 doi:10.20944/preprints202110.0165.v1

Article

Influence of COVID-19 Epidemic on Dark Web Contents

Abdul Razaque [1]*, Bakhytzhan Valiyev [1],Bandar Alotaibi [2,3]*, Munif Alotaibi [4]*, Saule Amanzholova [1], Aziz Alotaibi [5]

[1] Department of Cybersecurity, IITU, Almaty 050000, Kazakhstan; a.razaque@edu.iitu.kz (A.R.); 24795@edu.iitu.kz (B.V.); s.amanzholova@iitu.edu.kz (S.A)
[2] Sensor Networks and Cellular Systems Research Center, University of Tabuk, Tabuk 71491 , Saudi Arabia; b-alotaibi@ut.edu.sa
[3] Department of Information Technology, University of Tabuk, Tabuk 71491, Saudi Arabia
[4] Department of Computer Science, Sharqa University, Sharqa, Saudi Arabia; munif@su.edu.sa
[5] Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944. Saudi Arabia , Saudi Arabia; azotaibi@tu.edu.sa
* Correspondence: a.razaque@edu.iitu.kz, b-alotaibi@ut.edu.sa, munif@su.edu.sa

Abstract: The Dark Web is known as a place triggering a variety of criminal activities. Anonymization techniques enable illegal operations, leading to the loss of confidential information and its further use as bait, a trade product or even a crime tool. Despite technical progress, there is still not enough awareness of the Dark Web and its secret activity. In this study, we introduced the Dark Web Enhanced Analysis (DWEA), in order to analyze and gather information about the content accessed on the Dark Net based on data characteristics. The research was performed to identify how the Dark Web has been influenced by recent global events, such as the COVID-19 epidemic. The research included the usage of a crawler, which scans the network and collects data for further analysis with machine learning. The result of this work determines the influence of the COVID-19 epidemic on the Dark Net.

Keywords: Dark Net; Dark Web; COVID-19; data collection

1. Introduction

The Dark Net, referred to as the Dark Web, gains more attention from individuals who are concerned about their online privacy, since it is focused on providing user anonymity [1]. The Dark Web concept has been used since the early 2000s [2], and there have been many studies on terrorism activities.

The study conducted by [3] shows that the most common concern for the people involved in technological platforms is widespread data collection. Thus, the drawback of regular web searching is not feasible for users of the Dark Net, since the websites' tracking ability faces certain anonymization obstacles. Connection to the network is performed by using special browsers. They are focused on onion routing use. One of the most popular browsers is the TOR browser [4]. The majority of users show legitimate behavior, as the study of [5] states that most oDark Web users may have never visited websites ending with ".onion" and have used it instead for secure browsing. This is also proven by the low percentage of network traffic, corresponding to the range of 6-7% [5] leading to those sites.

However, the Dark Net is also widely used for committing criminal acts, such as the distribution of prohibited products and trade of illegally captured data [7]. The publication of [8] claims that anonymized and free-of-identity platforms became a perfect place for contraband sales. The network analysis of [9] identified that most threats could be due to computer worms and scanning actions. Users who attempt to use this infrastructure for legitimate purposes may lack knowledge about the crime scenes and their features. According to [10], more than 33% of suspected criminal websites located on the hidden

 doi:10.20944/preprints202110.0165.v1

side of the Dark Net cannot be classified. This prompted us to perform a network scan, one of the main purposes of this work.

The Dark Net is not stable and is likely to change quickly. There are different websites being created and deleted every month. Due to the recent events taking place in the world, such as the COVID-19 pandemic, the content may have been influenced by certain variations. This study is focused on identifying and describing the state of the Dark Net content. The Dark Net content of 2018 was rapidly changed in 2020. This allows us to understand the kinds of services the Dark Web hosts, their level of criminality and the level of impact from global events.

The research was performed by using an optimized web-crawler for information collection. Furthermore, the websites were accessed and categorized based on the content.

Recently, there have been many research works studying the Dark Net [11] in the field of illicit drugs, by collecting vendor names. Furthermore, the Pretty Good Privacy (PGP) protocol is a secure method, but it has also been found to be vulnerable. Anonymization techniques have enabled drug trafficking and other illegal businesses. This condition was described by [12] and [13] as an innovation and progression of illegal activities in their works.

The research presented in [14] analyzed the content of the Dark Web by implementing a web-crawler and performed categorization of received data. However, due to the fast-changing environment, the content may differ and require more recent analysis.

A crawler is a searching script that visits web pages and collects information about them. The crawler produces a copy of visited pages and provides captured time information [15]. Although the Dark Net is thought to be resistant to penetration [16], most of it can be accessed with relative ease.

1.1. Research contribution

Motivated by performed works, the contributions of this paper are as follows:

- Gathering itemized analytical information about the websites using a crawler by accessing them, analyzing their content, and identifying their types.
- Classification of the websites by topic based on collected information, enabling a better understanding of the Dark Net.
- Application of data science, in particular, machine learning, to preserve the accuracy of results.

1.2. Paper organization

The remaining parts of the paper are as follows. Section 2 contains the identification of issues. Section 3 generally covers the key elements of the relevant studies. Section 4 depicts the system model. Section 5 describes the Dark Web Enhanced Analysis (DWEA) plan of the research. Section 6 describes the experimental results and setup. Section 7 contains the discussion of the implementation, including its advantages and shortcomings. Finally, Section 8 provides a general summary.

2. Problem Identification

The greatest concern is the shortage of knowledge on the structure of the Dark Net and its criminal use. It is not easily accessed by most users, and therefore, is not well known.

The Dark Net contains a huge number of websites that cannot be accessed by regular search engines. The websites are harder to find and are not subject to the influence of local government laws. This creates a perfect basis for criminal activities, since it is harder to track the perpetrators.

