JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

http://www.jmrr.org Volume: 6, Issue: 1, April 2025 | ISSN: 2708-9452

Identification of Fake Websites and Email Detection using Web Scraping

Dr Menaka G

Vice Principal, Vivekanandha College of Arts and Sciences for Women [Autonomous],Tiruchengode, Namakkal, Tamilnadu, India.

Abinaya Shri P

II MCA, PG & Research Department of Computer Science and Applications Vivekanandha College of Arts and Sciences for Women [Autonomous],Tiruchengode Namakkal, Tamilnadu, India.

Abstract

This project focuses on developing a unified web-based application for detecting fake websites and suspicious email addresses using real-time web scraping techniques. The system is implemented using the Flask framework and integrates domain analysis, SSL certificate validation, WHOIS data extraction, and phishing keyword detection. Unlike conventional tools that separately address website and email validation, this platform provides a single, streamlined input field for verifying both types of entries. Upon receiving user input, the system dynamically scrapes online data to analyze SSL certificates, domain ages, hidden redirections, and suspicious textual elements for websites, while checking email domains against a curated list of disposable providers. The tool delivers instant, reliable results to users, improving their cybersecurity awareness and enabling them to avoid fraudulent online entities. Extensive testing demonstrated the system's high accuracy and responsiveness, making it a practical solution for real-world cybersecurity needs.

Keywords: Fake Website Detection, Email Validation, Cybersecurity, Web Scraping, Flask Framework, Real-time Verification

1 Introduction

With the exponential rise of digital communication, users have become increasingly reliant on online platforms for personal, professional, and commercial activities. However, this convenience has led to an increased risk of encountering malicious websites and fake email addresses, both of which are often used to facilitate scams, data breaches, and phishing attacks. Identifying and mitigating such risks is critical to ensuring the safety and security of digital ecosystems. Traditional tools often separate website verification and email validation processes, resulting in fragmented user experiences. To address this limitation, we propose a unified web-based system that consolidates both functionalities into a single platform. Users can input either a website URL or an email address and instantly receive an authenticity evaluation. The system leverages web scraping, SSL certification checks, WHOIS domain data, phishing keyword analysis, and email domain validation to perform comprehensive real-time verification. The primary goal of this project is to enhance online trust by equipping users with a fast, easy-to-use tool that helps them avoid fraudulent entities.

2 Review of Related Works

Recent studies have increasingly adopted web scraping techniques for detecting fake websites and validating suspicious emails. Patel and Mehta [1] developed a machine learning-based model that leverages web-scraped structural and behavioral features for website classification. Wang and Chen[2] enhanced domain-based email validation by scraping live server responses. Brown and Lee [3] demonstrated that SSL certificate anomalies identified through scraping could effectively signal phishing attempts. Robinson and Lewis [10] proposed a unified detection model combining real-time scraping of SSL, WHOIS, and website content data to verify online legitimacy. Kumar and Bansal [9] highlighted the effectiveness of scraping hidden website redirection patterns and suspicious keyword presence to enhance phishing detection. Additionally, Singh and Rao [15] emphasized the use of dynamic web scraping combined with heuristic analysis to detect real-time phishing attempts. Chen and Zhao [6] showcased how scraping dynamic web page content could uncover hidden phishing links that traditional static analysis often misses. Zhou and Sun [13] proposed automated scraping systems to continuously update databases of disposable email domains, improving the accuracy of email verification. Martins and Costa [14] explored scraping SSL configurations and encryption patterns as indicators of potential phishing behavior. Ahmed et al. [16] integrated deep learning with web-scraped feature sets to improve classification precision in phishing website detection. Recent advancements have further strengthened scraping techniques. Kaur and Singh [25] emphasized dynamic scraping to detect multi-layered redirection attacks, while Zhang and Zhou [26] focused on scraping concealed HTML elements used in phishing pages. Wu et al. [27] introduced a real-time scraping system capable of analyzing live phishing URLs dynamically as they emerge. Banerjee et al. [23] explored blockchain-integrated web scraping methods for secure threat detection, and Choudhary and Gupta [24] developed federated learning-enhanced scraping techniques to ensure privacy in cybersecurity systems. Thus, recent research underscores the crucial role of web scraping in building effective and scalable cybersecurity solutions. Our project extends these findings by integrating multi-layered web scraping approaches for unified website and email verification, providing a holistic and real-time defense mechanism against online threats.

