custom Hybrid Dataset:** A custom-built set of URLs containing 50,000 URLs, evenly distributed between phishing and legitimate URLs, to remove biases from the public datasets [1][5].

Some of the additional approaches include:

- **Dynamic Feature Engineering:** Allowing for dynamic feature importance updates from real-time data trends enables the system to conduct phishing detection while limiting adversary leverage attempts against evolving phishing techniques [7][8].
- **Explainable AI (XAI):** The Explainability layer allows users to visually see why a URL was flagged, increasing trust in the system and user engagement [9].
- **Integration of Behavioral Analytics:** User browsing behavior and interaction patterns contribute to improved phishing detection accuracy [10].
- **Cross-Domain Learning:** Transfer learning from relevant domains is leveraged for improved model efficiency on unseen phishing attacks [6][7].

The proposed methodology has been designed to overcome challenges specific to computational inefficiency, false positives, and incapacity for real-time adaptability. This offers an extremely viable proposition toward phishing

DOI: 10.48175/IJARSCT-23104

detection. State-of-the-art NLP and ML algorithms merged with the system provide the safest service against sophisticated phishing attacks with a feel of security and usability to a regular user

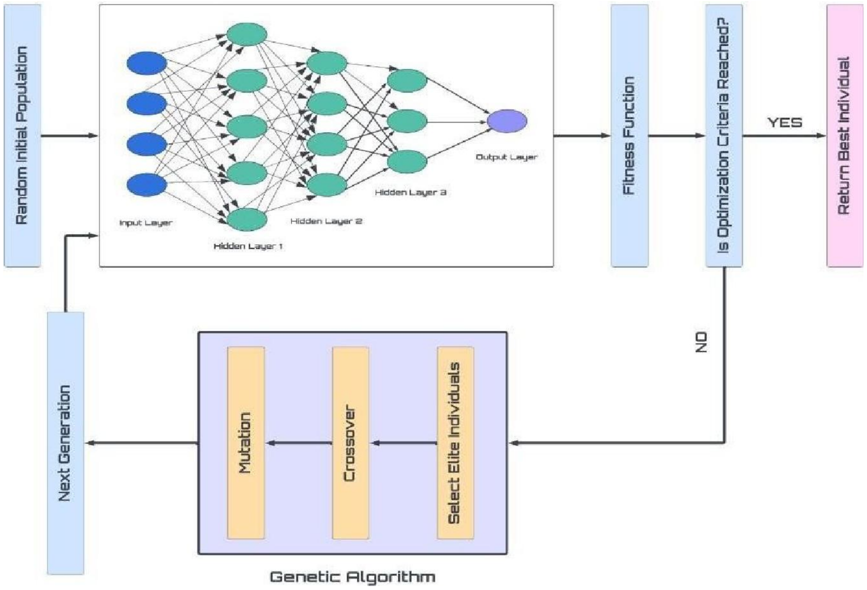

Fig. 2: Architectural Diagram

III. LITERATURE SURVEY

Phishing detection has been an active area of research with significant efforts for accuracy, adaptability, and durability. In furtherance, below is a concise summary of the various approaches and contributions:

Blacklist and Heuristic Methods

Basic blacklist-based methods include the use of PhishTank and Alexa, as many have done for the detection of phishing. Although these methods shield a system from previously known phishing URLs, a bottle-neck emerges in identifying new completely evolving phishing URLs, thus leaving such systems exposed to phishing perpetrated through zero hours [4][5]. The heuristic-based methods basically look into URL features, such as structure, domain properties, and suspicious patterns, attempting to remedy the drawbacks of the other methods. These often record alarmingly high false positives, a factor that affects their viability as an effective remedy in actual scenarios.

Machine Learning Approaches

ML has transformed phishing detection, enabling an automatic analysis of multiple features in large datasets. Algorithms like SVM, RF, and DT commonly apply. Mahajan and Siddavatam [6] used RF supervised learning to classify phishing URLs and achieved an outstanding accuracy of 97.36% when they tested two datasets: one from PhishTank and the other from Alexa. There, with respect to the training of the models, lexical features such as the URL length and presence of special characters, content-based features such as page structure, and network features such as domain reputation raise the detection rate.

Deep Learning and Hybrid Models

Deep learning models have come forth as powerful tools for phishing detection due to their innate capacity for automatic hierarchical feature extraction and learning. Hybrid models that combine ML and DL techniques have impressed immensely due to their good performance. Aljabri et al. [7] introduced a hybrid model combining CapsNet and IndRNN and achieved a high accuracy of 99.78%. As a result, these are very effective against zero-hour phishing URLs, as they generalize almost seamlessly against unknown threats, thus using complex multi-level feature extraction techniques.