Furthermore, the network can react to events happening in the world, which makes it possible to investigate if recent occasions, such as epidemics, change its structure.

Scarce awareness of the Dark Web creates many false beliefs about it. This could lead to the inaccurate use of its resources, leading to an increase in victims, which results in the

 doi:10.20944/preprints202110.0165.v1

further distribution of illicit schemes. This situation may be cyclic, as the last point may influence the first point.

There are several possible ways of identifying the Dark Net content, for instance, web-browsing using dictionary filling of the website address. This method includes checking every possible combination of symbols in a sequential manner. Another method is using recursive links found on specific websites and following them. This method is based on connection principles of distinct websites. An optimistic solution involves advantages of the previously mentioned methods whilst avoiding their drawbacks.

3. Related Works

The salient features of existing methods are briefly explained in this section. Research conducted by Gwern et al. [17] used a crawler that scanned the Dark Net websites and collected a large amount of records. However, the crawling was handled on a weekly basis, which reduced its comprehensiveness. Moreover, the crawler did not repeat the page load once it received an error.

Demant et al. [18] performed crawling to identify purchase sizes instead of products being sold. However, the work experienced certain drawbacks, such as incomplete crawls and the inaccurate deletion of presented duplications. Munksgaard et al. [19] proposed the automatic collection of data in their work. However, their method could be affected by uncertainties during information collection.

Moore et al. [5] conducted a research studying cryptography improvement effects and Tor's practicality. The crawling process followed a list of certain addresses. However, due to repetitions in the list, there could be speed issues. The research work of Kalpakis et al. [20] contributed to a crawler looking for products, guides showing how to make explosive materials, and their distribution places. The crawler operates with websites connected to a given initial set of pages.

Pannu et al. [21] developed a crawler operating with unsafe website detection. It also uses a given set of initial websites by loading their HTML code and going through links present on them. Once the specific page is loaded, it is scanned, and the process repeats. Al Nabki et al. [16] conducted an attempt to analyze the Dark Web. Their work included a collection of valid pages in the hidden portion of the Dark Net with approaches based on classification principles. However, they did not perform the network scan in a recursive manner, as they went through the initially given pages. Fidalgo et al. [22] performed a research of criminal act identification by image analysis by using classification methods. It improves the scanning process, but leads to ethical issues due to the storage of illegal materials.

4. System Model

In this section, the structure of the crawling system is described. The crawler's task is to scan a certain part of the Dark Net by following the links found on already scanned pages. The crawler is initially given a set of addresses to start the scanning process from. The system consists of several processes, as depicted in Figure 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

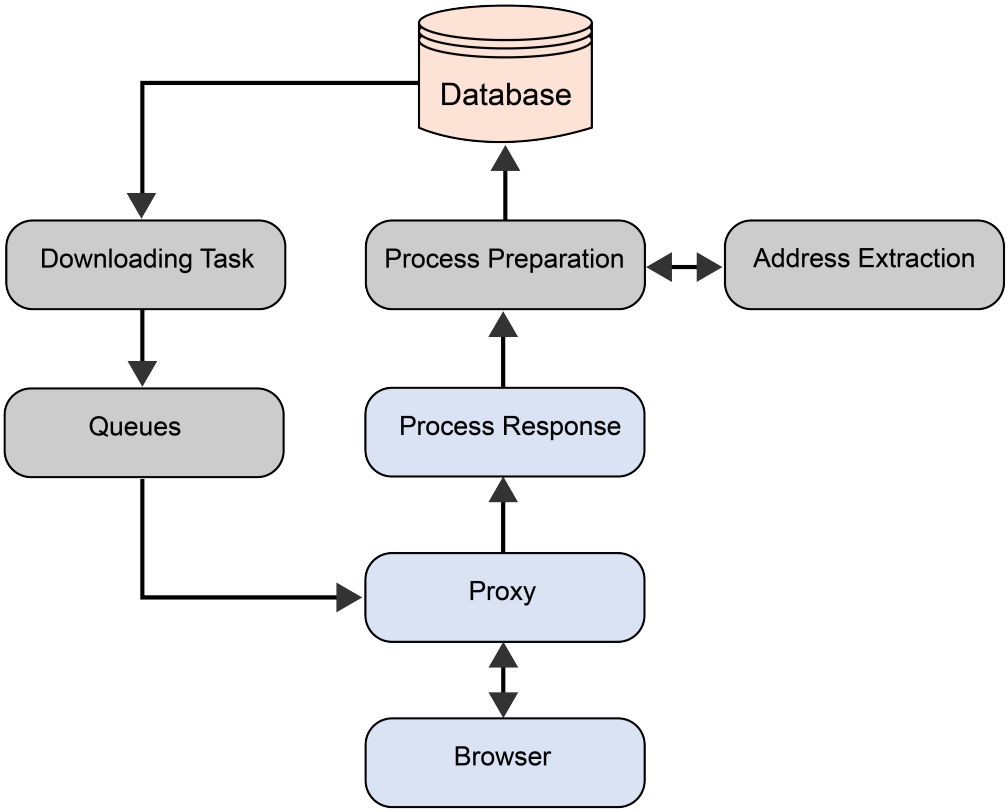


Figure 1. Proposed crawler architecture.

The crawling process starts from downloading tasks, which contain URL addresses. Once the process is completed, the task is to be sent to the proxy. The task is placed in a queue if there are other running tasks with the proxy. Connection to the network takes place through the proxy and browser. The system connects to the Dark Net by using the Tor browser and Tor proxy. It adds additional security and anonymity to the crawler by changing the source IP address.

Once the proxy returns an acknowledgement response, the response content is checked for the presence of illegal content. If the filter passes over the content, the page downloading completes and the content is attached to the result. In the next step, the result is sent to the preparation block. The preparation block sets the incoming data into a required state by leaving necessary information, such as page address and page content, and cutting explicit information.

