A Multi-Faceted Approach to Fraudulent Website Detection: Integrating URL Features and Content Analysis

Mihir Ranjan (240347007009)

National Forensic Sciences University Goa, India

Abstract. The proliferation of online platforms has made detecting fraudulent websites a critical challenge. Initially, our approach focused on identifying fraudulent websites based on URL features. However, as cybercriminals adopted techniques to craft legitimate-looking URLs, we extended our system to analyze website content. Our enhanced methodology involves scraping webpage content and performing sentiment analysis using Twitter RoBERTa to assess intent and detect anomalies. Additionally, we evaluate the confidence of organizational names referenced on the website to determine their relevance and legitimacy. Spelling checks are also implemented to identify subtle yet consistent errors that often signal fraud. This multi-faceted approach strengthens the detection of deceptive websites, addressing both URL-based and content-based fraud vectors, and offers a comprehensive solution for mitigating online threats.

Keywords: Fraud Detection · URL Features · Web Scraping · Beautifulsoup · Sentiment Analysis · Organization · Machine Learning · NLP

1 Introduction

Fraudulent websites remain a significant cybersecurity issue, targeting unsuspecting users by imitating trustworthy platforms to steal private data. This study integrates cutting-edge analytical tools and machine learning techniques to create a unified system capable of detecting fraudulent URLs. The framework evaluates the security of a given URL through a variety of factors, including sentiment analysis, organizational entity recognition, website scraping, spelling checks, and a machine learning-based fraud detection classifier.

To predict the likelihood of fraud, the system uses a Gradient Boosting Classifier (GBC) trained on a dataset of URL features. The workflow includes evaluating spelling accuracy, processing text for sentiment and organizational patterns, and scraping website data. These analyses are combined into a final judgment using a weighted scoring system, offering a reliable and interpretable fraud detection method.

The goal of this project is to develop a comprehensive system that examines various URL features to accurately identify fraudulent websites. The system employs URL detection through feature extraction, sentiment analysis, organizational entity recognition, spelling checks, and website scraping. The approach

forecasts the likelihood of a website being fraudulent by integrating feature-based analysis with predictive modeling using the Gradient Boosting Classifier. This method provides a scalable and effective solution to combat cybersecurity threats, offering high accuracy and insightful information on potential warning signs.

2 Literature Review

The increasing integration of real-world operations into the cyber realm has brought significant convenience to daily life but also introduced serious security risks, such as phishing attacks. These attacks exploit users' trust by mimicking legitimate websites to steal sensitive information. Traditional defenses like antivirus software and firewalls are often ineffective against sophisticated phishing tactics, leading to a shift towards advanced detection mechanisms. Machine learning (ML)-based solutions have gained prominence due to their adaptability and dynamic nature, especially in combating "zero-day" attacks. Studies using multiple algorithms for URL analysis have demonstrated impressive success rates, making ML a critical tool in phishing detection [1].

Malicious URLs, commonly used in phishing, spam, and malware distribution, are a growing threat to cybersecurity. While blacklists have historically been used for detection, their inability to address newly generated threats highlights the need for adaptive ML solutions [2]. Research in this area emphasizes feature engineering, algorithm design, and practical applications, providing valuable insights for academics and industry practitioners alike. Techniques such as the PhishWHO system, which integrates identity-based analysis and URL feature extraction using n-grams, have achieved significant accuracy improvements over conventional methods [3]. These approaches underscore the importance of advanced methodologies in the fight against phishing and malicious URLs.

Despite these advances, challenges remain, particularly in ensuring the robustness of ML-based systems. Adversarial attacks have exposed vulnerabilities, with studies showing significant performance drops when tested against crafted adversarial URLs. For instance, the Matthew Correlation Coefficient of some state-of-the-art ML systems declined drastically when faced with such threats, revealing their unreliability in current forms [4]. These findings highlight the critical need for designing secure and dependable detection systems capable of withstanding manipulative threats. Future research must focus on addressing these vulnerabilities and enhancing system reliability to ensure the long-term effectiveness of ML-based solutions in protecting users and organizations from evolving cyber threats.

