Internet & Web Systems
Term Paper
Harshitha Dodda-02072825
Harshitha_dodda@student.uml.edu

Detecting Malicious URLs Using Machine Learning Techniques

Abstract:

The abstract highlights how, as the digital world develops, there is an increasing risk of cyberattacks, especially through harmful URLs. Attackers use creative methods as more and more services move online, threatening user's security and resulting in large financial losses. The research paper operates a thorough analysis of the literature with a particular focus on machine learning models as an essential tool for identifying dangerous URLs. It examines the shortcomings of previous research, investigates feature kinds and detection systems, and evaluates the datasets used. The report focuses attention to a research gap related to the identification of harmful Arabic websites. The abstract ends with a discussion of the difficulties in detecting malicious URLs and some possible fixes based on a review of a few investigations.

Introduction:

As the internet develops and becomes a more useful tool for a wider range of purposes, including banking, social networking, and e-commerce, the introduction starts out by emphasizing how important it is to keep secure online. It highlights the frequency of attempts to deceive users into clicking on malicious URLs, which results in system hacks and unauthorized access to sensitive data, with a focus on Arabic-speaking internet users. Since attackers can take advantage of standard security measures like blacklisting and heuristic approaches, it is known that these measures are ineffective.

Next, as an alternative to identifying dangerous URLs, the paper discusses the use of artificial intelligence (AI) techniques, namely machine learning (ML) and deep learning (DL). It is highlighted that, especially in the context of cybersecurity, machine learning and deep learning have the capacity to adapt and learn from past events without constant human intervention. The paper highlights machine learning's shown efficiency in identifying recently created URLs as well as its capacity for automatic model upgrades. By citing a thorough examination of 91 papers released between 2012 and 2021, the introduction

provides context for the ensuing literature review. The studies concentrate on the use of ML and DL in the classification of malicious URLs, paying particular attention to linguistic variations and different facets of the detection procedure. The paper's stated contributions including the creation of taxonomies, comparisons, and talks about strategies and difficulties pertaining to malicious URL attacks and detection approaches.

Understanding URL Feature Types and Cyber Threats:

A.URL Features:

High-quality training data and features are necessary for machine learning models to be effective in identifying dangerous URLs. There are three various types of features that are important: lexical features, which are based on URL elements like characters and length; content features, which are derived from webpage content like scripts and HTML tags; and network features, which include host, DNS, and network attributes. Altogether, these characteristics improve the effectiveness of the model; authentic websites typically display a higher amount of content. Utilizing DNS data and keyword analysis, network features are useful for identifying fake sites hosted by less trustworthy providers. These capabilities help to provide a thorough evaluation of URL properties connected to potential dangers, which helps in accurate detection of malicious webpages**.**

B. URL Attack Techniques:

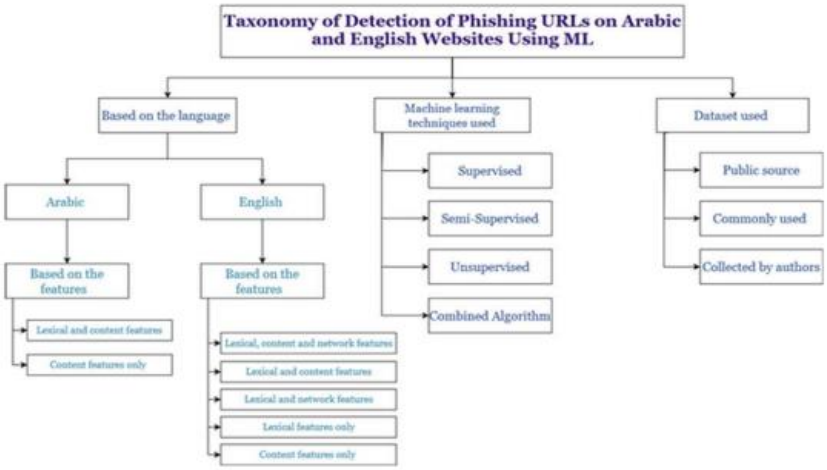
Data integrity, confidentiality, and system availability are put at risk by attack strategies that use malicious URLs. The attacks fall into four categories: defacement, phishing, malware, and spam. They all take advantage of user interactions with the URLs.