It is important to note the ground purpose of filter usage. Dark Net anonymity principles enabled it to become an area of storing media, trading offers, etc., strictly prohibited in most of countries. Since there is a database component, which stores retrieved data, its storage could become a criminal act. Therefore, illegal content, such as child pornography, must be excluded from the collected information.

The database stores data collected during the scanning. Its information is frequently updated, and new data are constantly inserted into it. Some types of database management systems, e.g., column-oriented databases, do not work well in the mentioned conditions. This is a reason to select relational databases, which are more suitable for frequent changes. The database includes a table of URLs with the path, time of access, and state of scan, and a table of contents, which stores the content retrieved from the web pages.

A proxy is used in order to establish more secure communication. The second reason is escaping a situation when a website may suspect the crawler of a Denial of Service (DoS) attack during accessing many pages. A browser is used in terms of the Tor concept, since it

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

serves as the only entrance to the Dark Net system. Queues are included in the structure to prioritize page extractions and prevent the system from resource overuse.

5. Proposed Dark Web Enhanced Analysis Process

This section describes the proposed DWEA process, which includes algorithms referring to certain stages of the crawling process. As mentioned in the previous section, the process includes crawling the Dark Net and collecting information hosted on visited websites. The sequence diagram in Figure 2 describes the process.

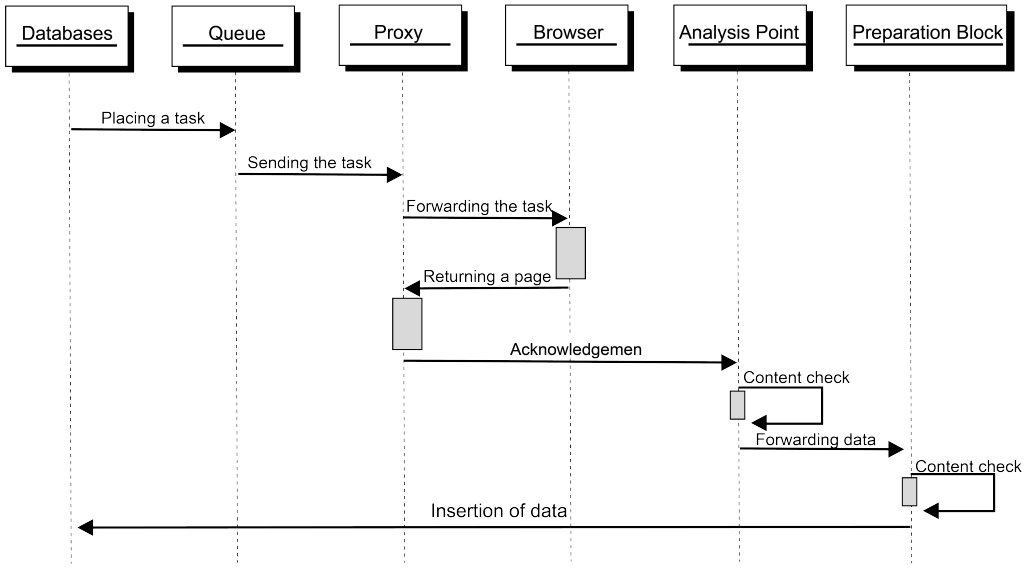


Figure 2. Sequence diagram of the system.

The process of scanning pages can be divided into the following components:

- Accessing the pages.
- Filtering the traffic.
- Classifying the pages.

5.1. Accessing the pages

Crawlers are frequently used in various cases, especially in search engines. Their main goal is to retrieve the newest information by copying pages for later operations. Web pages are scanned for the presence of certain types of information, such as harmful data or specific topics. The functional process of accessing the list of pages is explained in algorithm 1.

In algorithm 1, the accessing process of the list of pages is discussed. Step 1 initializes variables that the algorithm uses. A description of the values sent to the algorithm and declaration of the resulting value are shown at the beginning of the algorithm. Step 2 shows the process of adding initial URLs in the array. Step 3 starts the scanning process of all URLs. Step 4 sets a certain URL of the URL array. Step 5 checks if the webpage was not already scanned. Steps 6-8 reopen the page in case a page loading was not successful. Step 9 indicates the condition of successful page opening. Step 10 represents the crawler scanning a certain page. Step 11 adds the page to the list of scanned pages. Step 12 excludes recently scanned pages from the list of pages to be scanned.

Figure 3 represents the timing diagram of the crawler's scanning process. It shows how many pages the crawler scanned per minute in a 1000-minute period. The scanning speed variance can be explained by a page's complexity, since some pages may have only a short text, while others are full of content.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

Algorithm 1 Accessing the List of Pages

Input: $\{U_A\}$ in
Output: $\{S_{AWL}\}$ out
1: **Initialization:** $\{U_A$: *URL for Array; C: Crawler; S: Scan all URL; L: Link;* S_{AWL}: *Scanned Array of Web Lists;* P_l: *Page Load* $\}$
2: **Set** U_A
3: **for** $U_A = 0$ to $U_A = S$ **do**
4: Set $L \in U_A$
5: **if** $L \notin S_{AWL}$ **then**
6: **if** P_l ! $= true$ **then**
7: **Set** P_l
8: **end if**
9: **if** $P_l = true$ **then**
10: **Set** $C \leftarrow L$
11: **Add** L to S_{AWL}
12: **Remove** L from U_A
13: **end if**
14: **end if**
15: **end for**

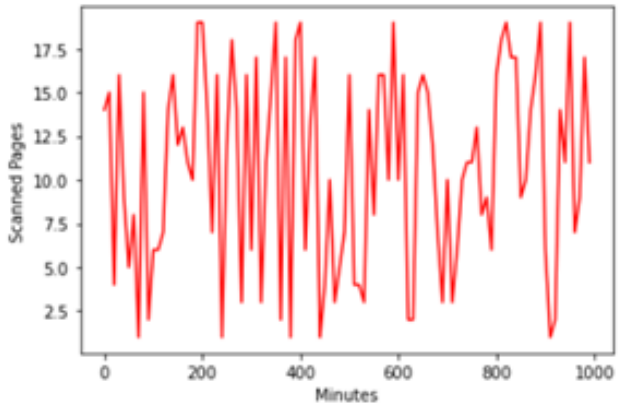


Figure 3. Crawler's scanning speed.