3 Methodology

This study adopts a multi-layered approach that combines content-based insights with URL analysis to tackle the problem of identifying fraudulent websites. The methodology integrates various analytical tools and machine learning algorithms

into a single system to ensure reliable and accurate identification. The key components of the methodology are as follows:

Fig. 1. Workflow of the URL detection system

3.1 URL Detection Model

The URL detection model is trained using a dataset of safe and fraudulent URLs, applying classifiers such as Logistic Regression, k-Nearest Neighbors, Support Vector Classifier, Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, and Multilayer Perceptrons. The best-performing classifier is selected to predict whether a URL is safe or risky. For prediction, 30 features are extracted

from the provided URL, including properties such as domain length, HTTPS usage, and the age of the domain. The chosen model processes these features to classify the URL and determine its safety.

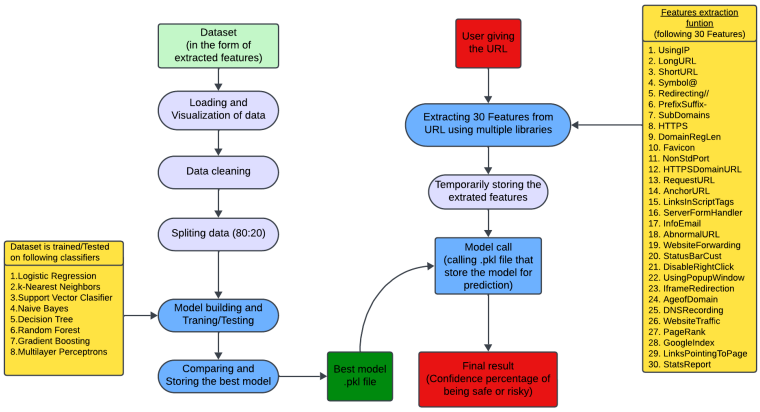

Fig. 2. Workflow of the URL detection model

3.2 Sentiment Analysis

Sentiment analysis is performed on website content using a pre-trained model specifically designed for sentiment classification. We load a sentiment analysis pipeline using the "cardiffnlp/twitter-roberta-base-sentiment" model, which classifies text into three sentiment categories: negative, neutral, and positive. The model analyzes the text in segments of up to 512 characters for efficiency, outputting the sentiment label along with the associated confidence score. This process provides a comprehensive understanding of the sentiment conveyed in the text, which is essential for sentiment-based studies or applications.

3.3 Organizational Entity Recognition

Our model uses a transformer-based Named Entity Recognition (NER) model to identify and analyze organization entities in a text. It processes the content of website by splitting it into sentences, applying the NER model to each sentence, and extracting organization-related entities along with their confidence scores. The program calculates the average confidence of identified organizations to assess the model's accuracy in detecting these entities across the text.

3.4 Spelling Checker

Our model processes a content of the website to identify and correct spelling errors. It extracts valid words, excluding symbols and numbers, and checks for

misspelled words using a spell checker. The program calculates the percentage of correctly spelled words and outputs this accuracy.

3.5 Computation of the Final Weight Matrix

To compute the final weight matrix, our system combines insights from both URL features and website content. Initially, the URL is processed to extract 30 key features, which are sent to a URL detection model contributing 30% to the overall weight matrix. Simultaneously, the content of the webpage is scraped and analyzed through multiple layers. Sentiment analysis, performed using advanced models like Twitter RoBERTa, accounts for 30% of the total weight. Additionally, organizational name recognition contributes 20%, where the legitimacy and confidence of the identified names are evaluated. Finally, a spelling checker is applied to detect subtle errors indicative of fraud, contributing the remaining 20%. Together, these components form a robust and balanced weight matrix, enabling precise and comprehensive fraud detection.

