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

Mihir Ranjan and Ranjit Kolkar

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 detection of fraudulent websites has become an increasingly critical challenge in the digital landscape, as cybercriminals constantly evolve their techniques to deceive users. Traditional methods focused on analyzing URL features, but these approaches alone have proven insufficient in the face of more sophisticated fraudulent tactics. This review examines various research contributions in the field of website fraud detection, which have expanded to incorporate both URL analysis and content evaluation, enhancing the effectiveness of detection systems.

2.1 Machine Learning-Based URL Analysis

Korkmaz et al. (2020) proposed a machine learning-based approach to detect phishing websites by analyzing URL features. Their method relied on extracting relevant characteristics from URLs to classify them as either safe or suspicious. The study emphasizes the effectiveness of machine learning algorithms in identifying phishing websites based on patterns and anomalies in URL structures, such as the use of suspicious domain names, IP addresses, or domain registration lengths. This approach laid the foundation for fraud detection by focusing on specific attributes of the URL, an essential step in early fraud detection [1].

Similarly, Sahoo et al. (2017) provided an extensive survey of malicious URL detection methods using machine learning, highlighting the importance of integrating various features such as URL length, domain information, and the presence of HTTPS to assess the legitimacy of a website. Their review underscores the growing reliance on machine learning models for URL-based fraud detection, reflecting the adaptability and efficiency of algorithms in learning from diverse patterns [2].

In addition, Sabir et al. (2022) analyzed the reliability and robustness of machine learning-based phishing URL detectors, noting that the performance of these systems could vary significantly based on the dataset and classifier used. The study calls attention to the need for ongoing improvements in feature selection and classifier training to increase the detection rate of phishing URLs while minimizing false positives [5].

2.2 Content Analysis and Sentiment Detection

However, URL-based detection alone may not suffice, as attackers frequently mimic legitimate domains to deceive users. To address this, content analysis

has emerged as a crucial aspect of fraudulent website detection. Marzuki et al. (2020) explored the integration of content analysis with fraud risk management, proposing a model that incorporates textual features and the overall structure of a website to assess its legitimacy. By analyzing website content, including text readability and structural elements, this approach adds a layer of complexity to detecting fraudulent websites beyond URL analysis [7].

In line with this, Othman et al. (2012) introduced the concept of text readability as a factor in fraud detection. Their study examined the relationship between the complexity of text on a website and its likelihood of being fraudulent. This insight suggests that fraudulent websites often contain poorly constructed content, with unusual sentence structures and inconsistent writing styles, which can be detected using readability algorithms [8].

Furthermore, the integration of sentiment analysis in fraudulent website detection has gained attention. Sentiment analysis, particularly through advanced models like RoBERTa, allows systems to assess the tone and intent behind the content of a website. This method can detect subtle anomalies and inconsistencies in content that may signal fraudulent activity. Although sentiment analysis was not explicitly highlighted in earlier studies, it has become a valuable tool in identifying deception by understanding the emotional cues present in website text.

2.3 Entity Recognition and Validation

The identification and validation of organizational names mentioned on websites is another key element in fraud detection. Ding and Xu (2021) focused on the role of organizational entity recognition in identifying fraudulent texts. Their research utilized deep learning models, such as Bi-GRU, to extract organizational entities from website content and determine their legitimacy. By cross-referencing these entities with known databases of legitimate organizations, they were able to flag websites that mention non-existent or suspicious organizations [9].

This entity recognition approach aligns with the work of Zahedi et al. (2015), who emphasized the importance of identifying elements within fake websites that promote user trust and enhance their fraudulent effectiveness. Their study identified the organizational entities and design elements that contribute to the perceived legitimacy of fake websites, offering a framework for detecting fraud based on these features [10].

2.4 Spelling and Grammar Analysis

A critical but often overlooked aspect of fraudulent websites is the presence of subtle spelling and grammar errors. These inconsistencies, while seemingly minor, are often indicative of fraudulent intent. As cybercriminals aim to mimic legitimate sites, they may overlook small but significant spelling mistakes that are not easily noticeable to users but can be detected by automated systems. The inclusion of spelling and grammar checks in fraud detection algorithms, as noted

by Alsubari et al. (2021), offers a valuable tool for identifying fraudulent websites, particularly those designed to deceive users through false representation [3].

