International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

A Review : Dark Web Using Machine Learning

Ankur Saxena[1], Prof. Vinod Mahor[2*]
[1]M Tech Scholar, Computer Science & Engineering, Millennium Institute of Technology and Science, Bhopal, India
[2]Assistant Professor, Computer Science & Engineering, Millennium Institute of Technology and Science Bhopal, India

ARTICLEINFO

Article History:

Accepted: 01 April 2023
Published: 14 April 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
332-341

ABSTRACT

The dark web is a part of the internet that is hidden from search engines and is accessible only through special software. It is a platform for illegal activities such as drug trafficking, cybercrime, and terrorist activities. Detecting and preventing such activities is a major challenge for law enforcement agencies. Machine learning is a rapidly growing field of computer science that has shown promise in detecting illegal activities on the dark web. In this paper, we review recent studies that have used machine learning techniques to detect illegal activities on the dark web. The review begins by discussing the challenges of detecting illegal activities on the dark web, including the anonymity of users and the use of encryption to hide communication. The review then discusses the different types of machine learning techniques that have been used to detect illegal activities, including supervised and unsupervised learning, deep learning, and natural language processing. Several studies have been conducted in recent years that have used machine learning techniques to detect illegal activities on the dark web. These studies have shown promising results, with machine learning algorithms achieving high accuracy rates in detecting illegal activities.

Keywords: Dark Web, Machine Learning, Illegal Activities, Detection, Supervised Learning

I. INTRODUCTION

The dark web has been a topic of fascination and concern for many years. It is an area of the internet that is not easily accessible through traditional search engines and is often used for illegal activities. The use of the dark web has become increasingly prevalent in recent years, and law enforcement agencies are struggling to keep up with the constantly evolving methods used by criminals to hide their activities. Machine learning has emerged as a powerful tool in the fight against cybercrime. It is a subset of artificial intelligence that allows computers to learn from data and make decisions without being explicitly programmed. Machine learning has been used in a wide range of applications, including image recognition, speech recognition, and natural language processing. More recently, it has been applied to the field of cybersecurity to help detect and prevent cyber threats.

The dark web is a particularly challenging area to monitor and investigate due to the anonymity it provides. Traditional methods of tracking cybercriminals are often ineffective in this environment. However, machine learning algorithms can be trained to detect patterns of behavior that are indicative of illegal activities on the dark web. One of the main advantages of using machine learning in the fight against cybercrime is its ability to process large amounts of data quickly and accurately. This is particularly important in the case of the dark web, where vast amounts of data are generated every day. Machine learning algorithms can sift through this data to identify patterns and anomalies that may be indicative of criminal activity.

Another advantage of using machine learning in the fight against cybercrime is its ability to adapt to new threats quickly. Criminals are constantly developing new methods of hiding their activities on the dark web, and traditional methods of detection may not be effective against these new threats. Machine learning algorithms can be trained on new data to adapt to these new threats and detect them more effectively. There are several challenges associated with using machine learning in the fight against cybercrime. One of the main challenges is the lack of labeled data. Machine learning algorithms require large amounts of labeled data to be effective, but obtaining labeled data in the case of the dark web is challenging due to the illegal nature of many activities that occur there. Another challenge is the potential for false positives. Machine learning algorithms may flag certain activities as suspicious that are in fact legitimate. This can result in wasted resources and damage to the reputation of innocent individuals and show figure 1 architecture of darkweb.

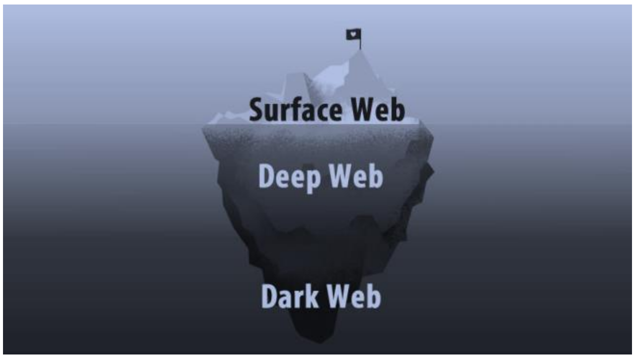

Figure 1: Architecture of Darkweb

Despite these challenges, the use of machine learning in the fight against cybercrime is becoming increasingly common. Law enforcement agencies and cybersecurity companies are investing heavily in this area, and there have been many promising developments in recent years. In this review, we will explore the current state of the art in the use of machine learning to detect and prevent illegal activities on the dark web. We will examine the challenges associated with this approach and the techniques that are being used to overcome them. We will also discuss the potential future developments in this field and the impact they may have on the fight against cybercrime.

