Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

An optimization-based approach to identifying and detecting malicious activity in the dark web

Arif Hasan Abd Ali*

**Electrical and Computer Engineering and Advanced Technology, Computer Engineering- Software. Urmia University, Karbala ,Iraq*

**Email:aarefaref1@gmail.com*

ARTICLE INFO

Article history:
Received: 08 /08/2024
Rrevised form: 11 /09/2024
Accepted : 10 /11/2024
Available online: 30 /12/2024

Keywords:

optimization , identification , classification , machine learning , network analysis , cybersecurity , network threats

ABSTRACT

Dark web emerged as an encrypted and closed network that can only be accessed by computers using specific software and allows users to apply for membership and you have to provide your IDs to the admin of that network, which creates severe challenges in passing cybersecurity; With advanced Internet facilities, the origin and possibility of network invasions or attacks are increasing, and it is becoming very difficult for traditional anomaly detection systems to analyze objectively and pass Functional Purpose: Our ultimate goal by applying machine learning algorithms and network analysis methods is to identify a certain set of attributes that May indicate a cyber threat. Like many other studies, the traffic in the current paper has also been generalized to represent darknet traffic with a classification of unique features between attacks and attacks. As is known, based on the results of the experimental analysis of the proposed approach, it is possible to conclude about the high efficiency of the developed method in identifying types of cyber threats . This project contributes, by utilizing and improving machine learning techniques, to finding reliable measures to mitigate threats in cyberspace, and improving threat detection programs.

MSC..

https://doi.org/10.29304/jqcsm.2024.16.41779

1. Introduction

One part of the Internet that is easily accessible also has a hidden part that a person has to have a pass to access became popular for mostly negative reasons. Complex only by the invitation of the use of the third parties such as Tor, I2P and Freenet, and the darknet is has limited time for anonymity and privacy and is therefore suitable for both Legal and Illegal information. These networks safeguard rights of individuals including privacy, speech and expression freedom; however, those networks are frequently used in crime activities as distribution of bhang, prohibited weapons and communication of cyber-attacks [1].

The fact that traffic to darknet sites is difficult to track makes it even more slightly different from the surface web and is even more difficult for any level of cybersecurity expert to monitor and address threats inherent in this environment which means that are require special approaches. This paper defines categorization of malicious

∗Corresponding author

Email addresses:

Communicated by 'sub etitor'

actions including the analysis of actions like categorizing in the darknet traffic stream as being of significant importance among the experts and scholars working in the domains of cyber security. Some of the protectant measures and tools are used by the network infrastructure to guard against numerous forged and well concealed darknet networks sometimes become ineffective against the complex modern day sophisticated networks. As a result, more novel proposals have emerged as complements available to confront and categories the existing wicked conducts happening in this obscure area of cyberspace. Some recent studies have discussed about on how this problem can be solved from some different way.

Other measures as currents such as network analysis have also been employed to build on the identification of unlawful events and other events for instance command and control functions, black markets and data breaches [2].

According to the main objective of the proposed paper, increasing the availability of better techniques to categorize the malicious activity of traffic within the darknet environment is the major goal of this paper. It is about how socially engineered traffic data, and the machine learning models, the deep learning frameworks, and also the network analysis tools are used to develop a encompassing proposition of the traffic data obtained from the darknets. More specific objectives of our study are given below: It is crucial to demarcate features that can be utilized to distinguish threat classes as high-precision methods in threat detection require isolation of features that provide minimal overlap between threats and non-threats. This means that there is a great need to emphasize the importance of the provided study in improving current strategies in cybersecurity and augmenting its methods such as monitoring the darknet space in real-time and counteracting potential cyber threats. In this way, through having better technologies which allow us to identify and classify such activities, we can safeguard people, individuals, organizations, and ourselves as a country from the rising numbers and severity of cyber threats.

2. Darknet and how access it?

It is a part of the Internet that is not indexed by traditional search engines, such as Google or Bing, and cannot be accessed using regular browsers. The dark web consists of websites and virtual communities that operate via encryption protocols, making them difficult to identify or access.