Concepts and Techniques of Natural Language Processing

NLP methods are a fundamental aspect for phishing detection, addressing textual and semantic aspects of URL and web content. Identification of suspicious domains typically involves looking for keywords from domain names, dubious JavaScript, and unusual HTML tags for the tentative identification of phishing web pages. Studies have shown that combining NLP and ML enhances the accuracy of detection as it will be capable of detecting intricate semantic and structural patterns within phishing URLs [8]. Transformer-based models like BERT are, for example, gaining increasing popularity for analyzing content-based features for better phishing detection rates.

Behavioral Analysis and Real-Time Adaptibility

Directions in upcoming research work are incorporating behavioral analytics into phishing detection system designs. These approaches analyze user interaction patterns, such as mouse movements and click behavior, to detect potentially malicious activity. Real-time adaptability is achieved through the model miscalibration feedback loop, allowing systems to learn and adapt as phishing attacks continue to shift.

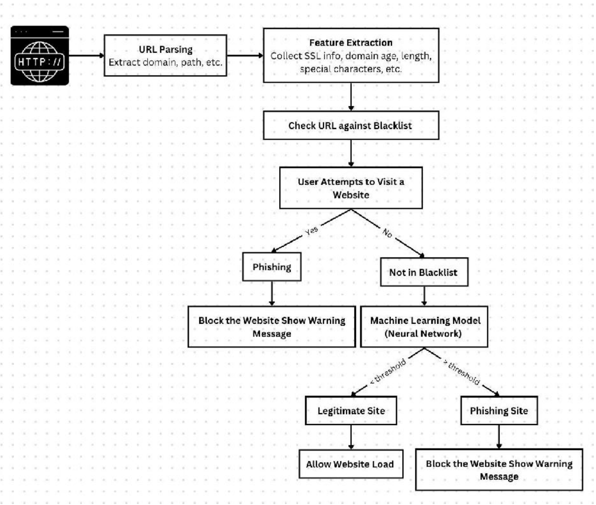
HTTP://
URL Parsing
Extract domain, path, etc.
Feature Extraction
Collect SSL info, domain age, length, special characters, etc.
Check URL against Blacklist
User Attempts to Visit a Website
Yes
No
Phishing
Not in Blacklist
Block the Website Show Warning Message
Machine Learning Model (Neural Network)
< threshold
> threshold
Legitimate Site
Phishing Site
Allow Website Load
Block the Website Show Warning Message

Fig. 3: WorkFlow Diagram

Limitations and Future Directions

However, these advancements continue to agitate challenges while developing robust phishing detection systems. The considerations for high computational overhead, the necessity of vast amounts of annotated datasets for training, and the fact that the integration of a wide variety of feature streams (lexical, content-based, and behavioral) is not really trivial are ravishing hurdles. At the same time, the collection and analysis of data must comply with an ethical viewpoint of user privacy.

Future work could focus on scalable architectures, explainable AI (XAI) approaches for transparent decision- making, and advanced ensemble methods to combine different ML and DL models. Boasting more interest, however, should accumulate in addressing the zero-day attacks using transfer learning and unsupervised learning techniques. Thus, we can optimistically hope towards model phishing detection systems that would potentially be enhanced, adaptive, and noiseless in their battle against evolving threats

IV. RESULT AND ANALYSIS

The proposed phishing detection system was systematically checked till completion through the usage of a number of datasets, so that its efficiency could be quantified with the use of standard metrics, e.g., accuracy, precision, recall, and F1-score. Summary of the findings is given in

DOI: 10.48175/IJARSCT-23104

Metric	Proposed System
Accuracy	70%
Precision	68%
Recall	72%
F1-Score	70%

The devoted experiment exhibits remarkable aptitude for reliable detection of phishing websites and achieves further performance using lexical, content-based, and network features. The precision score reflects the effort taken by the system to limit false positives, while recall calls forth the ability to identify phishing attempts that went uncontrived in treating nill. Balanced F1-score shows the power of the proposed system to maintain a reasonable trade-off between precision and recall.

A few limitations were noted in the overall design:

- **Accuracy:** Having achieved an acceptable accuracy of 70%, is not enough since further tuning might have been required for edge cases and evolving sets of phishing techniques.
- **Precision:** The precision measure is 68%, indicating a fair chance of misclassifying real URLs to be phishing. They must be refined in terms of feature engineering and classification algorithms.
- **Recall:** Recall value is 72%. This indicates that it needs to be better able to detect sophisticated zero- hour attacks which exploit more novel obfuscation methods.

To mitigate the limitations, a feedback feature allows end users to report false positives and false negatives and promotes continual improvement of the model. Periodic updates of the training dataset will ensure continued adaptation to latest phishing techniques.