1.Spam URL Attack:
To mislead browsers and attract to more people, spammers build misleading webpages to allegedly increase their reputation. Spam URLs are frequently distributed via spam emails that include harmful links, affecting victims' systems and installing malware and adware.

2.Phishing URL Attacks:
Attackers try to obtain sensitive data, including credit card details, by deceiving visitors into visiting fake websites by phishing URLs. Simple mistakes in the URL are easy to deceive non-expert users, making their data open to stealing.

3.Malware URL Attacks:
Users are sent to dangerous websites when malware is installed on their

systems, resulting in fraud, file corruption, and keystroke logging. A variety of dangers, including drive-by downloads, ransomware, keyloggers, trojan horses, spyware, scareware, computer worms, and viruses, are included in the category of malicious software, or malware.



(1) Taxonomy of malicious URL detection on Arabic and English websites using machine learning (ML).

4. Defacement URL Attacks:

By changing a website's look or content, hackers can deceive people into visiting a malicious version of the site. Websites can be attacked by hackers who take advantage of security holes to make changes without authorization. Technically speaking, this action—known as infiltrating a website—aims to take the site down for several reasons.

Malicious URL Detection on Arabic and English studies:

This part examines and highlights the literature on applying machine learning algorithms to identify dangerous websites in Arabic and English. For clarity, it divides features into three categories: lexical, content-based, and network-based. To provide a broader view of ML-based malicious URL identification in a variety of linguistic situations, the summarized studies are then categorized according to feature categories and the language of the website contents.

A. English Based Studies:

The part focuses lexical, content, and network-based elements by gathering and classifying studies on malicious attack detection in English content websites into five sections. A comprehensive summary of the

many strategies and feature combinations investigated in the subject of English website security is given by this methodical layout.

1. **Lexical, Content-Based, and Network-Based Features Studies:**
   A brief overview of several works on machine learning algorithms for malicious URL identification is provided. By utilizing lexical, network-based, and content-based elements in their self-learning method, Aldwairi and Alsalman were able to achieve 87% precision. With 54 features, Xuan et al. used the random forest technique to achieve 96.28% accuracy. With a focus on phishing website detection, Subasi et al. obtained 97.36% accuracy. A parallel neural joint model was introduced by Yuan et al. and achieved 99.78% accuracy. Yu developed a hybrid model that included support vector machines and deep belief networks to achieve 99.96% accuracy. With 97.4% accuracy, Zamir et al. developed a stacking model. With 95% accuracy, Alkhudair et al. used four machine learning techniques. Deebanchakkarawarthi focused on database independence and attained 97% accuracy.

   With a stacked restricted Boltzmann machine, Selvaganapathy et al. were able to obtain 75% accuracy.Using twin SVM and a heuristic approach, Rao et al. were able to achieve 98.05% accuracy. DL mechanisms have been studied by Vinayakumar et al., and LSTM achieved 99.96% accuracy. Patil and Patil's multi-class classification accuracy was 98.44%. A 98.99% accurate DL-based phishing detection technique was presented by Yang et al. A hybrid rule-based system with an accuracy of 97.945% was proposed by Mourtaji et al. With high gradient boosting, Chen et al. created a machine learning model that achieved 99.98% accuracy. Using naive bayes, Vundavalli et al. identified between benign and dangerous websites with 91% accuracy. Using network-based features and word embeddings, Crisan et al. achieved 95.81% accuracy with a multilayer perceptron.