The Dark Net content is an object of change. When the crawler scans a web page, it records the state of a page at that certain time. However, the content becomes different over a certain period. If that period can be approximately calculated according to either time value or probability, it greatly boosts the crawler's productivity. As the crawler knows when to rescan the page, it avoids excessive unnecessary scans of a page and does not involve pages whose content is still up-to-date according to the crawler's estimations.

Changes to pages occur randomly. According to the queuing theory, random event modeling may be conducted with the Poisson point process. The Poisson random measure is used for a set of random independent events occurring with a certain frequency. Telephone calls and webpage visits may be calculated using the Poisson point field.

The probabilistic properties of the Poisson flow are completely characterized by the function $\Lambda(B)$, which is equal to a decreasing function's increment in the interval S. Most frequently, the Poisson flow has an instantaneous value of the parameter $\lambda(t)$ with the points of continuity. It is a function, whose flow event probability value is $\lambda(t)\ dt$ in the interval $[t, t + dt]$. If S is a segment $[s_1, s_2]$, then:

$$\Lambda(B) = \int_{s_1}^{s_2} \lambda(t)\ dt \tag{1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

where
$\Lambda(B)$: function characterizing probabilistic properties of the Poisson flow; s_1, s_2: initial and final time values; $\lambda(t)$: parameter whose instantaneous value is in the Poisson stream; t: time.

Poisson flows can be defined for any abstract space, including multidimensional, where it is possible to introduce the measure $\Lambda(B)$. Stationary Poisson flow in multidimensional space is characterized by spatial density λ. Moreover, $\Lambda(B)$ is equal to the volume of the region B multiplied by λ, as shown in the following equation:

$$\Lambda(B) = V(B) \times \lambda \tag{2}$$

where
$V(B)$: volume of the region B; Λ: spatial density.

In order for an event, e.g., a page content change, to be a Poisson process, it has to satisfy certain conditions. One of them states that the periods between the points have to be independent and to have an exponential distribution as follows:

$$f_A(a) = \begin{cases} \lambda e^{\Lambda a}, & a \geq 0 \\ 0 & a < 0 \end{cases} \tag{3}$$

where
A: random value; f_A: probability density function; λ: rate parameter.

The probability density function needs to be integrated to obtain the value of the exponential function:

$$F_A(a) = \begin{cases} 1 - e^{\Lambda a}, & a \geq 0 \\ 0 & a < 0 \end{cases} \tag{4}$$

where
$F_A(a)$: exponential function of a random value A.

The Poisson process takes only non-negative integer values, which leads to the moment of the n^{th} jump becoming a gamma distribution $\Gamma(\lambda, n)$:

$$P(A_t = n) = \frac{\lambda^n t^n}{n!} e^{-\lambda t} \tag{5}$$

where
P: probability in a certain nth jump; n: non-negative value in range $[0; +\infty)$.

Changes to pages occur with a certain frequency, as follows:

$$F = \frac{N}{T} \tag{6}$$

where
F: value of λ change rate; N: number of changes; T: general access time.

F value is oriented to reaching λ during the sample growth:

$$\lim_{x \to \infty} P\{|F_x - \lambda| \leq C\} = 1 \tag{7}$$

5.2. Filtering the traffic

During the scanning, especially in minimally trusted environments, there is a high risk of facing data, the storage of which is not recommended or even prohibited. The Dark Net stores a large amount of prohibited information that needs to be checked before processing and saving it to the database. Algorithm 2, listed below, performs the filtering of illegal traffic.

In algorithm 2, scanned traffic is filtered for the state of being illegal. Step 1 initializes variables used in the algorithm. The description of values sent to the algorithm and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

Algorithm 2 Filtering Illegal Traffic

Input: $\{U\}$ in
Output: $\{D\}$ out
1: **Initialization:** $\{U$: *URL;* $D_T W$: *Whitelist of Data Types;* D_T: *Data Type;* T: *Timestamp;* S_{HTTP}: *Returned HTTP Status;* D: *Database;* C_T: *Text Content* $\}$
2: **if** $D_T \notin D_{TW}$ **then**
3: **Add** D_T, T, S_{HTTP} to D
4: **end if**
5: **if** $D_T \in D_{TW}$ **then**
6: **Add** C_T to D
7: **end if**

declaration of the resulting value are introduced at the beginning of the algorithm. Steps 2-4 describe a case when the traffic is not whitelisted. Step 3 explains the addition of data type, timestamp, and returned HTTP status to the database. Steps 5-7 provide a case when the traffic type is present in the whitelist. Step 6 describes the addition of the allowed content to the database.

Hypothesis 1. *Illegal content pages follow a common pattern.*

Proof. While the crawler performs the data extraction, it obtains the content in an unfiltered state:

$$C = \{double\ URLs,\ regular\ URLs\} \tag{8}$$

where
C: extracted content. □

The process of finding the features of illicit details starts from taking a sample set of unsafe and regular URLs. The goal is to find the best feature or a set of them that will give the most accurate partitions.

