International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Phishing Website Detection Using Machine Learning: A Comprehensive Study

Priyanshu Desai[1], Mann Shah[2]

[1,2]Student, G H Patel College of Engineering

Abstract

Phishing attacks continue to be a serious danger to consumer privacy and internet safety. Attackers use false websites that look like authentic ones to steal sensitive data, including login credentials and financial information. The ongoing evolution of fraudsters' strategies to get beyond conventional security measures makes it difficult to detect these malicious websites. Due to its capacity to recognize patterns and traits of such websites, machine learning has emerged as a promising method for spotting phishing websites.

In-depth analysis of the use of machine learning algorithms for phishing website prediction and detection is presented in this research report. To create accurate algorithms for detecting phony websites, we investigate numerous data taken from website content, structure, and user behavior. In order to ensure a thorough assessment of the suggested machine learning models, the study includes a broad dataset of both legitimate and phishing websites.

In the study, different machine learning techniques for categorizing phishing websites are evaluated, including decision trees,logistic regression, Multinomial NB. To improve model effectiveness and accuracy, feature selection strategies and data preprocessing techniques are also used. To evaluate the model's predicting ability, we additionally look into the effects of several evaluation criteria.

The findings indicate how well machine learning works at identifying phishing websites, with particularly strong performance in terms of precision, recall, and F1-score. This study also explores the potential integration of the generated models into online browsers and security systems, which would offer real-time defense against phishing assaults.

In conclusion, the study emphasizes the importance of machine learning in tackling the ongoing issue of phishing websites and offers insightful information about the creation of predictive models. The results of this study help to improve internet security, protect user data, and lessen the effect of phishing assaults on the digital world.

Keywords: Phising, Web Security, Machine Learning, Web Scrapping, Feature Enginnering.

Introduction

The widespread use of phishing websites is proving to be a serious problem in the ever-changing online environment, posing serious risks to data integrity, financial security, and personal privacy. The techniques used by bad actors to trick people and compromise their private data evolve along with technology. Application of machine learning techniques for the detection of phishing websites has emerged as an attractive option for strengthening our digital defenses in response to this expanding threat. This paper offers a thorough examination of the ongoing fight against phishing in the digital age by delving into the creative ways and procedures intended to discover and counter these fraudulent online platforms.

Background

- **Phishing:**
  Phishing is a cyber-attack method where deceptive tactics are used to trick individuals into revealing sensitive information like passwords or financial details.
- **Phishing Websites:**
  Phishing websites are fraudulent online platforms designed to impersonate legitimate sites, aiming to deceive visitors into disclosing confidential information.

- **Evolution of phishing attacks and techniques:**

1. Early Deceptive Emails:

Phishing attempts at first used fake emails that appeared to be from reputable companies. These emails were quite simple to recognize because they frequently contained spelling and punctuation mistakes. The receivers were asked to submit their credentials on fraudulent websites by clicking on embedded links.

2. Spear Phishing:

Cybercriminals invented spear phishing, a focused strategy, to increase the success of their attacks. Attackers carefully analyzed their targets before creating phishing emails that were targeted at particular people or businesses. The success rate of phishing attacks was dramatically raised by this approach.

3. Phishing Kits and Automation:

As phishing kits and automated tools emerged, even non-technical people could carry out phishing attacks. These kits streamlined the attack process by including pre-made templates, web hosting, and data capturing tools.

4. Clone Websites:

Clone websites that closely resembled real websites began to be made by cybercriminals. These fake websites deceived users into thinking they were communicating with reputable companies in order to collect sensitive data.

5. Credential Stuffing:

From data breaches, phishers gathered enormous databases of hacked usernames and passwords. They then used this information to conduct assaults known as "credential stuffing," which involved testing stolen login information from one platform across numerous domains.

- Overview of traditional methods of phishing detection:

1. Blacklist:

Blacklists are collections of known malicious IP addresses and websites. Blacklists are used by web browsers and security applications to alert users when they try to access a website that has been identified as a known phishing site. This method works effectively against well-known phishing sites, but it might not be as good at spotting newly constructed or frequently changing phishing pages.

2. Signature-based detection:

In order to detect phishing assaults, signature-based detection uses predetermined patterns or signatures. Incoming emails and web content are marked as suspicious when they match these signatures. This approach, however, performs less well against zero-day attacks or phishing operations that have certain characteristics.

3. Content filtering:

Content filtering software scans emails and web pages for problematic words, links, or attachments. This technique may be successful in preventing clear phishing efforts but may fall victim to more subtle attacks that escape detection.

4. Heuristics and Anomaly Detection:

These techniques examine departures from typical behavior. Heuristics look for anomalies in email or web content, whereas anomaly detection models provide a baseline of "normal" behavior and identify deviations as possible risks. Although they can be useful in some situations, they can also produce false positives and may not be able to keep up with new attack methods.