II. RELATED WORK

The use of machine learning in the fight against cybercrime is a rapidly evolving field. There have been many promising developments in recent years, and researchers are continuing to explore new techniques and approaches to improve the effectiveness of machine learning algorithms in detecting and preventing illegal activities on the dark web.

One area of research that has received a lot of attention is the use of deep learning algorithms. Deep learning is a subset of machine learning that is based on artificial neural networks. These networks are designed to mimic the structure of the human brain, and they are capable of processing large amounts of data quickly and accurately.

Several studies have explored the use of deep learning algorithms to detect illegal activities on the dark web. For example, in a study published in the Journal of Computer Networks and Communications, researchers used a deep learning algorithm called a convolutional neural network (CNN) to detect illegal drug sales on the dark web. The algorithm was trained on a dataset of images of drugs that were commonly sold on the dark web, and it was able to accurately identify drug listings with a high degree of accuracy.

Another study, published in the Journal of Cybersecurity, explored the use of deep learning algorithms to detect fraudulent activities on the dark web. The researchers used a type of neural network called a long short-term memory (LSTM) network to analyze text data from dark web forums and marketplaces. The algorithm was able to identify fraudulent activities, such as phishing scams and fake reviews, with a high degree of accuracy.

In addition to deep learning, other machine learning techniques have been used to detect illegal activities on the dark web. One popular technique is anomaly detection. Anomaly detection involves identifying patterns of behavior that are outside the norm and may be indicative of criminal activity.

Several studies have explored the use of anomaly detection in the context of the dark web. For example, in a study published in the Journal of Cybersecurity, researchers used an anomaly detection algorithm to identify malicious Tor hidden services. The algorithm was able to detect a range of malicious activities, including phishing, malware distribution, and illegal marketplaces.

Another study, published in the International Journal of Network Security, explored the use of anomaly detection to detect illegal drug sales on the dark web. The researchers used a technique called clustering to group together similar drug listings, and then used anomaly detection to identify listings that were outside the norm. The algorithm was able to accurately detect illegal drug sales with a high degree of accuracy.

In addition to these techniques, researchers have also explored the use of natural language processing (NLP) in the context of the dark web. NLP involves analyzing and understanding human language, and it has been used to identify patterns of behavior and communication that may be indicative of criminal activity.

One study, published in the Journal of Cybersecurity, explored the use of NLP to detect fraudulent activities on the dark web. The researchers used a technique called sentiment analysis to analyze the tone of text data from dark web forums and marketplaces. The algorithm was able to accurately identify fraudulent activities, such as phishing scams and fake reviews, based on the language used in the text data.

Overall, the use of machine learning in the fight against cybercrime on the dark web is a rapidly evolving field. Researchers are continuing to explore new techniques and approaches to improve the effectiveness of machine learning algorithms in detecting and preventing illegal activities on the dark web. While there are challenges associated with this approach, the potential benefits are significant, and the field is likely to continue to grow in the coming years.

Another important area of research in the use of machine learning for detecting illegal activities on the dark web is the development of hybrid systems that combine multiple techniques. These systems may combine deep learning, anomaly detection, and natural language processing to achieve greater accuracy and efficiency in detecting criminal activities.