2. **Lexical and Content Based Feature Studies:**
   A model that uses seven lexical and content-based variables from Sina Weibo was proposed by Cao et al. to detect fraudulent URLs in online social networks. Using a Bayesian network, the model achieved 84.74% accuracy. For the goal of identifying phishing attacks, Faris and Yazid created a dataset and used rules-based applications; the decision tree produced the best results, with an accuracy of 96.8%. Using data from Alexa PageRank and PhishTank, Rao and Pais

created a categorization model based on lexical and content variables. They used PCA-RF to reach 99.55% accuracy. Adewole and colleagues introduced a hybrid rule induction method designed to detect phishing attempts. Using a dataset from PhishTank, Yahoo, Alexa, CommonCrawl, and OpenPhish, the program achieved the maximum accuracy of 99.08%. Using a classification method based on association, Kumi et al. extracted 11 lexical and content-based features from datasets gathered from several source crawls, and they achieved 95.83% accuracy. Using 28 content-based and lexical variables from the WEBSPAM-UK2007 and UK-2011 datasets, Liu et al. developed a system for detecting web spam, with RF obtaining the maximum accuracy at 93%.

3. **Lexical and Network-Based features studies**

Using a dataset from Ma et al., Vanhoenshoven et al. obtained 98.26% accuracy with the same classifier as Manjeri et al. proposed, which used RF to categorize URLs with 96% accuracy. Rakotoasimbahoaka et al. used RF-CNN-LSTM to overcome overfitting in ML and DL combo models, achieving 93% accuracy. Using XGBoost, Rao et al. created an ML model that obtained 96.8% accuracy, whereas Rakotoasimbahoaka et al. used CNN-LSTM-RF to get 96% accuracy. Chiramdasu et al. detected phishing with 93% accuracy by using KNN. ELM was used by Shi et al. to identify malware domain names with a 95% accuracy rate. Using RF, Parekh et al. recognized phishing websites with 95% accuracy; whereas Butnaru et al. used an ML model to reach 98.86% accuracy.

Using 14 characteristics, Shantanu et al. performed with 99.7% accuracy. Astorino et al. used spherical separation to attain an accuracy of 86.3%. The JCLA model was used by Peng et al. with 98.26% accuracy. Using NN, Wadas identified phishing URLs with 78.4% accuracy. Patgiri et al. obtained 93.3% accuracy using RF, compared to 90.51% accuracy for Sadique et al. Prieto et al. achieved 89% accuracy with LR, while Rupa et al. achieved 99.61% accuracy with the same classifier. A Chrome addon that achieved 96.11% accuracy with RF was proposed by Desai et al. SVM was utilized by Akour et al. to detect phishing, with 96.3% accuracy. He and colleagues used RF to obtain 90.81% accuracy following feature selection. The hybrid DL models suggested by Ozcan et al. had an accuracy of 99.21% using DNN-BiLSTM. With their feature-

optimized NB and SVM models, Lee et al. were able to detect dangerous URLs with 99% accuracy.

4. **Lexical Studies:**
   Raja et al. presented a method that uses 20 chosen lexical features from the UNB dataset to detect malicious URLs with 99% accuracy. Vanitha N et al. used LR to achieve 98.42% accuracy using GitHub data, with the goal of autonomous learning. In their comparison of LR with DT, Aalla et al. found that LR had 97.5% accuracy. Using an FFNN with CICANDMAL2017, Ateeq and Moreb were able to detect 98.48% of URLs. When Shivangi et al. used a Chrome extension, the accuracy of the LSTM model was 96.89%. ID3 was utilized by Pingle et al. as a method for lexical feature-based web page damage detection. On Kaggle data, Lakshmanarao et al. employed RF with HV and achieved 97.5% accuracy. By combining ML techniques, Khan et al. were able to obtain 99.72% accuracy using 47 lexical characteristics**.** Abutaha et al. used SVM to achieve 99.896% accuracy. Zhao et al. used GRU NN to multiclassify data with 98.5% accuracy. NLP was utilized by Hai and Hwang to classify files as benign or malignant, with SVM achieving 97.1% accuracy. SVM was utilized by Banik and Sarma to identify phishing with 96.35% accuracy. Using 17 lexical characteristics, Sameen et al. built PhishHaven with 98% accuracy. Johnson et al. employed RF, and on ISCX-URL-2016, they achieved 96.99% accuracy. Proposing DBLSTM, Liang and Yan maintained a high degree of precision, 93–95%. Joshi et al. classified benign and harmful URLs with 92% accuracy using RF. Using five lexical features, the COVID-19 model by Ispahany and Islam achieved 99.2% accuracy. Using k-means, Afzal et al.'s URLdeepDetect achieved 99.7% accuracy.