Defining the best variant is carried out by comparing the entropies of partitions:

$$V_U(P) = I(P) - I(P_P) \tag{9}$$

where
V_U: variation of uncertainty; I: entropy; P: partition; P_P: previous partition.

This technique results in the creation of rules based on condition–action pairs. For example, if the page has a "drug" word and a photograph is detected, the photograph is likely to be illegal to store.

Combining the picture processing algorithms can improve the accuracy.

Image classification is frequently handled by using the bag-of-words model. The aim is to consider the pictures as a set of features, describing the picture's content. The initial step is to include the testing pictures in the database as follows:

$$I \in D \tag{10}$$

where
I: image; D: database.

The pictures are analyzed by a public feature extractor algorithm, such as scale-invariant feature transform (SIFT) [23] or KAZE (from Japanese Wind, an algorithm of feature detection used in nonlinear scale space) [24]. The result is a visual dictionary collected from a set of image features and descriptors as follows:

$$D \rightarrow \{I_f, I_d\} \rightarrow V_D \tag{11}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

where
I_f: image feature; I_d: image descriptor; V_D: visual dictionary.

Descriptors are used to create a cluster, i.e., a pattern based on all given data. K-means algorithms can be used here, as they identify a centroid as follows:

$$d(x,y) = d(y,x) = \sqrt{\sum_{i=1}^{n}(x_i y_i)^2} \tag{12}$$

where
d: Euclidean distance; x, y: points' coordinates; n: number of points; i: counting index.

While assessing an image during the crawling, its features are detected, and its descriptors are extracted and clustered as follows:

$$I \rightarrow \{I_f, I_d\} \rightarrow I_C \tag{13}$$

where I_C: clustered image.

In the next step, the clustered data are compared to the visual dictionary. The result is obtained by dictionary matching as follows:

$$I_C \in V_D \tag{14}$$

Corollary 1. *Image classification adds a certain complexity. However, it increases the accuracy of both classification and filtering stages, since some scanned webpages do not have enough text information. In this case, the presence of pictures and their analysis allows the categorization to be more precise.*

5.3. Classifying the pages

Scanned page information is stored in a text format. There are several variants of classifying the text information: Naïve Bayes, support vector machine, and deep learning algorithms.

Naïve Bayes requires the lowest amount of training data, but it also suffers from the lowest accuracy level during data classification.

Deep learning provides the highest accuracy. However, there is a need for millions of training samples.

The optimal variant for this situation is using the support vector machine algorithm. It does not require much data to output accurate results. Moreover, its accuracy level is improved when the data amount increases.

Since the training set does not have to be huge, its size was set at 1000, followed by working with the testing set. Algorithm 3 explains the classification process.

Algorithm 3 Page classification

Input: $\{T_r, T_s\}$ in
Output: $\{D_c\}$ out
1: **Initialization:** $\{$*K: Kernel; G: Gamma; C: Cost of Wrong Classification; Cl: Classifier; Tr: Training Set; Ts: Testing Set; Dc: Classified Data* $\}$
2: **Set** K
3: **Set** G
4: **Set** C
5: **Creat** Cl using K, G, C
6: **Set** $Cl \leftarrow T_r$
7: **Set** $D_c = Cl \leftarrow T_s$

In algorithm 3, scanned data stored in the database are sent for classification. In Step 1, variables mentioned in the algorithm are initialized. The description of the values sent to the algorithm and declaration of the resulting value are shown at the beginning of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

algorithm. Steps 2-4 describe the setting parameters for the classifier. In Step 5, the classifier is created by applying the parameters set in previous steps. Step 6 describes the classifier working with the training data as the preparation for the real data. Step 7 describes the process of the testing data being sent to the classifier, which results in classified data.

Hypothesis 2. *The Dark Net content has been significantly influenced by the COVID-19 epidemic.*

Proof. It is possible to carry out the analysis process of the Dark Net using a data science methodology. One of the methods is the inclusion of linear regression. Due to the fact that linear regression is a technique applied to find the correlation among variables and the resulting data, in the case of its selection, the correlation among the input data and the resulting output is also linear, as shown in the following equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +, ..., +\beta_{p-1} X_{p-1} + \varepsilon \tag{15}$$

where
$X_1, ..., X_{p-1}$: input variables; y: linear output; $\beta_0, ..., \beta_{p-1}$, ε: coefficients. □

Moreover, it is possible to change the output to the linear format by influencing the input.

The Dark Net content is subject to change. Its content is diverse and influenced by various characteristics. However, it is susceptible to Equation 16, describing the annual increase or reduction in content:

$$GR = \left(\frac{N_1}{N_0}\right)^{\frac{1}{d}} - 1 \tag{16}$$

where
GR: growth rate; N_0: initial number of contents; N_1: final number of contents; d: time difference between N_0 and N_1.

Logarithmic interpretation may be counted as the relative growth rate:

$$GR_R = (1 - 2^r) = \frac{ln(N_0) - ln(N_1)}{d} \tag{17}$$

where
GR_R: relative growth rate.

There is a characteristic showing the period of the two-fold information increase:

$$T_D = \frac{ln2}{GR_R} \tag{18}$$

where
T_D: information doubling period.

Classification issues can be solved by using the support vector machine (SVM) algorithm. Interest in this algorithm has been growing as its performance and theoretical basis satisfy requirements.

SVM assists in dividing the data into several categories. The sorting is held with a boundary that sets the border between the categories.

The given data follow the following rule:

$$v \in R^D \tag{19}$$

where
v: feature vector; R^D: vector space; D: dimension.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

It is important to note that there has to be a function mapping data points into the dedicated complex feature space from the input space:

$$\Phi(v) \in R^M \tag{20}$$

where R^M: mapping space; $\Phi(v)$: mapping function.