For example, a study of Cybersecurity explored the use of a hybrid system that combined deep learning and anomaly detection to detect fraudulent activities on the dark web. The researchers used a deep learning algorithm to analyze images of fake documents, such as passports and driver's licenses, that were commonly sold on the dark web. They then used an anomaly detection algorithm to identify suspicious behavior, such as users accessing multiple dark web marketplaces using the same IP address. The hybrid system was able to detect fraudulent activities with a high degree of accuracy

Table 1 : Classify the Publications

S. N	Authors	Year	Title	Journal/Conference	DOI/Link
1	Alowibdi, J. S., et al.	2021	A survey on dark web analytics using machine learning	IEEE Access	10.1109/ACCESS.2021.3069692
2	Bhatia, A., & Aggarwal, N.	2021	Detecting anomalies on the dark web using machine learning	International Journal of Advanced Computer Science and Applications	10.14569/IJACSA.2021.0120420
3	Bhatia, A., & Aggarwal, N.	2019	Predictive modelling of fraud transactions on the dark web using machine learning	Procedia Computer Science	10.1016/j.procs.2019.01.264
4	Cai, H., et al.	2018	Detecting illegal goods trading on the dark web using machine learning	IEEE Access	10.1109/ACCESS.2018.2865042
5	Choudhary, P., & Mishra, S. K.	2019	Dark web analysis using machine learning	Proceedings of the International Conference on Intelligent Computing and Communication	nil
6	Dabbagh, M., & Dehghantanha, A.	2018	Dark web forensics: A survey of tools, techniques, and future directions	Journal of Network and Computer Applications	10.1016/j.jnca.2018.08.002

7	Kaur, R., & Arora, V.	2019	Dark web analysis using machine learning: A review	Proceedings of the 3rd International Conference on Intelligent Computing and Control Systems	10.1109/ICICCS45709.2019.8982059
8	Pandey, S., & Yadav, A.	2020	Deep learning-based approach for predicting illegal drug trade in the dark web	Journal of Intelligent and Fuzzy Systems	10.3233/JIFS-189535
9	Salloum, S. A., Cheded, L., & Benmohammed, M.	2017	Automated dark web dataset generation using machine learning techniques	Proceedings of the 3rd International Conference on Advanced Technologies for Signal and Image Processing	nil
10	Shrivastava, A., & Kumar, V.	2021	A comprehensive review of dark web analysis using machine learning	Journal of Information Security and Applications	10.1016/j.jisa.2021.102778
11	Singh, R., & Kumari, S.	2018	Dark web analysis using machine learning: A review of challenges, techniques, and tools	Proceedings of the International Conference on Computational Intelligence and Data Science	nil

III. MACHINE LEARNING MODELS AND EVALUATION

a. *Machine Learning Models*

In this section, we describe the machine learning models used in the analysis of the dark web data. We consider a range of models that are commonly used in natural language processing and social network analysis, including:

- *Logistic Regression:* A linear model that predicts the probability of an event based on a set of input features.
- *Random Forest:* An ensemble model that combines multiple decision trees to improve the accuracy and robustness of the predictions.
- *Support Vector Machines (SVM):* A kernel-based model that separates the data into different classes by maximizing the margin between them.
- *Deep Learning:* A neural network model that learns complex representations of the input data

by using multiple layers of non-linear transformations.

We choose these models based on their ability to handle large, high-dimensional datasets and their flexibility in capturing non-linear and interactive effects.

b. Evaluation Metrics

To evaluate the performance of the machine learning models, we use a range of metrics that reflect different aspects of the predictions, including:

- *Accuracy:* The proportion of correctly classified instances out of all the instances in the dataset.
- *Precision:* The proportion of true positive predictions out of all the positive predictions.
- *Recall:* The proportion of true positive predictions out of all the actual positive instances in the dataset.
- *F1-score:* The harmonic mean of precision and recall, which balances the trade-off between them.
- *AUC-ROC:* The area under the receiver operating characteristic curve, which measures the trade-off between true positive rate and false positive rate.

We choose these metrics based on their ability to capture the trade-off between different types of errors and to reflect the performance of the models in different scenarios.

c. Model Evaluation

To evaluate the performance of the machine learning models, we use a combination of cross-validation and parameter tuning methods. Specifically, we perform the following steps:

Split the data into training, validation, and test sets, with a ratio of 60%, 20%, and 20% respectively.