5. **Content-Based Features Studies:**
   Using data from PhishTank and Alexa, Altay et al. presented a supervised machine learning approach for classifying web pages. By employing a keyword density extractor library, they were able to extract 8,000 content features. SVM-RBF attained a 98.24% accuracy rate. McGahagan et al. evaluated extra webpage elements to enhance malicious website identification. They chose 26 content-based features using an Alexa dataset and a dataset from Cisco Talos Intelligence Group. Without sampling, RF obtained the highest accuracy of 91.36%. A unique method for **spotting phishing risks** was presented

by Jain and Gupta, who looked at hyperlinks in HTML source code. PhishTank, the top websites listed on Alexa, Stuffgate Free Online Website Analyzer, and a list of online payment service providers were all included in their dataset. The LR model, which had an accuracy of 98.42%, had the best results.

Challenges and Recommendations:

Although machine learning (ML) has made great progress in detecting malicious URLs over the past ten years, there are still unanswered questions. Small data sample sizes are one of the limitations, highlighting the necessity of balanced datasets and effective balancing strategies. Scalable environments, such as cloud computing, are necessary to meet the issue of massive data collection. Several selection strategies can be used to address issues including feature-related issues and the absence of analysis of obfuscated JavaScripts. The dynamic nature of attributes that differentiate between reputable and suspicious URLs presents a problem, indicating the need to investigate Concept Drift detection methods. The effects of catastrophic forgetting can be reduced by group strategies. The significance of sustainability and validation models over time is underscored by the diversity of malicious URLs and real-world network traffic. To make sure ML models are appropriate, they should be tested on future data and trained on samples from a particular time period in order to represent the dynamic nature of cyberattacks.

Conclusion:

The paper discusses several research on the identification of malicious website links, with an emphasis on the application of machine learning (ML) to both Arabic and non-Arabic content. It highlights the significance of URL properties by classifying and contrasting various methods. Important discoveries include the extensive application of lexical characteristics, particularly in Arabic text, and a preference towards methods such as SVM, RF, CNN, and XGBoost, the latter two of which demonstrate good accuracy. The article points out that whereas non-Arabic studies employ open-source datasets, Arabic studies frequently create their own. Along with highlighting issues like dataset size, outliers, and feature selection, it also makes recommendations for future research topics aimed at enhancing ML detection methods.

Reference:

[1] https://ieeexplore.ieee.org/document/9950508
[2] M. E. H. V. S. Aalla and N. R. Dumpala, "Malicious URL prediction using machine learning techniques," Ann. Romanian Soc. Cell Biol., vol. 25, no. 5, pp. 2170–2176, 2021. Accessed: Jan. 19, 2022.[Online]. Available: https://www.annalsofrscb.ro/index.php/journal/article/view/4752
[3] M. Aljabri and S. Mirza, "Phishing attacks detection using machine learning and deep learning models," in Proc. 7th Int. Conf. Data Sci. Mach. Learn. Appl. (CDMA), Mar. 2022, pp. 175–180, doi: 10.1109/cdma54072.2022.00034.
[4] C. Johnson, B. Khadka, R. B. Basnet, and T. Doleck, "Towards detecting and classifying malicious URLS using deep learning," J. Wireless Mob Netw., Ubiquitous Comput. Dependable Appl., vol. 11, no. 4, pp. 31–48, Dec. 2020, doi: 10.22667/JOWUA.2020.12.31.