5. Sender Verification:

SPF (Sender Policy Framework) and DKIM (DomainKeys Identified Mail) are two sender verification methods that aim to validate the legitimacy of the sender's domain. They might not, however, stop attacks using phishing emails.

Machine Learning in Phishing Detection

- Machine learning is a subset of artificial intelligence that involves the development of algorithms and statistical models capable of enabling computer systems to learn from and make predictions or decisions based on data.
- Machine learning models are trained on huge datasets to identify patterns and relationships within the data, in contrast to traditional rule-based systems where explicit instructions are provided. For a number of important reasons, the adaptive nature of machine learning is highly relevant to the topic of cybersecurity.

1. Anomaly Detection:

In spotting anomalies from predicted behaviour, machine learning models excel. This is very useful in cybersecurity for picking up on suspicious activities like illegal access or data breaches. Machine learning algorithms can detect strange patterns and issue alarms by continuously monitoring enormous amounts of network traffic, potentially detecting cyber risks.

2. Pattern Recognition:

Cyberattacks frequently use complex strategies and methods. Whether they relate to malware signatures, phishing efforts, or other dangerous actions, machine learning algorithms are excellent at identifying these trends. This pattern recognition enables the quick identification of dangers that are both known and emerging.

3. Adaptive Defence:

Rapid cyberthreat evolution requires adaptable protection methods. By continuously learning from incoming data, machine learning models are able to adapt to new attack methods and plans. Because of their flexibility, cybersecurity systems are able to remain effective in the face of evolving threats.

4. Scalability:

Manual threat analysis is problematic in the digital age due to the growing number of data produced. Scalability in machine learning makes it possible to analyze huge datasets quickly. This is essential for spotting dangers in real time and taking immediate action.

5. Threat Intelligence:

Machine learning can include external data sources and threat intelligence feeds to improve its comprehension of new dangers. Machine learning models can actively adjust to defend against the most

recent vulnerabilities and attacks by analyzing this data.

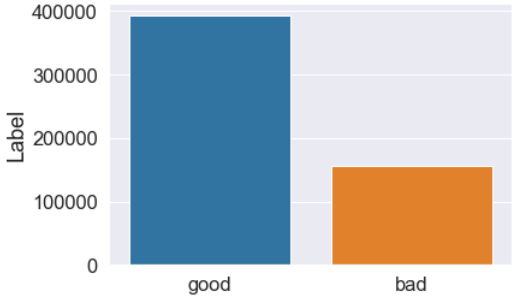
- Machine learning is a useful technology for the detection of phishing websites due to its versatility and capacity to examine large datasets. Its real-time analysis and continuous learning capabilities enable the detection of both well-known and unique phishing attacks. The incorporation of machine learning techniques is essential in the continuous battle to safeguard people against online fraud and data breaches as phishing websites become more complex.
- Advantages of using machine learning in this phishing website detection are as follows:

1. **Adaptability:**
   Machine learning models can adjust to new attack vectors and developing phishing strategies. They are ideally equipped to combat the constantly evolving field of phishing attacks because they continuously learn from new data.
2. **Real-time detection:**
   Machine learning allows for real-time analysis, enabling the rapid identification of emerging phishing websites and the prompt generation of alerts. This is crucial in preventing users from falling victim to phishing attempts.
3. **Pattern Recognition:**
   Pattern recognition in data is where machine learning really shines. It can spot tiny patterns in the content, structure, and behaviour of websites that can point to phishing. With the use of this capability, assaults might be discovered earlier.
4. **Scalability:**
   Machine learning gives scalability, enabling the analysis of enormous datasets at quick rates, which is necessary given the growing amount of websites and online content. For effectively spotting phishing websites in the plethora of online platforms, this is essential.
5. **Efficiency:**
   Automation of the detection process by machine learning can lessen the need for manual analysis. This facilitates faster detection and frees up security professionals to concentrate on more sophisticated threats.

Data Collection and Preprocessing

Dataset

The Dataset is taken from Kaggle.com.The data is containg 5,49,346 unique entries.There are two columns. Label column is prediction column which has 2 categories.The first category is Good - this signifies the urls is not carrying dangerous things and this site is not a Phishing Site and the second one is labelled Bad - this signifies the urls contains dangerous stuffs and this site is a Phishing Site. On top of that, there is no missing value in the dataset.



Data Preprocessing

As we have the data, we have to vectorize our URLs to move further on . With the help of Countvectorizer-create sparse matrix of words using regexpTokenizer, we gather the shorten words, since there are words in URLs that are more essential than other words e.g 'virus', '.exe' ,'.dat' etc. Another advantage of using Countvectorizer is that it helps to create a model for the machine learning algorithms to work on.