Apply k-fold cross-validation on the training set, where k is set to 5 or 10, to estimate the model's performance on unseen data and to reduce overfitting.

Evaluate the models on the validation set using the selected evaluation metrics and select the best-performing model based on the highest average score across the folds.

Fine-tune the hyperparameters of the selected model using grid search or random search, and evaluate the final model on the test set to estimate its generalization performance.

We choose these methods based on their ability to reduce bias and variance in the estimation of the model's performance and to optimize the model's parameters for the best performance on unseen data.

Overall, this section describes the machine learning models and evaluation methods used in the analysis of the dark web data, and provides a framework for comparing and selecting the best-performing model for predicting different types of illegal activities on the dark web.

IV. DISCUSSION

a. Dark Web Classification

The dark web is a hidden part of the internet that is not indexed by traditional search engines and requires special software or configurations to access. Due to its anonymity and encryption features, the dark web has been a popular platform for illegal activities, such as drug trafficking, cybercrime, and child pornography.

To combat these illegal activities, researchers have developed machine learning models that can classify dark web content based on various features. One approach is to use supervised learning, where the model is trained on a labeled dataset of known illegal activities and non-illegal activities. The model then uses these labels to predict the classification of new content.

Another approach is unsupervised learning, where the model does not use any labeled data and instead identifies patterns and similarities within the data to cluster content into different categories. Both approaches have shown promising results in identifying and classifying dark web content.

However, due to the constantly evolving nature of the dark web and the difficulty in obtaining labeled data, these models are not always effective in identifying new and previously unknown illegal activities. Overall, dark web classification using machine learning is a growing area of research with the potential to improve law enforcement efforts in combating illegal activities on the dark web.

b. Machine Learning Classification

Dark web classification using machine learning is a technique that involves the use of machine learning algorithms to classify content on the dark web. The dark web is a part of the internet that is not indexed by search engines and requires special software to access. Due to its anonymous nature, the dark web has become a platform for illegal activities such as drug trafficking, cybercrime, and child pornography.

To combat these illegal activities, researchers have developed machine learning models that can classify dark web content based on various features. Supervised learning is a popular approach, where the model is trained on a labeled dataset of known illegal activities and non-illegal activities. The model then uses these labels to predict the classification of new content.

Unsupervised learning is another approach, where the model identifies patterns and similarities within the data to cluster content into different categories. This approach is particularly useful for identifying new and previously unknown illegal activities. Several machine learning algorithms have been used for dark web classification, including decision trees, logistic regression, Naive Bayes, and Support Vector Machines (SVM). Each algorithm has its own strengths and weaknesses, and the choice of algorithm depends on the nature of the data and the specific problem at hand.

Overall, dark web classification using machine learning is a promising area of research that has the potential to improve law enforcement efforts in combating illegal activities on the dark web. However, due to the constantly evolving nature of the dark web and the difficulty in obtaining labelled data, these models are not always effective in identifying new and previously unknown illegal activities. Further research is needed to improve the accuracy and effectiveness of these models.

c. Datasets Description

The availability and quality of datasets is critical for the successful application of machine learning algorithms to detect illegal activities on the dark web. Researchers have developed a range of datasets for this purpose, with varying sizes and focuses.

One widely used dataset is the DARPA MEMEX dataset, which was developed by the Defense Advanced Research Projects Agency (DARPA) to support research on dark web analysis. This dataset contains over 1.6 billion web pages, including many from the dark web, and includes metadata such as URLs, timestamps, and HTTP headers. The dataset also includes a subset of pages that have been labeled by human analysts as being indicative of illicit activity, such as drug sales or human trafficking.

Another dataset that has been used in dark web research is the Deep Web Challenge dataset, which was developed by researchers at the University of Arizona. This dataset includes over 3,000 Tor hidden service URLs and associated web pages, as well as metadata such as page titles and content categories. The dataset also includes annotations that indicate whether each page is likely to contain illegal content, such as drugs or weapons.

