Borneo International Journal eISSN 2636-9826; Vol. 6 (1); 2023; 61-64
Published by *Majmuah Enterprise*
www.majmuah.com

Exploring and Evaluating a Malicious Site on the Dark Web Using Machine Learning

SITI NURULNADIRAH AZHAR and MOHAMAD FADLI ZOLKIPLI
[1]*School Of Computing, College Arts And Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia*
Email: [1]siti_nurulnadirah@uum.edu.my , [2]m.fadli.zolkipli@uum.edu.my

Received: February 23, 2023
Accepted: February 26, 2023
Online Published: March 01, 2023

Abstract

Various web-based assaults, especially Drive-by-Download operations, were getting more significant in this era of globalisation. It's indeed crucial that gather data regarding harmful websites that might offer a blacklist-based detecting tool throughout order to safeguard genuine users. They suggest a technique through this research to collect the URLs of harmful websites mostly on dark web. This suggested method works by scanning dark web resources as well as gathers harmful URLs which are verified through using Gred algorithm or VirusTotal. To use a content embedded and gradient boosting decision tree model, they additionally forecast risky subcategories of gathered websites that seem to be possibly harmful. We show through trials that the suggested method has an F1-score accuracy of 0.82 for predicting harmful site types. Harmful websites, commonly referred to as malicious unified resource locators (URLs), serve as the foundation for a variety of online offences include phishing, spamming, identity theft, financial fraud, and malware. A regular and severe danger to cyber security has been found. Researchers found several artificial learning methods, such as delisting systems, to effectively identify and classify hazardous URLs online. Delisting is entirely useless for detecting malicious URL variants or freshly created URLs. In real-time settings, it also requires human input and takes a while to complete.

Keywords: Era of globalization, harmful sites, dark web, embedded, URL.

I.Introduction

The term "dark web" describes communication infrastructure that use the Internet and certain content of the World Wide Web that is only accessed without software application, configuration, or authorization. On the dark web, personal computer systems may communicate and carry out transactions in privately without disclosing sensitive data like a specific location. Even while the term "dark web," which refers to the section of the Web that browsers do not monitor, is regularly used wrongly, it only makes up a minor piece of the dark web. It shouldn't be unexpected that perhaps the way we just use internet and other linked technologies is changing to more closely resemble how we use the physical universe because desktop products have become present in practically every part of our life. The dark web are sometimes confused with one another. In 2009, the phrase "dark web" first appeared. The exact beginning of the dark web. However, is uncertain.

Many internet users only access the information available through standard web browsers, or the surface web. Although it only makes up a minor portion of the dark web, the dark web requires specialised software to access its information. This misunderstanding began at least in 2009. Since then, despite suggestions that they should be differentiated, the two names have frequently been used interchangeably, notably in reporting about Silk Road. In reality, this web consists of each individual host, workstation, or other piece of equipment interconnected to another beyond a network of networks. The Surface Web and the Dark Web seem to be the consequent two components. The Surface Web that's what most people consider to be "the internet." It consists of a group of sites that may be freely distributed using common browsers and access technologies and are searched through such search engines as Google, and Yahoo. Even though it might appear to be a tonne of data here, the Surface Web is simply the top of the iceberg.

This entire iceberg, known as the Dark Web, seems to be mostly inaccessible to clients of the web site. Even though it is difficult to gauge this Dark Web's size, experts believe it will be around 4000 and 5000 times greater than the Surface Web (Finklea 2015). Some people were surprised to learn that 90% of web traffic originates from the Dark Web due to their unaware that frequently use it (Greenberg 2014). Since it may only be accessible via software development kit connections, information on websites like Facebook, Twitter, or Snapchat is referred to as being on the Dark Web. Online messaging information and document platforms such Dropbox and Google Drive are two other significant data sources (Greenberg 2014).

Authorities are in a difficult situation due to the opposite ways that the dark web platform is used. Consequently, it is necessary to get insight into the underlying workings of the dark web in order to perhaps improve the surveillance of activities taking place there. Finding the central nodes that support the network structure may be made easier with the aid of the dark web's web graph. It could also shed light on how the network architecture facilitates illegal activity.