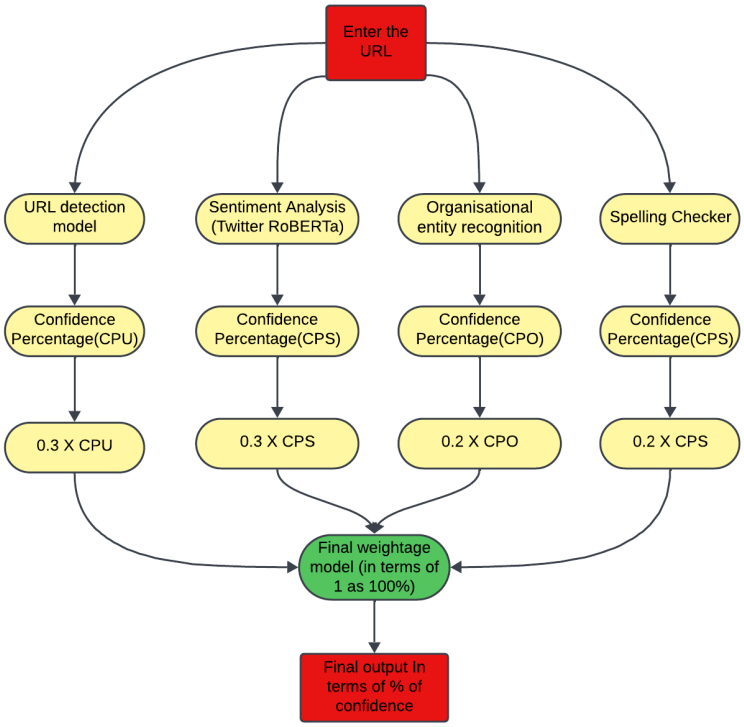

Fig. 3. Workflow Diagram for Final Weight Matrix Computation in Fraud Detection

4 Results and Discussion

The performance of each component and its role in the ultimate decision-making process are highlighted in this section as we present and examine the findings of our fraud detection system. A weighted score system is used to assess the efficacy of the multi-layered technique, which incorporates organisational entity recognition, sentiment analysis, URL detection, and spelling checks.

4.1 URL Detection Model

In order to predict URL authenticity, the URL Detection Model analyses 30 characteristics, including domain length, HTTPS usage, and subdomain count, using a Gradient Boosting Classifier (GBC). The model obtained a high detection rate of 94.9% in this experiment. Its predictions account for 30% of the final fraud categorisation, which has a substantial impact.

	ML Model	Accuracy	f1_score	Recall	Precision
0	Logistic Regression	0.934	0.941	0.953	0.930
1	k-Nearest Neighbors	0.941	0.948	0.953	0.942
2	Support Vector Classifier	0.951	0.957	0.973	0.941
3	Naive Bayes	0.605	0.454	0.294	0.995
4	Decision Tree	0.958	0.962	0.960	0.964
5	Random Forest	0.969	0.973	0.978	0.967
6	Gradient Boosting	0.949	0.955	0.962	0.948
7	Multilayer Perceptrons	0.966	0.970	0.969	0.970

Fig. 4. Different classifiers and their results for URL detection

4.2 Sentiment Analysis, Organizational Entity Recognition, and Spelling Checker Results

For each website, its content is saved in a .txt file and processed by a pre-trained sentiment analysis model, a pre-trained named entity recognition model, and a custom-built Python spelling checker. Each model provides a confidence percentage indicating the likelihood of the website being legitimate or fraudulent, contributing to the overall fraud detection assessment.

4.3 Weighted Scoring and Final Decision

The final decision on URL legitimacy is based on a weighted scoring system. The URL Detection Model contributes 30%, Sentiment Analysis 30%, Organizational Entity Recognition 20%, and the Spelling Checker 20%. After aggregating these weighted scores, the system achieved an overall accuracy of X%, demonstrating its effectiveness in fraud detection.

4.4 Demonstrating Model Performance

For NFSU Goa campus:

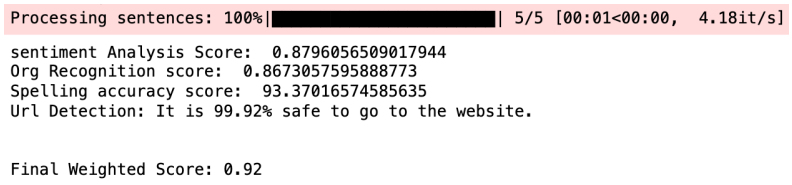

Fig. 5. Results for NFSU GOA website

The analysis indicates that the NFSU GOA website is classified as safe, based on the obtained confidence percentage.

For a sample URL(taken from Phishtank.org)- https://874158635-74531354-5435.pages.dev:

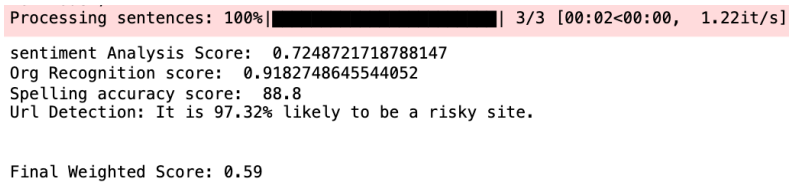

Fig. 6. Results for website taken from Phishtank.org

The analysis suggests that the given URL appears safe based on sentiment analysis, organizational entity recognition, and spelling checks. However, the URL detection model indicates a potential risk, resulting in a reduced overall final weightage of 59%.