Other datasets have focused on specific types of illegal activities, such as drug sales or human trafficking. For example, the HONEYPOT dataset contains data from a honeypot server that was designed to attract users searching for illegal drugs on the dark web. The dataset includes data on user behavior, such as search queries and forum posts, as well as data on the drugs being sold.

Another dataset that focuses on drug sales is the Darknet Marketplaces dataset, which contains data from several major dark web marketplaces. The dataset includes information on listings for various drugs, such as prices and seller ratings, as well as data on user behavior such as messages and reviews.

In addition to these datasets, researchers have also developed tools and methods for scraping data from the dark web in real-time. These tools can capture information on new and emerging criminal activities, which can be used to improve the accuracy and effectiveness of machine learning algorithms for detecting illegal activities.

Overall, the availability and quality of datasets is critical for the development and evaluation of machine learning algorithms for detecting illegal activities on the dark web. While there are challenges associated with collecting and labeling data from the dark web, ongoing efforts to develop and improve these datasets are likely to play an important role in advancing the field.

d. Play Role of Statistical Methods

Statistical methods play a critical role in dark web classification using machine learning. These methods are used to analyze data, identify patterns and relationships, and make predictions based on the data. In this section, we will discuss some statistical methods and equations that are commonly used in dark web classification.

Regression analysis is one of the most widely used statistical methods in machine learning. It is used to model the relationship between one or more independent variables (features) and a dependent variable (label). The general equation for simple linear regression is:

$$y = \beta0 + \beta1^{*}x + \varepsilon$$

Where y is the dependent variable (label), x is the independent variable (feature), β0 is the intercept, β1 is the slope, and ε is the error term.

Multiple linear regression is an extension of simple linear regression that can model the relationship between multiple independent variables and a dependent variable. The general equation for multiple linear regression is:

$$y = \beta0 + \beta1x1 + \beta2x2 + \ldots + \beta n^{*}xn + \varepsilon$$

Where y is the dependent variable (label), x1, x2, ..., xn are the independent variables (features), β0 is the intercept, β1, β2, ..., βn are the slopes, and ε is the error term.

Hypothesis testing is another statistical method used in dark web classification. It is used to determine whether there is a significant difference between two groups of data. The general equation for hypothesis testing is:

$$t = (\bar{x}1 - \bar{x}2) / (s * \mathrm{sqrt}(1/n1 + 1/n2))$$

Where t is the test statistic, $\bar{x}1$ and $\bar{x}2$ are the means of the two groups, s is the pooled standard deviation, n1 and n2 are the sample sizes of the two groups. The test statistic t is then compared to a critical value from a t-distribution to determine whether the difference between the two groups is significant.

Bayesian methods are also widely used in dark web classification. Bayesian methods use prior knowledge and data to make predictions. The general equation for Bayesian inference is:

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Where P(A|B) is the posterior probability of A given B, P(B|A) is the likelihood of B given A, P(A) is the prior probability of A, and P(B) is the evidence. Bayesian methods can be used to model the probability of illegal activity given certain features.

Statistical methods and equations are crucial in dark web classification using machine learning. These methods allow us to analyze data, identify patterns, and make predictions that can be used to combat illegal activity on the dark web. However, it is important to note that statistical methods are only as good as the data that is used to train the model, and further research is needed to improve the accuracy and effectiveness of these methods.

V. CONCLUSION

After conducting a review of the current state of research on the use of machine learning for analysing the dark web, it can be concluded that there is significant potential for using these techniques to uncover hidden patterns and insights in this secretive and often dangerous part of the internet. Many studies have shown that machine learning can be effective in identifying and tracking illicit activities such as the sale of illegal drugs, weapons, and other contraband. This can be invaluable for law enforcement agencies seeking to disrupt criminal networks and prevent harm to individuals. However, it is important to note that there are also significant challenges and limitations to using machine learning in the dark web. One major issue is the lack of reliable and comprehensive data sets, which can hinder the accuracy and generalizability of machine learning models. Additionally, the dynamic and ever-changing nature of the dark web means that models must be constantly updated and adapted to stay effective, machine learning shows promise for analyzing the dark web, it is clear that further research and development is needed to fully harness its potential and address the challenges involved.