2.Background

Users of the darknet may keep their location hidden and ensure secrecy thanks to layered encryption technologies. By routing users' data via a significant number of intermediary sites, online black data encryption protects users' identities and guarantees secrecy. The data supplied can only be decoded by the node that follows the exit node in the scheme. Due to the intricate structure, it is practically impossible to repeat the node path and decode the data layer by layer. Due to the high level of encryption, websites are unable to track users' IP addresses and geolocations, and visitors are unable to obtain this information about the host. Artificial learning-based techniques recover manually created characteristics like grammatical, morphological, situational, or philosophical data, summary statistics of URL strings, n-grams, bag-of-words, link architectures, material structure, DNS information, network traffic, etc. They also depend in part on the product implementation stage. Machine learning-based systems' feature engineering has to adapt to the new dangerous URLs. Due to its impressive performance on a number of artificial intelligence (AI) issues in the fields of image processing, audio processing, natural language processing, and other disciplines, deep learning has attracted some of the most attention in recent years. These do have the ability to dynamically retrieve information from the raw source text.

Researchers test various recurrent neural networks, such as recurrent neural network (RNN), identity-recurrent neural network (I-RNN), long short-term memory (LSTM), convolution neural network (CNN), and convolutional neural network-long short-term memory (CNN-LSTM) architectures, in order to take advantage of this and apply algorithms towards the challenge of identifying malicious URLs. The optimal parameter for deep learning architecture is determined experimentally utilising numerous network variables and communication network topologies.

3.Uses Of Darkweb

The dark web occasionally serves unethical or even unlawful goals because of its anonymity. This would include trading in illegally obtained narcotics, firearms, identities, and credentials, as well as illegal obscenity and other potentially harmful goods. Government authorities have recently taken down a number of websites that hosted illicit content, including Silk Road, AlphaBay, and Hansa. Over the past two decades, the darkness of the dark web has also exacerbated cybersecurity worries and a number of data thefts.

4.Browsing, Communicating, and Using Methods on the Dark Web

The term "anonymity" in the Dark Web comes from the Greek word "anonymia," which means disguising one's identification through others. Consumers leave digital traces of every activity we take online, which are stored online as information. The assurance of anonymity is established if the Internet Protocol address cannot be traced. Information on the Internet is sent globally by TOR clients using voluntary server networks. By doing this, user information may be hidden and tracking of activity is completely eliminated. By enabling cybercriminals to conduct crimes online and hide their tracks, the dark web also has detrimental impacts. It is seen as a suitable avenue enabling authorities to trade confidential information, for journalism to get around surveillance, as well as for dissidents to "escape" from brutal regimes. Through a system of computers, the onion approach allows for anonymous communication. The asymmetric encryption method is used to send messages, which are subsequently routed through onion routers, which are network nodes.

This same cryptographic layer is removed from each onion router whenever the signal is transmitted to them in exactly the same way that the onion peeling has been removed in order to prevent the routing instructions from being discovered. A signal is then sent to the other router, and this process is repeated until it is sent to a particular location. By using this method, the intermediary nodes are shielded from "information" on the origin, recipient, and statement's contents.

5.Conclusion

Different options when it comes to identifying network risks based on cyberterrorism attacks encounter problems such as insufficient specificity and a lack of hypothesis. Unknown cyber terrorist threats are impossible to combat, and it is very simple to evade rule-based surveillance of cyber terrorist attacks. As a result, we propose a novel approach to discovering vulnerabilities on the dark Web network that makes use of artificial intelligence and the Commercial Internet of Things platform. The method is targeted at analysing datasets related to cyber terrorist attacks, where very minimal preparation is required and automated feature extraction is handled by the LSTM itself. Excellent experimental results were obtained using the LSTM dataset. This system has generated cutting-edge results in the identification of cyber terrorist threats, having a maximum classification performance yet maintaining a low proportion of false alarms. But we must keep in mind that some of these defects could be more dangerous than their ranking led us to assume.

Acknowledgement

The authors would like to thank to all School of Computing members who involved in this study. This study was conducted for the purpose of Systems & Network Security Project. This work was supported by University Utara Malaysia.

