


Chapter 8

Malicious URL Detection Using Machine Learning

Ferhat Ozgur Catak

 <https://orcid.org/0000-0002-2434-9966>
Simula Research Laboratory, Oslo, Norway

Keyser Sahinbas

Istanbul Medipol University, Turkey

Volkan Dörtkardeş

Şahıs Adına, Turkey

ABSTRACT

Recently, with the increase in Internet usage, cybersecurity has been a significant challenge for computer systems. Different malicious URLs emit different malicious software and try to capture user information. Signature-based approaches have often been used to detect such websites and detected malicious URLs have been attempted to restrict access by using various security components. This chapter proposes using host-based and lexical features of the associated URLs to better improve the performance of classifiers for detecting malicious web sites. Random forest models and gradient boosting classifier are applied to create a URL classifier using URL string attributes as features. The highest accuracy was achieved by random forest as 98.6%. The results show that being able to identify malicious websites based on URL alone and classify them as spam URLs without relying on page content will result in significant resource savings as well as safe browsing experience for the user.

INTRODUCTION

The significance of the World Wide Web (WWW) has attracted increasing attention because of the growth and promotion of social networking, online banking, and e-commerce. While new development in communication technologies promote new e-commerce opportunities, it causes new opportunities for attackers as well. Nowadays, on the Internet, millions of such websites are commonly referred to as mali-

DOI: 10.4018/978-1-7998-5101-1.ch008

Malicious URL Detection Using Machine Learning

cious web sites. It was noted that the technological advancements caused some techniques to attack and scam users such as spam SMS in social networks, online gambling, phishing, financial fraud, fraudulent prize-winning, and fake TV shopping (Jeong, Lee, Park, & Kim, 2017). In recent years, most attacking methods are applied by spreading compromised URLs and fishing, and malicious Uniform Resource Locators (URLs) addresses are the leading methods used by hackers to perform malicious activities. Common types of attacks using malicious URLs can be categorized into Spam, Drive-by Download, Social Engineering, and Phishing (Kim, Jeong, Kim, & So, 2011). Spam is called to be sent to unsolicited messages by force for advertising or phishing, which we do not request and do not want to receive. These attacks have caused a tremendous amount of damage (Verma, Crane, & Gnawali, 2018). The download of malware while visiting a URL is called as Drive-by download (Cova, Kruegel, & Vigna, 2010). Lastly, Social Engineering and Phishing attacks guide users to reveal sensitive and private information by acting as genuine web pages (Heartfield & Loukas, 2015). The attackers create copies of the popular web pages used by users such as Facebook and Google and compromise victim computers by placing various pieces of malicious code in the manipulated web site's HTML code. Besides, the ubiquitous use of smartphones encourages the increase of mobile and Quick Response (QR) code phishing activities, especially to deceive the elderly that encode fake URLs in QR codes. The dark side of the Internet has attracted increasing attention and bedeviled the world (Patil & Patil, 2015). Internet security software cannot always detect malware from malicious websites and drive-by downloads. It can, however, prevent you from getting them in the first place (Symantec, 2020). Malicious URLs detection is not adequately addressed yet and causes enormous losses each year. In the fourth quarter of 2019, more than 162,000 unique phishing URLs were detected globally (Statista, 2020).

Even though the security components used today are trying to detect such malicious sites and web addresses, these components are evading by using different methods implemented by the attackers. Researchers have studied to gather effective solutions for Malicious URL Detection. One of the most popular ways is the blacklist method that uses records of known malicious URLs to filter the incoming URLs. However, blacklists have some limitations, and this approach useless for new malicious sites that are created continuously. Security components have started to use innovative applications of machine learning and artificial intelligence-based prediction models to cope with this problem, during the last decades (Garera, Provos, Chew, & Rubin, 2007) (Kuyama & Kakizaki, 2016) (Ma, Saul, Savage, & Voelker, Beyond blacklists: learning to detect malicious web sites from suspicious URLs, 2009) (Ma, Saul, Savage, & Voelker., Learning to Detect Malicious URLs, 2011). They have started to prefer machine learning and artificial intelligence prediction instead of being signature-based for Malicious URL Detection. Machine Learning approaches apply a set of URLs as training data and learn a prediction function to classify whether a URL is malicious or benign. This approach allows them to generalize to new URLs, unlike blacklisting methods. Soon, these solutions will need to be used in Cyber-Physical Systems (CPS), and the other area will be to identify harmful sites and URL addresses. As a result, it can be noted that Artificial Intelligence-based antimalware tools will aid to detect recent malware attacks and develop scanning engines.

This chapter aims to present the basics of machine learning-based malicious URL detection. The rest of the chapter is organized as follows. In the Background section, a review of the existing approach and a summary of the literature in the field of URL classification is presented, In Dataset and Analysis section, the publicly available dataset is discussed. In Method Section, the fundamentals of machine-learning models are explained. In the Experiment section, the detection performance comparisons of different algorithms are evaluated. Lastly, conclusion and future directions remarks are given.

19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/malicious-url-detection-using-machine-learning/266138

Related Content

Trust based Dynamic Multicast Group Routing Ensuring Reliability for Ubiquitous Environment in MANETs

Shobha Tyagi, Subhranil Somand Qamar Parvez Rana (2017). *International Journal of Ambient Computing and Intelligence* (pp. 70-97).

www.irma-international.org/article/trust-based-dynamic-multicast-group-routing-ensuring-reliability-for-ubiquitous-environment-in-manets/176714

Psychological Effects of Dominant Responses to Early Warning Alerts

Thomas Jack Huggins, Lili Yang, Jin Zhang, Marion Lara Tanand Raj Prasanna (2021). *International Journal of Ambient Computing and Intelligence* (pp. 1-15).

www.irma-international.org/article/psychological-effects-of-dominant-responses-to-early-warning-alerts/279583

Problem-Solving Manager: Creating an Innovative Learning Organization

(2022). *Socrates Digital™ for Learning and Problem Solving* (pp. 234-254).

www.irma-international.org/chapter/problem-solving-manager/290570

A Valid Scheme to Evaluate Fuzzy Definite Integrals by Applying the CADNA Library

Mohammad Ali Fariborzi Araghiand Samad Noeiaghdam (2017). *International Journal of Fuzzy System Applications* (pp. 1-20).

www.irma-international.org/article/a-valid-scheme-to-evaluate-fuzzy-definite-integrals-by-applying-the-cadna-library/189126

A User Authentication Schema Under the Integration of Mobile Edge Computing and Blockchain Technology

Feng Xueand Fangju Li (2023). *International Journal of Ambient Computing and Intelligence* (pp. 1-20).

www.irma-international.org/article/a-user-authentication-schema-under-the-integration-of-mobile-edge-computing-and-blockchain-technology/327027