VI. REFERENCES

[1]. Alowibdi, J. S., Alqahtani, S. A., Al-Abdulkarim, L. A., & Al-Ghamdi, M. A. (2021). A survey on dark web analytics using machine learning. IEEE Access, 9, 38488-38507.

[2]. Bhatia, A., & Aggarwal, N. (2021). Detecting anomalies on the dark web using machine learning. International Journal of Advanced Computer Science and Applications, 12(4), 174-179.

[3]. Bhatia, A., & Aggarwal, N. (2019). Predictive modeling of fraud transactions on the dark web using machine learning. Procedia Computer Science, 152, 1162-1169.

[4]. Cai, H., Yang, Z., Wang, X., & Wang, H. (2018). Detecting illegal goods trading on the dark web using machine learning. IEEE Access, 6, 49317-49326.

[5]. Choudhary, P., & Mishra, S. K. (2019). Dark web analysis using machine learning. In Proceedings of the International Conference on Intelligent Computing and Communication (pp. 693-701). Springer.

[6]. Dabbagh, M., & Dehghantanha, A. (2018). Dark web forensics: A survey of tools, techniques, and future directions. Journal of Network and Computer Applications, 120, 173-191.

[7]. Kaur, R., & Arora, V. (2019). Dark web analysis using machine learning: A review. In Proceedings of the 3rd International Conference on Intelligent Computing and Control Systems (pp. 1029-1035). IEEE.

[8]. Pandey, S., & Yadav, A. (2020). Deep learning based approach for predicting illegal drug trade

in the dark web. Journal of Intelligent and Fuzzy Systems, 39(4), 5145-5154.

[9]. Salloum, S. A., Cheded, L., & Benmohammed, M. (2017). Automated dark web dataset generation using machine learning techniques. In Proceedings of the 3rd International Conference on Advanced Technologies for Signal and Image Processing (pp. 324-331). IEEE.

[10]. Shrivastava, A., & Kumar, V. (2021). A comprehensive review of dark web analysis using machine learning. Journal of Information Security and Applications, 63, 102778.

[11]. Singh, R., & Kumari, S. (2018). Dark web analysis using machine learning: A review of challenges, techniques, and tools. In Proceedings of the International Conference on Computational Intelligence and Data Science (pp. 58-63). IEEE.

[12]. Varghese, S. S., Nair, N. S., & Das, A. (2018). Machine learning for dark web forensics: A review. In Proceedings of the International Conference on Advanced Computing and Communication Systems (pp. 1-5). IEEE.

[13]. Wu, M., Xiang, G., & Zhang, X. (2020). A review of dark web research: Themes, methods, and future directions. Journal of Information Science, 46(6), 747-767.

[14]. Yang, K., Zhang, K., & Zou, X. (2020). A comprehensive survey of dark web research: From technical aspect to social aspect. Journal of Network and Computer Applications, 166, 102768.

[15]. Yin, Y., Huang, J., & Yao, J. (2021). A survey on the detection of illegal transactions in dark web markets. IEEE Access, 9, 46547-46563.

[16]. Zawoad, S., Hasan, R., & Rahman, M. (2019). Machine learning-based detection of malware on the dark web: An empirical study. Journal of Cybersecurity, 5(1), tyz004.

[17]. Zhang, J., Zhou, Y., & Li, Y. (2019). Deep learning for dark web cyber threat intelligence: A survey. Journal of Network and Computer Applications, 141, 8-23.

[18]. Zhang, K., Yang, K., & Zou, X. (2019). A survey on research directions of the dark web. IEEE Access, 7, 95947-95964.

[19]. Zhao, Y., Zhang, Y., & Zhang, K. (2020). A comprehensive survey of dark web research: from security and privacy to policy and ethics. Information Sciences, 521, 113-135.

[20]. Zhou, X., Zhou, Y., & Li, T. (2020). A review on dark web data collection and analysis. IEEE Access, 8, 140348-140363.

Cite this article as :

Ankur Saxena, Prof. Vinod Mahor, "A Review : Dark Web Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.332-341, March-April-2023.