References

About. HeinOnline. (2021, March 8). Retrieved from https://heinonline.org/HOL/LandingPage?handle=hein.journals%2Fwnelr41&div=11&id=&page=

Affect intensity analysis of Dark Web Forums. IEEE Xplore. (n.d.). Retrieved from https://ieeexplore.ieee.org/abstract/document/4258712

Balhara, A., Ubba, S., Sharma, Y., & Chawla, P. (2021, July 12). Exploring and analyzing dark web. SSRN. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3879619

Covrig, B., Mikelarena, E. B., Rosca, C., Goanta, C., Spanakis, G., & Zarras, A. (1970, January 1). Upside down: Exploring the ecosystem of dark web data markets. SpringerLink. Retrieved from https://link.springer.com/chapter/10.1007/978-3-031-06975-8_28

Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts). IEEE Xplore. (n.d.). Retrieved from https://ieeexplore.ieee.org/abstract/document/8004907

Exploring the dark web for Cyber Threat Intelligence using machine leaning. IEEE Xplore. (n.d.). Retrieved from https://ieeexplore.ieee.org/abstract/document/8823360

Exploring the evolution of exploit-sharing hackers: An unsupervised graph embedding approach. IEEE Xplore. (n.d.). Retrieved from https://ieeexplore.ieee.org/abstract/document/9624846

Exploring the topological properties of the Tor Dark Web. IEEE Xplore. (n.d.). Retrieved from https://ieeexplore.ieee.org/abstract/document/9340182

Exploring the use of zcash cryptocurrency for illicit or ... - Rand. (n.d.). Retrieved from https://www.rand.org/content/dam/rand/pubs/research_reports/RR4400/RR4418/RAND_RR4418.pdf

Faizan, M., & Khan, R. A. (n.d.). Exploring and analyzing the dark web: A new alchemy. First Monday. Retrieved from https://journals.uic.edu/ojs/index.php/fm/article/view/9473

Google. (n.d.). Dark web. Google Books. Retrieved from https://books.google.com.my/books?hl=en&lr=&id=a6t1ptKndvAC&oi=fnd&pg=PR3&dq=exploring%2Bdark%2Bweb&ots=sGJDRmiUTc&sig=SSwyw0cwGjUMnge5Dm3_nItFnBc&redir_esc=y#v=onepage&q&f=false

Guetler, V. F. (n.d.). Exploring cyberterrorism, topic models and social networks of Jihadists Dark Web Forums: A computational social science approach. The Research Repository @ WVU. Retrieved from https://researchrepository.wvu.edu/etd/11253/

Jan Kietzmann - Researchgate. (n.d.). Retrieved from https://www.researchgate.net/profile/Jan-Kietzmann

Jianwei Ding Science and Technology on Communication Security Laboratory, Ding, J., Science and Technology on Communication Security Laboratory, Xiaoyu Guo Science and Technology on Communication Security Laboratory, Guo, X., Zhouguo Chen Science and Technology on Communication Security Laboratory, Chen, Z., & Metrics,

O. M. V. A. (2020, January 1). Big data analyses of ZeroNet sites for exploring the new generation Darkweb: Proceedings of the 3rd International Conference on Software Engineering and Information Management. ACM Other conferences. Retrieved from https://dl.acm.org/doi/abs/10.1145/3378936.3378981

Research explorer. Research Explorer | The University of Manchester. (n.d.). Retrieved from https://www.research.manchester.ac.uk/portal/

Ven, K. van de, & Koenraadt, R. (2017, November 22). Exploring the relationship between online buyers and sellers of image and Performance Enhancing Drugs (IPEDS): Quality issues, trust and self-regulation. International Journal of Drug Policy. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0955395917302827

Exploring open source information for cyber threat intelligence. IEEE Xplore. (n.d.). Retrieved from https://ieeexplore.ieee.org/abstract/document/9378220

Hensmans, M. (2021, April 5). Exploring the dark and bright sides of internet democracy: Ethos-reversing and Ethos-Renewing Digital Transformation. Technological Forecasting and Social Change. Retrieved from https://www.sciencedirect.com/science/article/pii/S0040162521002092?casa_token=yjgY3dfaWB8AAAAA%3AcAzykOaIc8f8YWJG6kbNwBFkaSZ_wszwho_2KgMKBkkIAwXaEu7rgCwox6rQlKk8GoWHdjjpCxsq

Naveen, I. N. V. D., Manamohana, K., & Verma, R. (2019, January 1). Detection of malicious urls using machine learning techniques. Manipal Academy of Higher Education, Manipal, India. Retrieved from https://manipal.pure.elsevier.com/en/publications/detection-of-malicious-urls-using-machine-learning-techniques and Applied Sciences, 2(8), 257840