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:

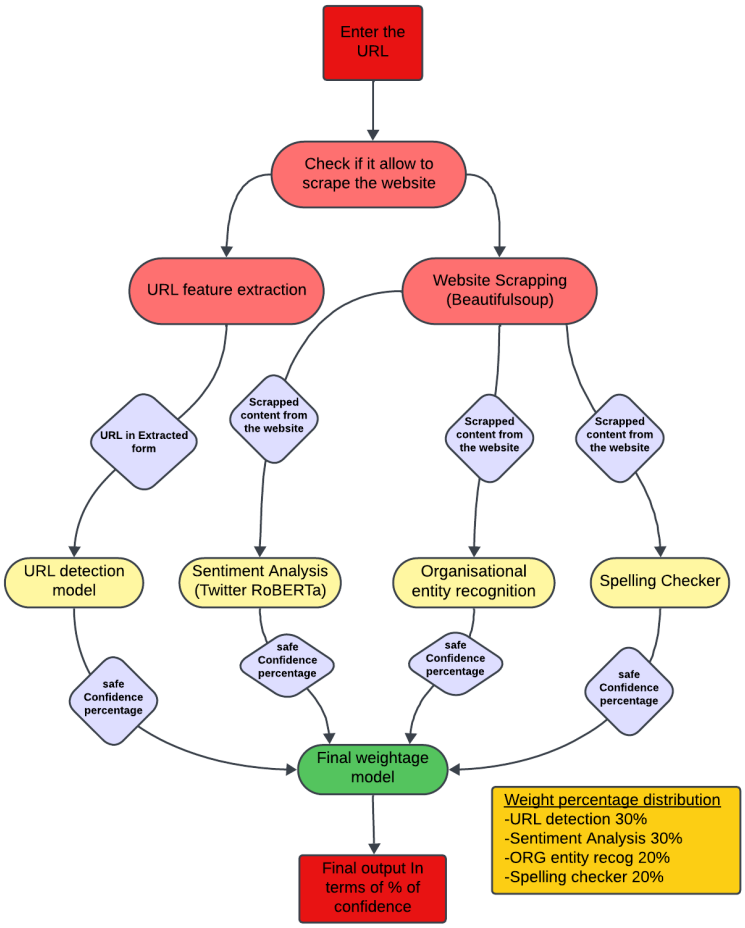

Fig. 1. Workflow of the URL detection system

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 [1].

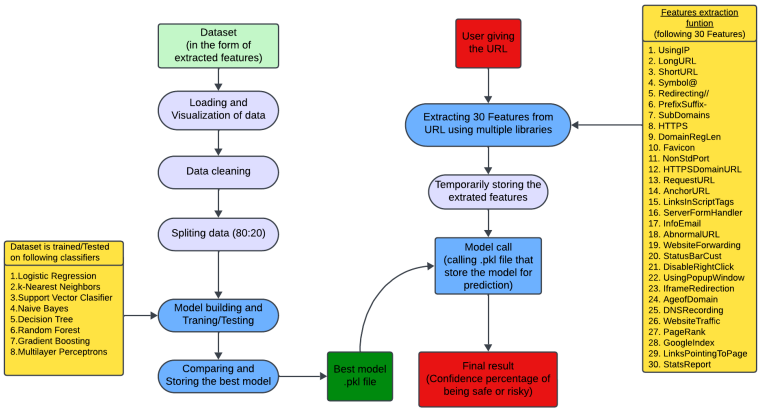

Fig. 2. Workflow of the URL detection model

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 [8].

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 the 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 [9].

3.4 Spelling Checker

Our model processes the 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 [6].

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 [7]. Additionally, organizational name recognition contributes 20%, where the legitimacy and confidence of the identified names are evaluated [9]. Finally, a spelling checker is applied to detect subtle errors indicative of fraud, contributing the remaining 20% [6]. Together, these components form a robust and balanced weight matrix, enabling precise and comprehensive 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 [1].

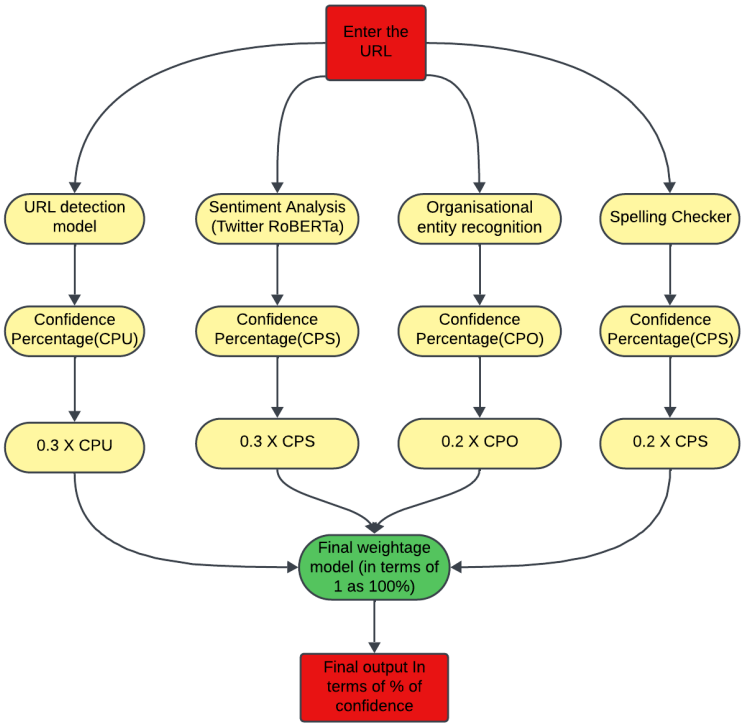

Fig. 3. Workflow Diagram for Final Weight Matrix Computation in Fraud Detection

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 [8] [9] [6].