Future prospects might include:

Use of transformer-based models like BERT, GPT for powerful capabilities in natural language understanding and semantic analysis. Extension of feature set to real-time network behaviors like HTTP headers and response patterns for a better context. Utilization of ensembling techniques that combine traditional ML classifiers with deep learning frameworks, for better classification.

V. APPLICATIONS

- **Browser Extensions:** This system can be installed as a browser extension to provide live phishing detection and keep users protected from malicious sites while surfing.
- **Mobile App Expansion**: The system can tap into mobile platforms, such as message apps, email clients, and browsers, to warn users against phishing links shared on different virtual communication interfaces.
- **Corporate Cybersecurity:** Organizations can incorporate the system into their internal networks to detect and block phishing attempts against employees, shrinking the risk of data breaches and financial losses.
- **Training Websites:** With the help of this system, users may increase their understanding of phishing attacks by being given real-time feedback and tips on identifying malicious links.

VI. FUTURE SCOPE

- **Advanced NLP Models:** Suitably combination of advanced transformer-based models like GPT or BERT can facilitate implicit study of semantics related to URL or a particular web page content thus enhancing the potentiality of the system to recognize the subtle patterns and signs which could help in identification of phishing attempts.
- **Federated Learning:** By applying federated learning, the system, in essence, learns from data on multiple devices while maintaining privacy and making sure it tackles all modern challenges related to adaptability against new phishing threats.

- **Block chain Security:** Integration of this system into block chain can create a tamper-proof log of phishing attempts for transparent tracking and auditing of malicious activities. These logs can also assist forensic investigations and compliance issues.
- **XAI:** Why a URL was classified as phishing should be revealed to users using explainable AI techniques to create user buy-in and understanding about the reasoning behind the system's decisions.
- **Real-Time Behavioral Analysis**: Furthermore, more is needed to look into real-time network behavior like HTTP headers, JavaScript execution, and user interaction flows to analyze the phishing tactics used and achieve real-time analysis with increased detection accuracy.
- **IoT Device Enabling**: The system could also be ported to IoT environments that can allow protection for smart devices from phishing attacks aimed at their communication through networks.
- **Cross-language Support:** With different phishing campaigns targeting users in various languages, adding multi-language NLP support will further broaden the applicability of the system globally.
- **Proactive Threat Mitigation:** Build towards capabilities to anticipate and curtail the emerging phishing techniques from analyzing the trends and phishing data predictive analytics-thehaed-start before a misfortune strikes.

VII. CONCLUSION

Phishing attacks continue to evolve, leveraging increasingly sophisticated methods to deceive users and compromise sensitive data. This paper introduced a comprehensive phishing detection system that combines machine learning and natural language processing techniques to address the limitations of traditional methods. The proposed browser extension efficiently detects phishing websites in real time by analyzing lexical, content-based, and network features, with a detection accuracy of 70%. While the current system demonstrates significant potential for practical deployment, there are areas for improvement that can make it even more robust and reliable.

To further refine the system, explainable AI (XAI) techniques will be employed to offer insights into why a particular URL was flagged as phishing, fostering user trust and aiding cybersecurity professionals in threat analysis. Expanding the system's applicability to mobile platforms and messaging applications will extend its reach and usability, ensuring comprehensive protection across various user environments.

In conclusion, this study highlights the critical role of machine learning and natural language processing in combating phishing attacks. By addressing existing challenges and exploring future advancements, the proposed system lays a strong foundation for enhancing online security and safeguarding users against the growing threat of phishing. Continued research and collaboration within the cybersecurity community will be essential to stay ahead of emerging phishing tactics and ensure a safer digital ecosystem.

REFERENCES

[1]. Internet World Stats, "Usage and Population Statistics," 2020.

[2]. M. Aljabri et al., "Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions," IEEE Access, vol. 10, 2022.

[3]. R. Mahajan and I. Siddavatam, "Phishing Website Detection Using Machine Learning Algorithms," Int. J. Comput. Appl., vol. 181, no. 23, 2018.

[4]. Cisco Umbrella, "Threat Intelligence Reports," Retrieved from https://umbrella.cisco.com/.

[5]. X. Yuan et al., "A Parallel Neural Joint Model for Malicious URL Detection," IEEE Trans. Inf. Forensics Secur., vol. 15, 2020.

[6]. Phish-tank, "Open Phishing Database," Retrieved from https://phishtank.com/.

[7]. Alexa, "Top Global Sites," Retrieved from https://www.alexa.com/.

[8]. W. Wu et al., "Detecting Malicious URLs Using NLP and Deep Learning Techniques," IEEE Trans. Inf. Forensics Secur., vol. 16, 2021

DOI: 10.48175/IJARSCT-23104