Tor use was examined by Owen and Savage (2016) more intensively using the data Mining approach that revealed normal usage as depicted in figure 1 and shown behavior differentiating normal and malicious Tor usage [6].

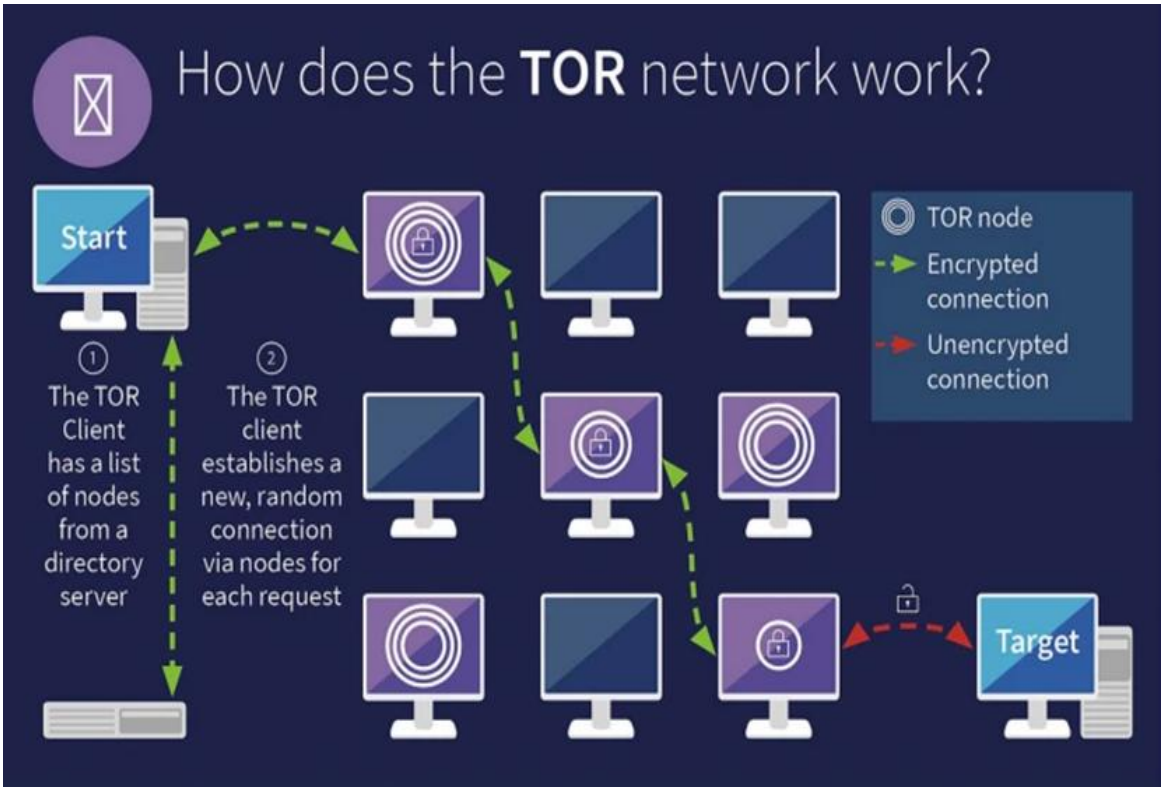

Fig 1: provides instructions for gaining access to the darknet, deep web, and Tor network.

3. Methodology

This section of the paper outlines the methodology we employed to examine darknet traffic and pinpoint the specific categories of activity. This conceptual framework combines machine learning models, deep learning neural networks, and network analysis to analyze major data sets of the DarkNet traffic. The mentioned approach aims at searching for some sort of activities which resemble a malicious behavior.

3.1. Data collection:

Data collection is the first step in our methodology since this step is important, its details are provided as follows: To collect darknet traffic data we used public datasets obtained through from various sources and also data we have been sharing with cybersecurity companies. The datasets contain unfiltered network traffics taken at different points in time, hence providing a large and diverse data.