A hyperplane separates the pieces of data placed on the field. They are partitioned as categories. The process is written as follows:

$$H : w^T(v) + b = 0 \tag{21}$$

where
b: interception value; H: hyperplane; w^T: transposed vector normal to the hyperplane.

As obtaining the least errors is vital, the hyperplane must be placed in a certain way with a certain distance:

$$d_H\Big(\Phi(v_0)\Big) = \frac{\Big|w^T\Big(\Phi(x_0)\Big) + b\Big|}{||w||_2} \tag{22}$$

where
d_H: hyperplane distance; $||w||_2$: Euclidean norm for w length as follows:

$$||w||_2 = \sqrt{w_1^2 + w_2^2 + ... + w_n^2} \tag{23}$$

where
n: finite length value.

A hyperplane needs to be maximally far from the points of different classes, i.e., it must have the biggest margin. The focus is on the points that are closest to the hyperplane. The distance is calculated as follows:

$$w = arg_w max\Big[min_n d_H\Big(\Phi(v_n)\Big)\Big] \tag{24}$$

Correctness of classification is checked by modified Equation 16:

$$y_n[w^T(v) + b] \tag{25}$$

If the classification is correct, the value of Equation 25 is greater than or equal to 0. If the classification is incorrect, the value is negative.

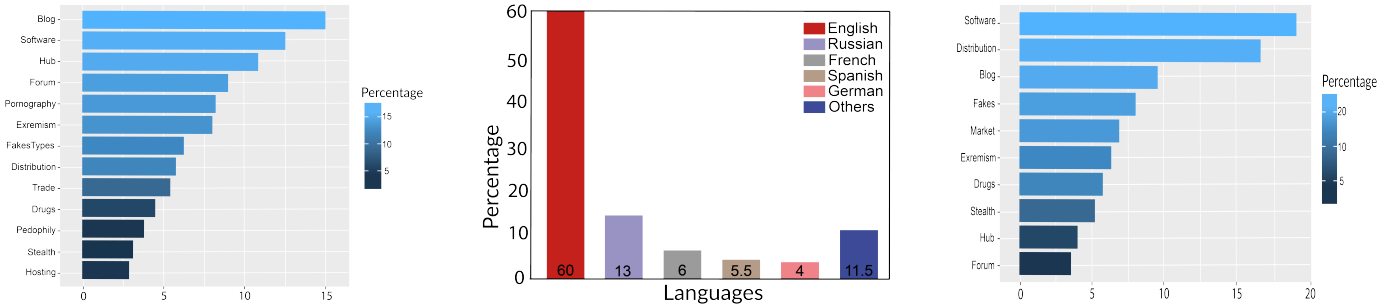
6. Experimental Results and Setup

This section describes the scanning results using the principle of classification based on websites' contents. The crawler was written by using the Python programming language. In order to store and retrieve the gathered data, the PostgreSQL relational database was used. Connection to the network is performed by using the Tor browser and ExpressVPN Proxy. We used the improved support vector machine-enabled radial basis function classifier to analyze the data for topics and state of legality and non-legality [25]. Table 1 describes the characteristics of the computer used during the experiment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

Table 1. Characteristics of the machine.

OS	MS Windows
Edition	10.0.19041 Build 19041
Processor	Intel Core i7-4750HQ, 4 Cores
Processor bit capacity	64 bit
Hard drive	256 GB SSD
RAM	16.0 GB



(**a**) (**b**) (**c**)

Figure 4. (**a**) Content distribution by types. (**b**) Content classification by language. (**c**) Distribution of hidden services.

The testing scenario is as follows. The scanning machine with a crawling program is turned on. The Virtual Private Network (VPN) is enabled. The Tor browser is executed, providing connection to the Dark Net. The crawling program is given a set of web addresses to start the scanning from and a database to store the results. The experiment starts with the execution of the crawler.

- Content distribution by types in visible and hidden parts of the Dark Net.
- Visible network legality and non-legality accuracy.
- Hidden network legality and non-legality accuracy.

6.1. Content distribution by types

It was identified that almost one-third of the non-hidden Dark Net contains webpages with no content. Since the empty pages do not carry any useful information, they needed to be excluded from the collected sample. It was identified that most websites in the visible Dark Net, without counting the empty pages, do not contain illicit content, as shown in Figure 4a.

The X-axis of Figure 4a defines webpage categories classified according to their contents. The Y-axis shows the percentage of the categories.

It is worth mentioning that the blog category is leading. This can be explained by pages that could not belong to other groups, being classified as blog pages. The result additionally proves that the majority of the content in the visible part of the Dark Net is legal.

The content is present in different languages, as shown in Figure 4b. The x-axis corresponds to the percentage, while the y-axis contains language bars.

Figure 4c illustrates data corresponding to the hidden part of the Dark Web. Axes represent the same characteristics as in the Figure 4a.