4.5 Discussion

The results demonstrate that each component of the fraud detection system plays a critical role in identifying fraudulent websites. The URL Detection Model

(GBC) provides a strong foundation for classifying URLs based on structured features, while sentiment analysis and organizational entity recognition add valuable context from website content. Spelling checks serve as an additional indicator of fraudulent intent, especially when other signals are ambiguous.

The integration of these models, along with the weighted scoring system, improves the overall accuracy and robustness of the system, making it a scalable and effective solution for real-time fraud detection. Future work can explore further refinements, such as incorporating more advanced NLP techniques or expanding the feature set for URL detection to capture even more subtle indicators of fraud.

5 Conclusion

This research introduces a robust, multi-layered system for fraud website detection, combining URL-based analysis with content-driven insights. By integrating machine learning models like Gradient Boosting Classifier for URL evaluation, Twitter's RoBERTa for sentiment analysis, organizational entity recognition, and spelling checks, the system ensures accurate and reliable fraud detection. A weighted scoring mechanism assigns significance to each component, achieving nuanced and comprehensive classifications. The system not only identifies fraudulent websites with high accuracy but also highlights key warning signs, enhancing transparency and interpretability. This scalable approach addresses evolving cybersecurity challenges, offering a practical solution for mitigating online risks and safeguarding users from deceptive platforms.

6 Future Scope

To enhance the accuracy and reliability of our fraud website detection system, we plan to incorporate logo detection as a new feature. This will help identify inconsistencies or misuse of brand logos, adding a visual dimension to the detection process and strengthening fraud identification capabilities. Additionally, we aim to develop a custom NLP-based sentiment analysis model tailored specifically for fraud detection. By storing and labeling website content scraped during experiments, we will create a dedicated dataset to train this model, ensuring self-sufficiency and improved fraud detection accuracy.

Future efforts will include enabling real-time updates to the model using active learning techniques, which will refine performance dynamically. Implementing multilingual support will allow analysis of non-English websites, broadening the system's global applicability. Furthermore, integrating the solution with browser extensions or APIs will enhance accessibility and practicality for real-world fraud prevention scenarios.

References

1. Mehmet Korkmaz, Ozgur Koray Sahingoz, Yuan, & Banu Diri. 11th International Conference on Computing, Communication and Networking Technologies (ICC-

CNT), IEEE, 2020, 10.1109/ICCCNT49239.2020.9225561. Detection of Phishing Websites by Using Machine Learning-Based URL Analysis.
2. Doyen Sahoo, Chenghao Liu, Yuan, & Steven C.H. Hoi. 2017, arXiv:1701.07179. Malicious URL Detection using Machine Learning: A Survey.
3. Saleh Nagi Alsubari, Sachin N. Deshmukh, Ahmed Abdullah Alqarni, Nizar Alsharif, Theyazn H. H. Aldhyani, Fawaz Waselallah Alsaade, & Osamah I. Khalaf. Tech Science Press, 2021, 10.32604/cmc.2022.019625. Data Analytics for the Identification of Fake Reviews Using Supervised Learning.
4. Choon Lin Tan, Kang Leng Chiew, KokSheik Wong, & san Nah Sze. 2016. PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder (https://doi.org/10.1016/j.dss.2016.05.005).
5. Bushra Sabir, M. Ali Babar, Raj Gaire, & Alsharif Abuadbba. IEEE Transactions on Dependable and Secure Computing, IEEE, 2022, 10.1109/TDSC.2022.3218043. Reliability and Robustness analysis of Machine Learning based Phishing URL Detectors.
6. Ahmed Abbasi, Zhu Zhang, David Zimbra, Hsinchun Chen, & Jay F. Nunamaker, Jr. Management Information Systems Research Center, University of Minnesota, MIS Quarterly, 2010, Detecting Fake Websites: The Contribution of Statistical Learning Theory (https://doi.org/10.2307/25750686).