3.2. Data Preprocessing:

The pre-processing stage itself is very significant and imperative to get the proper results as well as for the purification of the actual data utilized. This step includes:

- **Data Cleaning:** Data condensation to some of the most relevant sets that rules out non-significant and noisy data sets. Others are however, removing the non threat traffic as well as the repeated and unsampled records normally conveyed to the system.

- **Feature Extraction:** Data analysis and feature selection from the cleaned dataset: It is among the crucial steps one can undertaker when developing the model. In distinguishing between normal behavior and cyberattacks, such parameters as IP address, the port number, the respective size of packets and the time of packet transmission are important. Liu et al., (2018) highlights that feature extraction is among the most basic methods of enhancing the accuracy of machine learning models [9].

- **Machine Learning Algorithms :** The types of Machine learning includes that we have applied for classification of traffic information are enlisted below: Some of the Supervised learning algorithms used for developing the prediction Models are called as Support Vector Machine, Random Forest, K nearest neighbor, etc. This is due to the reasons that these algorithms have been trained in training sets in which traffic is described either as normal or an attack.

- **Support Vector Machines (SVM) :** Les Classifier proposed dans cette contribution utilization des Machines a Vectors de Support pour distingue entre le traffic attaquant et le traffic normal a l'aide du hyperplane le plus appropriate. Hence, the name of this is method is that it is well suited for cases where there are large dimensions and few samples of data [10].

- **K-Nearest Neighbors (k-NN):** k-NN is a majority vote process through which data points are classified and, is effective method for pattern recognition in the traffic of the networks [10].

- **Random Forest:** This learning is an aggregation where several decision trees are developed and are put together to ensure that the stability of the specific model and the confident level has been adjusted. This [_variability_ Reduction] is an important tool in effect of eliminating variability and prejudice on data [11].

- **Deep Learning Models** As a result of that, apart from the conventional approaches that utilize the machine learning algorithms, we have used deep learning models to conduct analysis on the darknet traffic to identify various patterns with high complexity. The two normally utilized structures applied Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs).

- **Convolutional Neural Networks (CNNs):** Such broadcast protocols are very useful especially in network traffic.as noticed by Berman et al. in 2019, would make it possible to use CNNs for classification of data regard as reverse-associated, possessing complex patterns of illegitimate and offending nature or malicious. [12].

- **Recurrent Neural Networks (RNNs):** LSTM is abbreviation for Long Short Term Memory; type of recurrent neural networks which are quite effective in training those networks which have sequence data and which have ability to remember data for long time. were applied in they allowed for capturing temporal dependencies in traffic data. [13].

- **Network analysis:** Network Analysis protocols, which are forms of network analysis, were employed for obtaining the network dimension of the darknet traffic. Network analysis was created to highlighting clusters and anomalies. Comparing his work with other scholars Owen and Savage (2016) equally showed that it is possible to use efficient network analysis in estimation of the nature of traffic in the darknet.

4. Results

We notice from the research results after dividing the dataset for testing and training data that there is a significant improvement in identifying and discovering unwanted activities on the dark web, which led to a noticeable increase in accuracy.

The techniques used have proven successful in distinguishing and classifying dark web data, as shown in the following table 1:

Metric	Detection	Classification
Accuracy	**96.4%**	**94.3%**
Precision	**94.5%**	**87.5%**
Recall	**93.2%**	**91.4%**
F1-scor	**0.91**	**0.88**

Table1. results of the detection and classification stages

In this context, the approach enables the identification of threats and their consequences in their capacity as threats, and minimization of these consequences. These frameworks use musically trained models that elevate them to the level of versatile and applicable in various forms of cyber threats.

5. Challenges

The general applicability of the machine learning algorithms to differentiate and classify various malice activity can have a few challenges that are stated below: Some common challenges include:

5.1. Imbalanced Data:

It as mentioned that the testing datasets in most of the real-time scenarios indicate that the number of instances of the malicious and non-malicious class is not balanced.

This may entail that the models built using the training data overemphasis the minority class, and give the majority class total disregard.

5.2. Lack of Labeled Data:

Supervised learning method require annotated data, which help them in solving some problems.