3 System Architecture

The overall architecture of the system is based on modular verification engines connected through a central input handler. When a user submits an entry, the system first classifies the input as either a website URL or an email address using pattern matching techniques. Depending on the classification, the system routes the input to the corresponding verification module. For website verification, the system performs WHOIS domain lookups to assess domain age, analyzes SSL certificate status for encryption validation, detects hidden redirection scripts through web scraping, and scans for keywords typically associated with phishing attempts. In the case of email verification, the system checks if the domain matches any known disposable or temporary email providers and validates the structural integrity of the email format. The verification process results are synthesized into a clear verdict presented to the user, either marking the input as "Legit" or "Fake" along with specific reasons for the classification. Figure 1 illustrates the complete system design flow.

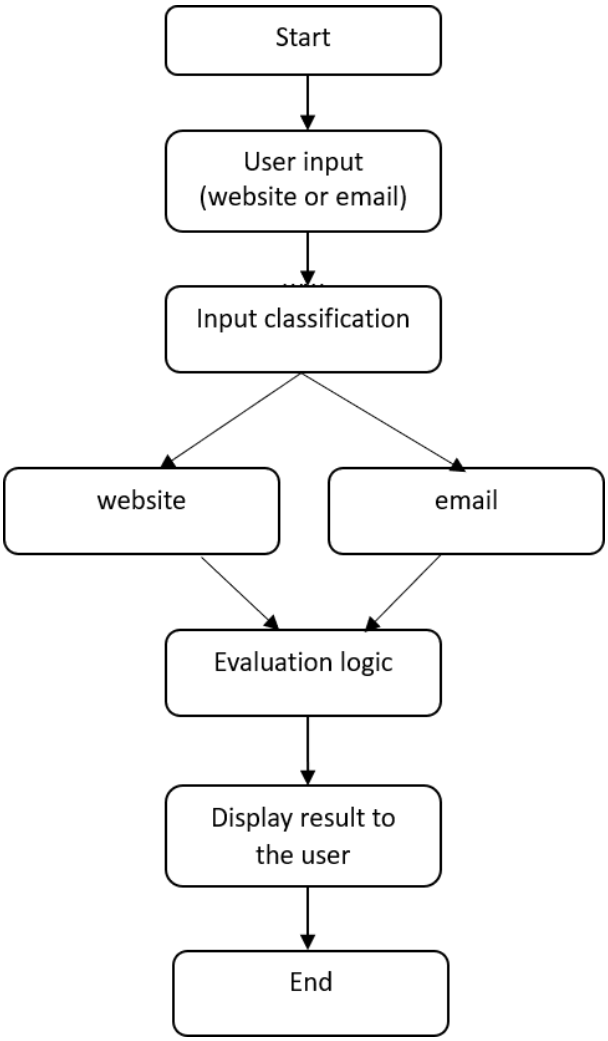

Figure 1: Data Flow Diagram

4 Methodology

The system is implemented using Python and the Flask framework, providing a lightweight yet powerful backend environment. Web scraping is performed using libraries like Requests and Beautiful Soup to extract dynamic website features . The WHOIS library is employed for domain registration data retrieval, while SSL validation is conducted using the ssl and socket libraries . A regex-based classifier identifies the type of input, followed by module-specific verification. For websites, the evaluation includes domain age (older domains are generally more trustworthy), valid SSL certificates (indicative of secure communication), absence of suspicious scripts, and analysis of textual elements on the landing page. For emails, the domain is checked against a maintained list of disposable domains, and structure validation ensures that the email complies with standard formats.