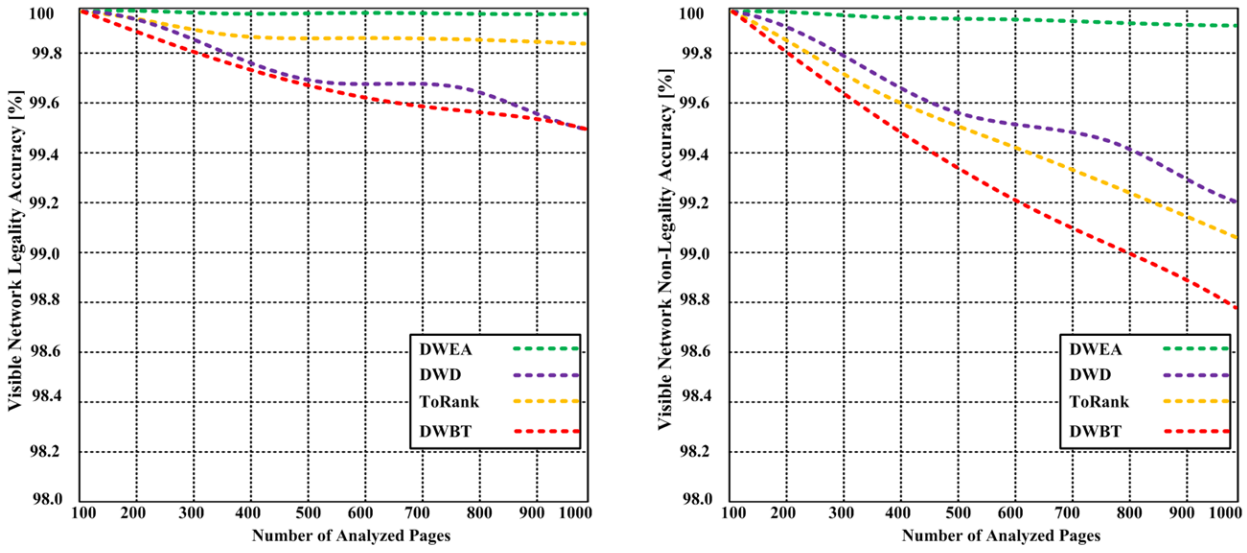
It is observed, based on the result, that the category that collected the highest number of websites was software. This is explained by the fact that many web pages use software for different purposes. Furthermore, webpages tend to collect data of users and store them. This action affects the pages included in this category.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

According to the collected results, it was identified that visible and hidden sides of the Dark Web host different contents. Deception, e.g., fraudulence, is the second group after the software in the hidden network, e.g., 17%. However, it is not even in the top five in the visible section. This means that the hidden section is likely to be a more dangerous place for users rather than the visible part.

6.2. Visible Network Legality and Non-Legality Accuracy

The content is generally classified as legal or illegal. The page detection accuracy is different depending on whether the contents are legal or illegal. Figures 5a and 5b illustrate the ratio between legal and illegal pages on the visible Dark Net based on the collected information. In this experiment, the proposed DWEA was compared to the state-of-the-art counterparts: Dark Web in Dark (DWD) [6], ToRank [10], and Dark Web-Enabled Bitcoin transactions (DWBT) [26]. Based on the results, it is observed that the proposed DWEA shows better accuracy within the visible network when detecting the number of legal pages. The DWEA obtains 99.98% visible network legality accuracy, while the counterparts, ToRank, DWD, and DWBT, obtain 99.82%, 99.51% and 99.51% respectively. Furthermore, the proposed DWEA also provides better accuracy for non-legality page detection, that is, 99.93%, whereas the contending counterparts, DWD, ToRank, and DWBT, yield 99.2%, 99.07%, and 98.78%, respectively.



(**a**) (**b**)

Figure 5. (**a**) Legality accuracy detection of the proposed DWEA and counterparts: DWBT, DWD and ToRank, with visible network. (**b**) Non-legality accuracy detection of the proposed DWEA and counterparts: DWBT, DWD and ToRank, with visible network.

6.3. Hidden network legality and non-legality accuracy

The hidden network shows almost completely opposite information, with illegal page domination. In this experiment, a maximum of 1000 pages were analyzed in the hidden dark network. Legality and non-legality accuracy were greatly affected due to hidden networks. However, the performance of the proposed DWEA is better than its counterparts. Based on the results, it is observed in Figures 6a and 6b that the proposed DWEA obtains 87.2% legality accuracy and 77.42% non-legality accuracy, whereas the counterparts are greatly affected. ToRank yields 76.67% legality accuracy and 72.09% non-legality accuracy, with a similar number of pages. On the other hand, the remaining two contending methods have lower legality and non-legality accuracy. It is proved that the proposed DWEA yields better results, despite the negative impact of the hidden network, as compared to its counterparts.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 October 2021 doi:10.20944/preprints202110.0165.v1

7. Discussion of Results

The proposed method for scanning the Dark Net consists of three stages. The first stage is the retrieval of pages to scan, the second is the scanning and collection of new pages, and the last is analysis and classification. The advantages of DWEA based on the results of the study are the broad classification and extensive analysis of information. Another advantage of this tool is that it accesses the pages several times if the page could not be loaded at the first time.