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 [1] [8] [9] [6].

	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

4.4 Demonstrating Model Performance

For NFSU Goa campus:

```
Processing sentences: 100%|████████████████████| 5/5 [00:01<00:00,  4.18it/s]
sentiment Analysis Score:  0.8796056509017944
Org Recognition score:  0.8673057595888773
Spelling accuracy score:  93.37016574585635
Url Detection: It is 99.92% safe to go to the website.


Final Weighted Score: 0.92
```

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:

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

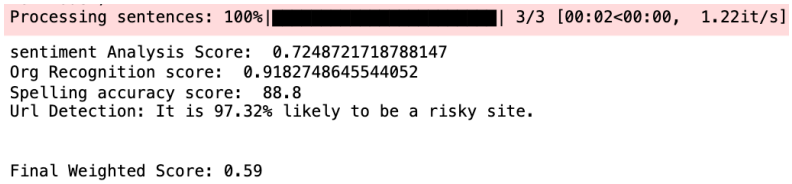

Fig. 6. Results for website taken from Phishtank.org

features, while sentiment analysis and organizational entity recognition add valuable context from website content [8] [9]. Spelling checks serve as an additional indicator of fraudulent intent, especially when other signals are ambiguous [6].

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 [1].

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. Korkmaz, Mehmet, Sahingoz, Ozgur Koray, Yuan, and Diri, Banu. "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis." *11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2020, doi:10.1109/ICCCNT49239.2020.9225561.
2. Sahoo, Doyen, Liu, Chenghao, Yuan, and Hoi, Steven C.H. "Malicious URL Detection using Machine Learning: A Survey.", 2017, arXiv:1701.07179.
3. Alsubari, Saleh Nagi, Deshmukh, Sachin N., Alqarni, Ahmed Abdullah, Alsharif, Nizar, Aldhyani, Theyazn H.H., Alsaade, Fawaz Waselallah, and Khalaf, Osamah I. "Data Analytics for the Identification of Fake Reviews Using Supervised Learning." *Tech Science Press*, 2021, doi:10.32604/cmc.2022.019625.
4. Tan, Choon Lin, Chiew, Kang Leng, Wong, KokSheik, and Nah, Sze San. "PhishWHO: Phishing Webpage Detection via Identity Keywords Extraction and Target Domain Name Finder." *Decision Support Systems*, 2016, doi:10.1016/j.dss.2016.05.005.
5. Sabir, Bushra, Babar, M. Ali, Gaire, Raj, and Abuadbba, Alsharif. "Reliability and Robustness Analysis of Machine Learning-Based Phishing URL Detectors." *IEEE Transactions on Dependable and Secure Computing*, IEEE, 2022, doi:10.1109/TDSC.2022.3218043.
6. Abbasi, Ahmed, Zhang, Zhu, Zimbra, David, Chen, Hsinchun, and Nunamaker, Jay F. "Detecting Fake Websites: The Contribution of Statistical Learning Theory." *MIS Quarterly*, Management Information Systems Research Center, University of Minnesota, 2010, doi:10.2307/25750686.
7. Marzuki, Marziana Madah, Majid, Nik Abdul, Zurina, Wan, Azis, Nur Kamaliah, Rosman, Romzie, Abdulatiff, Haji, and Kamaruzaman, Nik. "Fraud Risk Management Model: A Content Analysis Approach." *The Journal of Asian Finance, Economics and Business*, 2020, doi:10.13106/jafeb.2020.vol7.no10.717.
8. Othman, Intan Waheedah, Hasan, Hazlina, Tapsir, Roszana, Abdul Rahman, Norhafizah, Tarmuji, Indarawati, and Majdi, Suria. "Text Readability and Fraud Detection." *IEEE Symposium on Business, Engineering and Industrial Applications (ISBEIA)*, 2012, doi:10.1109/ISBEIA.2012.6422890.
9. Ding, Xiangwu, and Xu, Cheng. "Research on the Identification of Organizational Entities in Fraudulent Texts via Bi-GRU-Flat." *2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)*, 2021, doi:10.1109/CISAI54367.2021.00013.
10. Zahedi, Fatemeh Mariam, Abbasi, Ahmed, and Chen, Yan. "Fake-Website Detection Tools: Identifying Elements that Promote Individuals' Use and Enhance Their Performance." *Journal of the Association for Information Systems*, 16(6), 2015, doi:10.17705/1jais.00399.
11. Gopal, Ram D., Hojati, Afrouz, and Patterson, Raymond A. "Analysis of Third-Party Request Structures to Detect Fraudulent Websites." *Decision Support Systems*, 2021, doi:10.1016/j.dss.2021.113698.