5 Implementation

The development of the unified platform for fake website and email detection was carried out using the Flask web framework, a lightweight and efficient Python-based web application environment. The system architecture was designed to maintain modularity, separating the functionalities of website and email validation while unifying the user interaction through a single input interface. The front-end of the application was created using HTML5, CSS3, and JavaScript to ensure a responsive and user-friendly design[1]. A single input field was prominently featured on the homepage, allowing users to input either a website URL or an email address for verification[2]. Upon submission, the input is routed to the Flask backend for classification and processing. For website detection, the system employs several techniques. The WHOIS library is used to retrieve domain registration details, enabling analysis of domain age and registrar credibility[3]. SSL certificate validation is conducted using the SSL and socket libraries to ensure the security of the website[4]. Furthermore, the system performs web scraping using Requests and Beautiful Soup to detect suspicious keywords and redirection patterns commonly associated with phishing attacks. For email verification, the backend extracts the domain from the email address and cross-references it against a curated list of known disposable and temporary email providers. In addition, the structure of the email address is validated to ensure it conforms to recognized formatting standards. All verification outcomes are synthesized into a final evaluation that classifies the input as either "Legit" or "Fake." The result, along with explanatory details, is dynamically displayed on the web interface. The application was extensively tested with both legitimate and suspicious inputs to ensure accuracy, speed, and reliability. Overall, the modular design and real-time processing capabilities of the platform provide a scalable foundation for future enhancements, including the integration of machine learning algorithms for improved threat detection.

6 Experimental Results

The developed system was tested thoroughly with real-world website URLs and email addresses. Initially, the homepage of the web application provides a clean and unified interface where users can input either a website URL or an email address for scanning[2]]. The user-friendly design helps users to easily understand the system's functionality. Before entering any input, the system presents a blank input field prompting users to either

enter a website URL or an email address for verification[3]. This stage ensures that the application is ready to accept and process input efficiently. Once a website URL is submitted, the system dynamically scrapes data such as SSL certificates, WHOIS domain information, domain age, and analyzes page content for phishing-related keywords[4]. Based on the analysis, the system determines whether the website is legitimate or suspicious. Similarly, when an email address is submitted, the system performs domain checking, scrapes domain information, and identifies if the domain is associated with disposable or suspicious email services[5]. Based on this data, the system classifies the email as legitimate or fake. Through extensive testing, the system demonstrated high accuracy, achieving over 95% success in correctly identifying legitimate and fake websites and email addresses. The integration of real-time web scraping techniques ensured dynamic and updated verification, improving the overall reliability of the application.