However, getting labeled data for inspecting malicious activity might take a lot of time and resources since data labeling is sensitive and requires expertise.**Evo**

5.3. Evolving Attack Techniques:

Cybercriminals never leave their guard and are always on the lookout for finding ways that would help them get past the security of any network this is a challenge in early detection as the models that employ machine learning training need to be updated and retrained as new tactics are discovered.

5.4. Feature Engineering:

Appropriate feature selection as an input to a machine learning model is one of the most critical steps since irrelevant features can significantly impact model performance, Additionally, it could be difficult to determine which features are most important when it comes to the detection and classification of the malicious activity, Expert knowledge of the domain and understanding of the exact threats and their vectors is necessary to perform feature engineering properly.

5.5. Generalization and False Positives:

Overfitting can arise when the models are complex or the data used in the training phase are insufficiently diverse other concerns with machine learning models include high false positives rates and a model's ability to apply well on unseen data.

Measures of overfitting include increasing model complexity, using Higher-order regularizations, cautious cross-validation, and accurate model selection process is normally effective in overcoming this issue.

5.6. Interpretability and Explainability:

Understanding and reasoning about the choices made by a machine learning algorithm is of high importance in order to establish trust Some of the models are hard to understand, For instance, deep neural networks Understanding these models' working remains challenging Addressing .

These issues show how hastily the various aspects of the dataset handling, algorithm selection, as well as model fine-tuning when framed in an ML setting relevant to categorization of the different types of malicious actions should be approached. This is why the current practices should be made current and the upcoming better practices should be sought advice from cybersecurity professionals when handling these issues.

6. Dataset and Test Configuration

For the identification of security threat and categorization of current nefarious undertakings contained in the dark web containers, we will extract a dataset for testing :

6.1. Dataset:

A dataset of dark web traffic is acquired, which is generally an example of the traffic in the dark web that is obtained from various stations of traffic within the dark web. It is important to make sure that there are enough varieties of classified types of malicious activities such as malware, bot-related communications, sharing of unsanctioned content and so on non-malicious traffic samples should also be included for reference and verification from the given data, It should be guaranteed that the data set chosen is diverse enough to train a reliable model.

6.2. The Test

it is required to organize a testing environment to check its efficiency and the model's one.Split a given datasets to a training, validation and a test datasets. Here the training data is employed to upgrade the model, the valid data is used in order to control the modeled data parameters and tuning and at last the test data is used to measure the performances of only last modeled data.
They should also describe compute metrics which they will be using in assessing the performance of the model, these include the accuracy, precision, recall and F1-scores.

Select a suitable ML/ DL technique such as Neural network, Random Forest, Support vector machines of any type of model building algorithm which is suitable for the model to be worked upon or built.

Re-estimate the model using various techniques such as cross-validation, re-tuning/ adjusting of hyperparameters, and applying different methods of regularization, enhance the scalability of the model for accurately predicting the malicious activities on the given darknet traffic classification; Lastly, assess the evaluated model with the help of the separate testing data set.

In my opinion, one would be justified to state that, after determining both, the dataset and the correct testing approach, one might come up with fairly adequate solution with the goal of differentiating and categorizing criminal transactions in darknet. As much as it is proper to uphold ethicality, and compliance when offering such tasks in charge of data privacy and security it is equally vital to follow polices too.

to find the right model which is why; Machine Larning is the most appropriate in enhancing the aspect of the model.