The words are shortened and converted into tokens of a sort using regexpTokenizer which is specifically used for that purpose. A RegexpTokenizer is a tokenizer that splits a string into substrings using a regular expression. For example, the tokenizer can form tokens out of alphabetic sequences, money expressions, and any other non-whitespace sequences. After they are tokenized, the feature extraction process is to be initiated where, with the help of Snowball Stemmer, the words are the words are stemmed accordingly.

FEATURE EXTRACTION

Feature extraction refers to the process of translating raw data into numerical features that may be handled while keeping the information in the original data set. It gives better results than applying machine learning directly to the raw data.Here features of the URL will be extracted in the form of stemmed words which will be done with the help of Snowball Stemmer.

Snowball Stemmer: It is a stemming algorithm which is also known as the Porter2 stemming algorithm as it is a better version of the Porter Stemmer since some issues of it were fixed in this stemmer where stemming is reducing a word to its base word or stem in such a way that the words of similar kind lie under a common stem. For example – The words care, cared and caring lie under the same stem 'care'. Thus,it helps with recognizing the URL's for training the dataset accordingly.

Machine Learning Models

Logistic Regression-

Logistic regression is a supervised machine learning technique generally used for classification tasks where the goal is to estimate the chance that an instance of belonging to a specified class. It is used for classification algorithms its name is logistic regression. it's referred to as regression since it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. The distinction between linear regression and logistic regression is that linear regression output is the continuous value that can be anything whereas logistic regression predicts the likelihood that an instance belongs to a specified class or not.

MultinomialNB-

Multinomial Naive Bayes algorithm is a probabilistic learning method that is widely utilized in Natural Language Processing (NLP). The method is based on the Bayes principle and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution generally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also operate. MNB works on the concept of Bayes theorem and assumes that the characteristics are conditionally independent given the class variable.

Decision Tree-

one of the rigorous algorithms for machine learning that is most frequently employed is the Decision Tree algorithm . The decision tree method is simple to apply and comprehend. The decision tree starts its process by selecting the best splitter—regarded as the tree's root—among the qualities that can be used for classification. The algorithm keeps building the tree until it comes to a leaf node. Each internal node of the tree belongs to an attribute, and each leaf node of the tree belongs to a class label. Decision trees are used to generate training models that are used to predict target values or classes in tree representations. These nodes in the decision tree methodology are computed using the gini index and information gain techniques**.**

Implementation and Result

Machine learning methods have been imported using the Scikit-learn tool. For the decision tree algorithm, the dataset is split into training and testing sets in a 50:50 ratio; for the other two, the ratio is 70:30. Training sets are used to train each classifier, and testing sets are used to assess the classifiers' performance. The accuracy score ,precision score,f1 score, recall ,support , false negative rate, and false positive rate of classifiers have been used to assess their performance.

Confusion matrix for Multinomial NB

CONTENT	POSITIVE (actual)	NEGATIVE (actual)
POSITIVE (predicted)	95661	3394
NEGATIVE (predicted)	2449	35833

‘

Confusion matrix for Logistic Regression

CONTENT	POSITIVE (actual)	NEGATIVE (actual)
POSITIVE (predicted)	93247	3483
NEGATIVE (predicted)	2421	36758

Confusion matrix for Decision Tree

CONTENT	POSITIVE (actual)	NEGATIVE (actual)
POSITIVE (predicted)	91173	3864
NEGATIVE (predicted)	2707	35539

Other comparative values are also as follows:

Content	Accuracy score	Precision score	F1 score	recall	support
Logistic regression	0.98	0.98	0.97	0.97	99055
Multinomial NB	0.97	0.98	0.96	0.96	98352
Decision Tree	0.95	0.96	0.97	0.96	97752

Thus from the above tables, after evaluating values it can be observed that logistic regression is the most optimal order with most number of true positives and an accuracy score of **98%** .

Challenges and Limitations

Some of the limitations and challenges this model has are follows:

- Limited to only a few machine learning algorithms.
- Dataset only contains two attributes ,so the model is only generated on the basis of feature engineering from those two attributes.
- An upgraded or an advanced URL can break this detection model and can go through undetected.
- The dataset is only limited to a several type of phishing websites and thus not all websites can be detected by this model.

Countermeasures and Future Directions:

The countermeasures that can be taken are given as follows:

1. **Data Analysis in Real Time:**
   Real-time analysis should be used to spot anomalies and respond rapidly to new phishing attacks. To handle enormous amounts of data in real time, use streaming data processing algorithms.
2. **Collective Learning:**
   Using ensemble approaches, you can combine the capabilities of numerous models to improve overall accuracy and robustness. For diversity, combine decision trees, neural networks, and other methods.