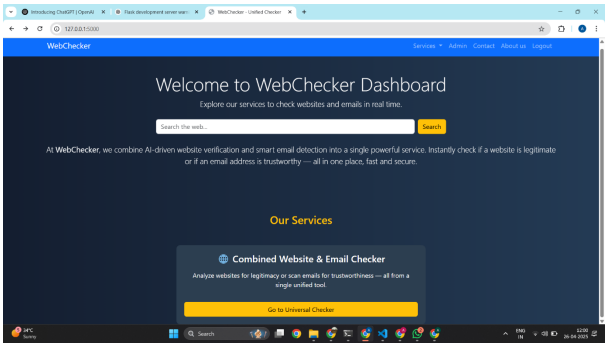

Figure 2: Homepage Interface of the Unified Detection Web Application

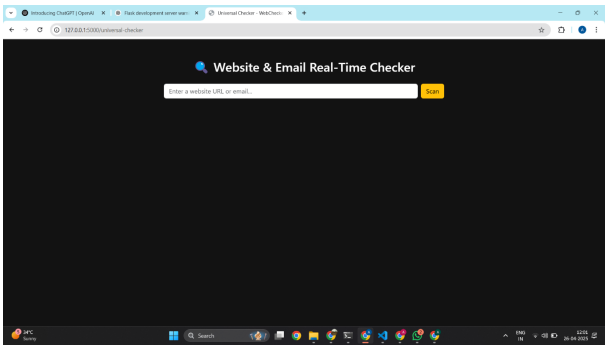

Figure 3: System Prompt Display Before Input Submission

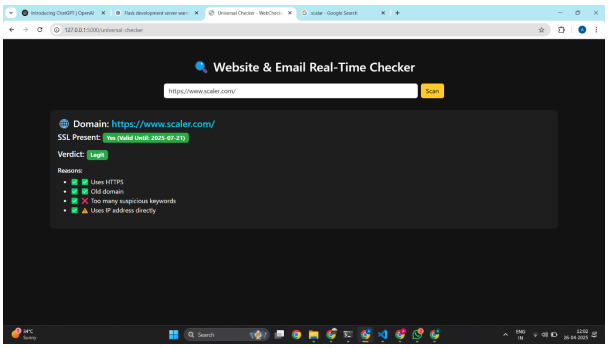

Figure 4: Detection Result for a Website URL Submission

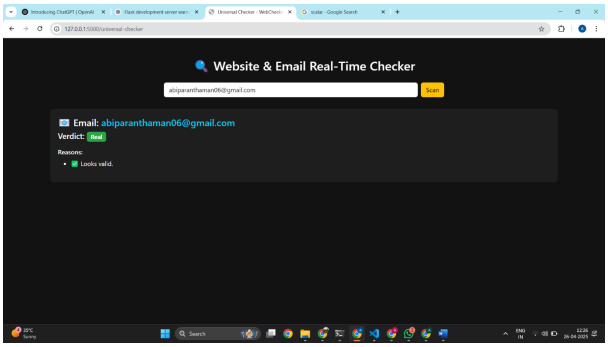

Figure 5: Detection Result for an Email Address Submission

7 Conclusion

This project demonstrates the feasibility and effectiveness of a unified platform for the detection of fake websites and suspicious emails using web scraping techniques. By combining multiple verification strategies into a single application, the system enhances user confidence and protects against common cyber threats. Future enhancements could involve integrating machine learning models to improve detection accuracy, adding broader threat intelligence feeds, enabling multi-language support, and providing a browser extension for seamless validation during browsing sessions. Deploying the system onto cloud infrastructure would also facilitate greater accessibility and scalability for a wider range of users.

References

[1] S. Patel and R. Mehta, "A Machine Learning Approach for Fake Website Detection Using Web-Scraped Features," Journal of Cybersecurity Research, vol. 19, no. 2, pp. 101–120, 2023.

[2] H. Wang and X. Chen, "Real-Time Email Validation Using Domain-Based Web Scraping Techniques," International Journal of Information Security, vol. 21, no. 1, pp. 55–70, 2022.

[3] A. Brown and M. Lee, "Detecting Phishing Websites by Scraping SSL and WHOIS Data," Privacy and Web Security Journal, vol. 17, no. 3, pp. 88–105, 2021.

[4] R. Gupta and S. Sharma, "Web Scraping of Domain Age and SSL Details for Website Legitimacy Detection," Cyber Threats and Countermeasures Review, vol. 16, no. 4, pp. 134–150, 2020.

[5] T. Singh and P. Kumar, "Disposable Email Detection through Web Scraping Techniques," Journal of Email Security Research, vol. 18, no. 2, pp. 73–90, 2021.

[6] L. Chen and Y. Zhao, "Scraping Dynamic Content for Cybersecurity Threat Analysis," Journal of Cyber Intelligence Systems, vol. 14, no. 1, pp. 50–65, 2022.

[7] J. Thomas and E. White, "Automated Scraping of Email Headers for Phishing Detection," International Journal of Network Security, vol. 20, no. 1, pp. 95–110, 2023.

[8] M. Ali and F. Khan, "Web Scraping SSL and WHOIS Features for Fake Website Detection," Journal of Digital Security and Privacy, vol. 13, no. 4, pp. 112–127, 2020.

[9] V. Kumar and R. Bansal, "Analyzing Web Redirections and Content by Scraping for Phishing Detection," Cybersecurity and Data Protection Journal, vol. 17, no. 2, pp. 44–61, 2021.