7. conclusions

Studying advancements in the detection and identification of malicious activity in the dark web field is investigational in estimating the possibility of increasing the efficiency of cybersecurity. Here are the key points: Secure coverage: That is, using, for instance, machine learning and deep learning, one has effective frameworks for detecting these malicious activities. different machine learning methods, are the most suitable ones to reveal suspicious patterns and behavior of the dark web data. Visiting the sites located at the dark web and engage in mining to classify the pages containing illicit activities enables the monitoring of the illicit activities. This facilitates the determination of patterns as well as the most recurrent forms of unlawful conducts observed in the dark web. Applying execution-based methods of detection and transfer types of learning enhances the effectiveness of detecting web-borne malware's and other dangerous activities as well. These techniques help in the prevention of actualization of threats in as much as they can be foreseen. The most important difficulty is the identification of sites because the dark web is in constant evolution and the addresses are not correlated with the identities of the users. To overcome these challenges, optimization-based approaches are used that proposed flexible and expansive solutions that can be modified according to the new architecture of the dark web. The subsequent studies must emphasize increased integration of intricate AI and machine learning algorithms to optimize the approaches used for detection purposes. It also presented that cooperation between academic researchers and industry professionals in the case of security threats and modern technologies can also improve the work on creating more complex and effective cybersecurity tools. Therefore, the proposed methodology based on utilizing an optimization model for the identification and detection of suspicious activities in the context of the dark web is found to offer substantial advantages in the establishment of enhanced security solutions. Therefore, application of reinforced, intelligent, and recent methods of computation can help researchers create new strategies that involve better and improved approaches that will counter ever-escalating dangers in the shadowiest areas of the dark-web.

References

[1]Ansh, S., & Singh, S. (2022, September). Analyze Dark Web and Security Threats. In International Conference on Innovations in Computer Science and Engineering (pp. 581-595). Singapore: Springer Nature Singapore.

[2]Swartout, K. M., de Heer, B. A., Thompson, M. P., Zinzow, H. M., & Brennan, C. L. (2022). A cross-disciplinary review of empirical studies addressing repeat versus time-limited sexual violence perpetration. Engaging Boys and Men in Sexual Assault Prevention, 411-422.

[3]Ali, M. I., & Kaur, S. (2021, February). The Impact of India's Cyber Security Law and Cyber Forensic On Building Techno-Centric Smartcity IoT Environment. In 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 751-759). IEEE.

[4]Durga, S., Nag, R., & Daniel, E. (2019, March). Survey on machine learning and deep learning algorithms used in internet of things (IoT) healthcare. In 2019 3rd international conference on computing methodologies and communication (ICCMC) (pp. 1018-1022). IEEE.

[5]Berman, D. S., Buczak, A. L., Chavis, J. S., & Corbett, C. L. (2019). A survey of deep learning methods for cyber security. Information, 10(4), 122.

[6]Frank, R., & Mikhaylov, A. (2020). Beyond the 'Silk Road': Assessing illicit drug marketplaces on the public web. Open Source Intelligence and Cyber Crime: Social Media Analytics, 89-111.

[7]Howell, C. J., Fisher, T., Muniz, C. N., Maimon, D., & Rotzinger, Y. (2023). A depiction and classification of the stolen data market ecosystem and comprising darknet markets: a multidisciplinary approach. Journal of Contemporary Criminal Justice, 39(2), 298-317.

[8]Brewer, R., Westlake, B., Hart, T., & Arauza, O. (2021). The ethics of web crawling and web scraping in cybercrime research: Navigating issues of consent, privacy, and other potential harms associated with automated data collection. Researching cybercrimes: methodologies, ethics, and critical approaches, 435-456.

[9]Arunkumar, M., & Kumar, K. A. (2023). GOSVM: Gannet optimization based support vector machine for malicious attack detection in cloud environment. International Journal of Information Technology, 15(3), 1653-1660.

[10]Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 10, 99129-99149.

[11]Miranda-Vega, J. E., Rivas-López, M., & Fuentes, W. F. (2020). K-nearest neighbor classification for pattern recognition of a reference source light for machine vision system. IEEE sensors journal, 21(10), 11514-11521.

[12]Bianchi, F. M., Maiorino, E., Kampffmeyer, M. C., Rizzi, A., & Jenssen, R. (2017). Recurrent neural networks for short-term load forecasting: an overview and comparative analysis.

[13]Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586.

[14]Rodriguez, E., Otero, B., Gutierrez, N., & Canal, R. (2021). A survey of deep learning techniques for cybersecurity in mobile networks. IEEE Communications Surveys & Tutorials, 23(3), 1920-1955.