Malicious Webpage Detection Based on Feature Fusion Using Natural Language Processing and Machine Learning

Publisher: IEEE

Pradeepa G ; Devi R **All Authors**

Abstract:
Malicious websites are purposefully designed to deceive internet users to steal sensitive personal information, infect the victim's system with malware, cause financial l... **View more**

Metadata

Abstract:
Malicious websites are purposefully designed to deceive internet users to steal sensitive personal information, infect the victim's system with malware, cause financial losses, and damage the victim's reputation. Finding these pages or links is hard for internet users. Such websites are discovered using detection tools. The majority of detection techniques use blacklisting or whitelisting strategies to find and prevent malicious websites. However, compiling such a sizable list of website links is a time-consuming job that is challenging to update regularly. Therefore, the researchers employ machine learning-based methods to identify these fraudulent connections. These methods are based on the features taken from URLs or web pages. Additionally, features such as DNS details, webpage reputation, and visual similarity data are used. However, these features are few and do not fully utilize the URLs or website contents. This work focuses on merging URL lexical features and content-based features for malicious webpage detection in order to fully exploit the dataset's potential. Natural language processing methods like Hashing, Count, and Term Frequency - Inverse Document Frequency (TF-IDF) vectorizers are employed to extract features from the content of Web pages. The suggested approach's efficiency is evaluated by using the most well-known machine learning methods. The outcome shows that the Count vectorizer with Random Forest achieves a higher accuracy of 91.17% with 500 features.

Published in: 2023 2nd International Conference on Edge Computing and Applications (ICECAA)

Date of Conference: 19-21 July 2023

DOI: 10.1109/ICECAA58104.2023.10212120

Date Added to IEEE *Xplore*: 16 August 2023

Publisher: IEEE

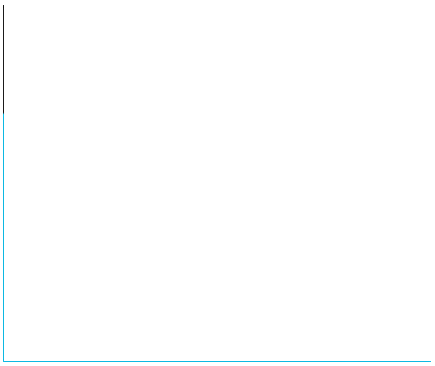
ISBN Information:

Conference Location: Namakkal, India

Contents

Malicious websites are those that are intended to harm or exploit users who visit them. They usually contain information, code, or links designed to mislead users into taking activities that pose risks to their computers or compromise their personal information [1]. Malicious websites can take many different forms, such as Phishing websites trying to get users to reveal personal information including passwords, credit card details, and other sensitive data. They impersonate banks, social networking platforms, and other legitimate websites to steal users' information. Malware distribution websites spread viruses, Trojan horses, and spyware to users' computers and mobile devices. Users may accidentally download malware by clicking on a link or downloading a file from a malicious website. Malware distribution websites spread viruses, trojan horses, and spyware to users' computers and mobile devices. Users may accidentally download malware by clicking on a link or downloading a file from a malicious website. Scam websites offer bogus products or services, such as bogus antivirus software or lottery scams, to dupe consumers into handing over money or personal information. Drive-by download websites leverage browser vulnerabilities to install malware on users' devices without their consent. Rogue security software websites provide bogus security software that purports to protect a user's PC against malware but is malware. Consequently, malicious websites pose significant risks to users and organizations. Malicious website detection is crucial because it shields individuals and organizations from a variety of risks, such as data theft, financial loss, brand damage, and the propagation of malware. Researchers introduce several detection approaches employing distinct features and detecting techniques. URL lexical features, content-based features, DNS-based features, and website reputation-based features are most typically utilized [2] [3]. Researchers use a variety of detection methods, including blacklisting, rules-based, and machine learning/deep learning techniques [4]. Most detection approaches employ blacklisting or whitelisting tactics to identify and block harmful websites. However, compiling such a lengthy inventory of website links is a time-consuming task that is difficult to maintain. Rules-based methods [5] use a small number of features and a threshold value of critical features to find malicious websites. However, domain knowledge is needed to choose the best features and threshold values. Therefore, the researchers use a method based on machine learning or deep learning. Machine learning techniques use the features that are extracted from URLs or content of the web pages. In addition, features like DNS information, webpage reputation, and information on visual similarity are also utilized. However, the dataset's potential was not fully realized due to the limited number of features utilized. To address this issue, this paper fusion the URL lexical features (26 features) and web content features by using text encoding methods in natural language processing (NLP) and machine learning (ML) models. To create features from the webpage content, natural language processing techniques like Count, TF-IDF, and Hashing Vectorizer are used [4]. The effectiveness of the provided approach is assessed using the most popular machine learning algorithms. Count vectorizer with random forest achieves greater accuracy.

Metrics



More Like This

Visualization of Driving Behavior Based on Hidden Feature Extraction by Using Deep Learning
IEEE Transactions on Intelligent Transportation Systems
Published: 2017

Interactive Exploration of Surveillance Video through Action Shot Summarization and Trajectory Visualization
IEEE Transactions on Visualization and Computer Graphics
Published: 2013

Show More