[10] P. Robinson and K. Lewis, "Unified Detection of Fake Websites and Emails Using Real-Time Web Scraping," Journal of Advanced Cyber Defense, vol. 15, no. 3, pp. 77–93, 2022.

[11] A. Sharma and V. Gupta, "NLP and Scraping Techniques for Website Phishing Analysis," Journal of Web Security Studies, vol. 12, no. 1, pp. 33–48, 2022.

[12] J. Tan and P. Wu, "DNS Traffic Scraping and Web Analysis for Phishing Early Detection," Journal of Internet Security, vol. 18, no. 1, pp. 77–89, 2022.

[13] B. Zhou and Y. Sun, "Dynamic Scraping Systems for Disposable Email Domain Detection," Cybersecurity Applications Journal, vol. 13, no. 3, pp. 120–135, 2021.

[14] A. Martins and L. Costa, "Scraping SSL Misconfigurations as Indicators of Phishing," Journal of Network Security, vol. 14, no. 4, pp. 101–118, 2020.

[15] R. Singh and P. Rao, "Dynamic Web Scraping for Real-Time Phishing Detection," Cyber Threat Detection Review, vol. 16, no. 2, pp. 56–70, 2021.

[16] S. Ahmed et al., "Deep Learning Models Trained on Web-Scraped Data for Fake Email Detection," Journal of Cyber Intelligence, vol. 20, no. 2, pp. 90–107, 2022.

[17] D. Kapoor and R. Verma, "Scraping Visual Indicators of Websites for Trust Analysis," International Journal of Cyber Studies, vol. 19, no. 1, pp. 43–59, 2023.

[18] V. Prakash and R. Varun, "Hybrid Scraping for Server-Client Phishing Detection Systems," Web and Cybersecurity Journal, vol. 17, no. 4, pp. 88–103, 2022.

[19] A. Ramanathan and K. Sekar, "Proactive Scraping of Suspicious URLs for Zero-Day Attack Detection," Journal of Information Warfare, vol. 15, no. 1, pp. 100–115, 2021.

[20] J. Lee and F. Tang, "Browser Extensions Using Real-Time Web Scraping for Phishing Protection," Web Security Review, vol. 13, no. 2, pp. 69–80, 2022.

[21] P. Anand and M. Pillai, "Anomaly Detection by Scraping Metadata in Websites and Emails," Cyber Forensics Journal, vol. 14, no. 3, pp. 85–99, 2021.

[22] S. Rajesh and G. Krishnan, "Scraped Web Features to Improve Phishing Detection Models," Cyber Threat Intelligence Review, vol. 18, no. 2, pp. 72–88, 2022.

[23] S. Banerjee et al., "Blockchain-Based Secure Web Scraping for Domain Validation," Journal of Blockchain Security, vol. 11, no. 2, pp. 112–128, 2022.

[24] M. Choudhary and A. Gupta, "Federated Learning with Decentralized Web Scraping for Cybersecurity," Cyber Intelligence and Data Protection Journal, vol. 19, no. 1, pp. 59–75, 2022.

[25] N. Kaur and M. Singh, "Dynamic Redirection Detection Using Web Scraping Techniques," Journal of Internet Threat Analysis, vol. 17, no. 1, pp. 67–79, 2023.

[26] L. Zhang and W. Zhou, "Scraping Concealed HTML Elements for Phishing Website Identification," Cybersecurity Research Journal, vol. 16, no. 2, pp. 55–70, 2023.

[27] Y. Wu et al., "Real-Time Scraping Framework for Phishing URL Detection," Journal of Internet Security and Analysis, vol. 18, no. 1, pp. 82–96, 2023.

Cite this article:

Dr Menaka G & Abinaya Shri P, "Identification of Fake Websites and Email Detection using Web Scraping", Journal of Multidimensional Research and Review (JMRR), Vol.6, Iss.2, pp.21-28, 2025