Future directions that can used to improve and enhance the model are given below:

1. **Advances in Deep Learning:**
   Investigate the use of advanced deep learning approaches to capture complicated relationships in data, such as attention mechanisms, transformers, and graph neural networks.
2. **Phishing Detection on the First Day:**
   Investigate ways for detecting zero-day phishing assaults early on by spotting novel patterns and behaviors before they propagate.
3. **Federated Education:**
   Use federated learning to train models across numerous devices without transferring raw data, protecting user privacy and increasing model accuracy.

Case Studies

Following are some of the real-world examples of successful phishing detection using machine learning:

1. **Google's safe browsing:**
   Using machine learning algorithms, Google's Safe Browsing service detects phishing websites and alerts users when they attempt to access them. Users can escape phishing attempts with the help of this service, which is integrated into well-known web browsers.
2. **PayPal's Machine Learning Approach:**
   To identify phishing emails, PayPal deployed a machine learning-based system. To determine whether emails were real or phishing efforts, their model examined the email content and the sender information. The technology increased the precision of phishing email detection and significantly decreased false positives.
3. **Microsoft's Office 365 Advanced Threat Protection:**
   Machine learning is used by Microsoft's Office 365 Advanced Threat Protection (ATP) to recognize and stop phishing emails. To find malicious emails, it examines sender activity, email content, and numerous indicators. For users who are ATP-protected, Microsoft has noticed a marked decline in successful phishing attacks.
4. **IronScales Phishing Detection Platform:**
   Machine learning is used by the cybersecurity platform IronScales to identify phishing emails. To detect phishing attempts, it examines the characteristics, content, and attachments of emails. According to a case study, IronScales users had shorter phishing reaction times and higher detection rates.
5. **Sophos Phish Threat:**
   Employees can receive training and simulations using the Sophos Phish Threat solution, which employs machine learning to help them identify and report phishing attacks. After that, it examines user feedback to enhance its detection abilities and efficiently instruct people.

Conclusion

In this work, we investigated the practicality and the efficiency of using machine learning for phishing detection. We developed thre machine learning models based on multinomial NB, logistic regression and decision trees (DTs) techniques. We then selected the most outperforming model of the three, the ,logistic regression and compared its performance with other solutions in the literature. The overall results show this model achieved the highest performance and outperforms other schemes in the literature.

References

1. Alswailem, B. Alabdullah, N. Alrumayh and A. Alsedrani, "Detecting Phishing Websites Using Machine Learning," 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 2019, pp. 1-6, doi: 10.1109/CAIS.2019.8769571.
2. M. M. Uddin, K. Arfatul Islam, M. Mamun, V. K. Tiwari and J. Park, "A Comparative Analysis of Machine Learning-Based Website Phishing Detection Using URL Information," 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Chengdu, China, 2022, pp. 220-224, doi: 10.1109/PRAI55851.2022.9904055.
3. N. Binti Md Noh and M. N. Bin M. Basri, "Phishing Website Detection Using Random Forest and Support Vector Machine: A Comparison," 2021 2nd International Conference on Artificial Intelligence

and Data Sciences (AiDAS), IPOH, Malaysia, 2021, pp. 1-5, doi: 10.1109/AiDAS53897.2021.9574282.

4. G. S. Gopika, M. Sreekrishna, K. Karthik and C. Reddy, "Privacy Preserving Secure and Efficient Detection of Phishing Websites Using Machine Learning Approach," 2023 2nd International Conference on Edge Computing and Applications (ICECAA), Namakkal, India, 2023, pp. 252-255, doi: 10.1109/ICECAA58104.2023.10212349.
5. S. Jain and C. Gupta, "A Support Vector Machine Learning Technique for Detection of Phishing Websites," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023, pp. 1-6, doi: 10.1109/ISCON57294.2023.10111968.
6. A. Bhavani, R. S. Lakshmi, P. Harshavardhini, P. V. Prakash, N. V. Behara and V. A. Kumar, "Detection of Legitimate and Phishing Websites using Machine Learning," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 366-371, doi: 10.1109/ICSCSS57650.2023.10169697.
7. M. M. Vilas, K. P. Ghansham, S. P. Jaypralash and P. Shila, "Detection of Phishing Website Using Machine Learning Approach," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru, India, 2019, pp. 384-389, doi: 10.1109/ICEECCOT46775.2019.9114695.
8. J. Rashid, T. Mahmood, M. W. Nisar and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Riyadh, Saudi Arabia, 2020, pp. 43-46, doi: 10.1109/SMART-TECH49988.2020.00026.
9. R. A. A. Helmi, M. G. M. Johar and M. A. S. B. M. Hafiz, "Online Phishing Detection Using Machine Learning," 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), Jeddah, Saudi Arabia, 2023, pp. 1-4, doi: 10.1109/ICAISC56366.2023.10085377.
10. M. D. Bhagwat, P. H. Patil and T. S. Vishawanath, "A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 1505-1508, doi: 10.1109/ICICV50876.2021.9388441.