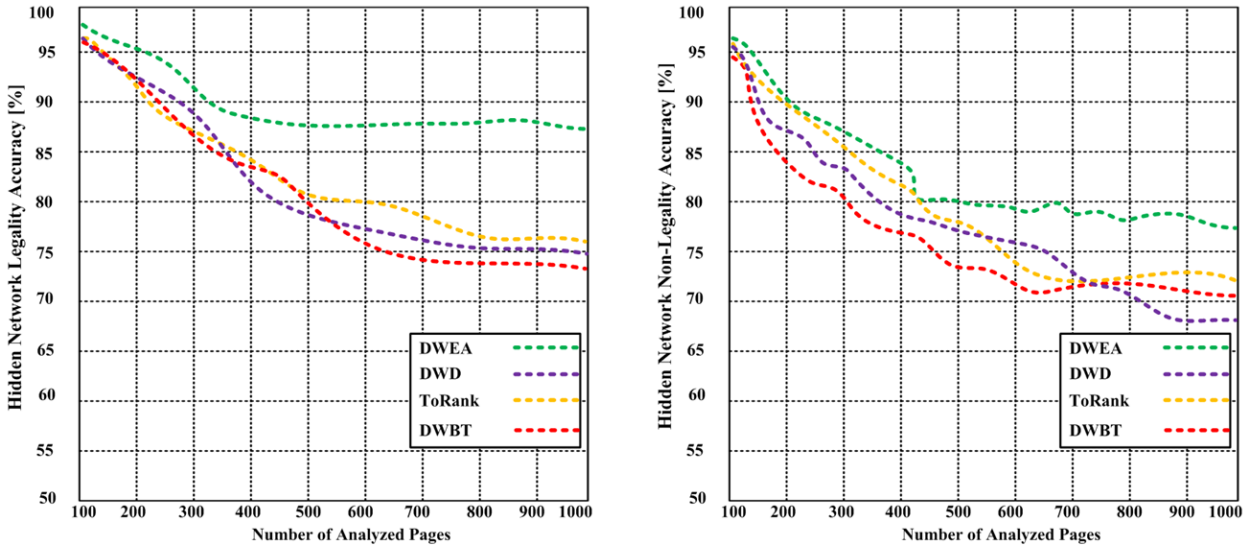


Figure 6. (**a**) Legality accuracy detection of the proposed DWEA and counterparts: DWBT, DWD and ToRank, with hidden network. (**b**) Non-Legality accuracy detection of the proposed DWEA and counterparts: DWBT, DWD and ToRank, with hidden network.

In accessing the pages that required scanning, the crawler was given a sample of pages. This is a necessary step, since the crawler needs to have a starting point. The bigger the sample is, the faster the crawler can obtain new websites. An advantage of this stage is the rescanning of pages after a certain time in case of changes. It is not a random value, but a calculation based on Poisson process points suitable for random events over a long-term period.

The classification stage involved the use of a machine learning algorithm, as this has the optimal ratio of setting complexity and calculation accuracy. The contemporary analysis of the proposed DWEA and its counterparts is given in Table 2. A shortcoming of this is that the crawler did not continuously scan the network and thus did not record all possible data. However, using samples instead of the whole data usually shows sufficient results when the sample is properly taken.

Table 2. Evaluation of the proposed DWEA and counterparts (DWBT, DWD and ToRank).

Method	Legality accuracy with visible network	Non-legality accuracy with visible network	Legality accuracy with hidden network	Non-legality accuracy with hidden network
DWD	99.51%	99.2%	75.02%	67.59%
ToRank	99.82%	99.07%	76.67%	72.09%
DWBT	99.51%	98.78%	73.83%	70.32%
Proposed DWEA	99.98%	99.93%	87.2%	77.42%

8. Conclusions

We conducted a wide analysis regarding to the design of the Dark Web and its content. In this research, DWEA was introduced to analyze the content and the composition of the Dark Web. The system performed a scanning process, and based on the collected information, it conducted a further classification on a page-by-page basis. As a result, we observed that there is a major difference between legal and illegal pages' accuracy in visible and hidden Dark Net segments. The process was based on legality and content examination. It is remarkable that the Dark Net, in general, hosts more legal resources than originally perceived. This is due to the fact that half of its web pages are classified as legitimate web resources. The most common type of crime was identified as fraud. This could be explained by people spending more time at home during the pandemic compared to the pre-pandemic period, and thus being more likely to become victims, especially when not following security rules on the net. The investigation experienced drawbacks, such as covering a relatively small portion of the Dark Net, but we are planning to improve this in the future by performing more frequent and comprehensive scans.

Acknowledgment

Taif University Researchers Supporting Project number (TURSP-2020/302), Taif University, Taif, Saudi Arabia. The authors gratefully acknowledge the support of SNCS Research Center at the University of Tabuk, Saudi Arabia. Also, the authors would like to thank the deanship of scientific research at Shaqra University for supporting this work.

Author Contributions: A.R. and B.V., conceptualization, writing, idea proposal, methodology, and results; B.A. and M.A., conceptualization, draft preparation, editing, and visualization; S.A., writing, and reviewing; A.A. conceptualization, draft preparation, editing, reviewing All authors have read and agreed to this version of the manuscript.

.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bancroft, A.; Reid, P. Concepts of illicit drug quality among darknet market users: Purity, embodied experience, craft and chemical knowledge. *International Journal of Drug Policy* **2019**, *35*, 42–49.
2. Nazah, S.; Huda, S.; Abawajy, J.; Hassan, M. M. Evolution of Dark Web threat analysis and detection: a systematic approach. *IEEE Access* **2020**, *8*, 171796-171819.
3. Dencik, L.; Cable, J.The advent of surveillance realism: Public opinion and activist responses to the Snowden leaks. *International Journal of Communication* **2017**, *11*, 763–781.
4. Mador, Z. Keep the dark web close and your cyber security tighter. *Computer Fraud & Security* **2021**, *1*, 6-8.
5. Moore, D.; Rid, T. Cryptopolitik and the Darknet. *Survival* **2016**, *58*, 7–38.
6. Tsuchiya, Y.; Hiramoto, N. Dark web in the dark: Investigating when transactions take place on cryptomarkets. *Forensic Science International: Digital Investigation* **2021**, *36*, 301093.
7. Chaudhry, P. E. The looming shadow of illicit trade on the internet. *Business Horizons* **2017**, *60*, 77–89.
8. Ladegaard, I. We know where you are, what you are doing and we will catch you: Testing deterrence theory in digital drug markets. *The British Journal of Criminology* **2018**, *58*, 414–433.
9. Fachkha, C.; Debbabi, M. Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization. *IEEE Communications Surveys & Tutorials* **2015**, *18*, 1197–1227.
10. Al-Nabki, M. W.; Fidalgo, E.; Alegre, E.; Fernández-Robles, L. Torank: Identifying the most influential suspicious domains in the tor network. *Expert Systems with Applications* **2019**, *123*, 212–226.
11. Broséus, J.; Rhumorbarbe, D.; Mireault, C.; Ouellette, V.; Crispino, F.; Décary-Hétu, D. Studying illicit drug trafficking on Darknet markets: structure and organisation from a Canadian perspective. *Forensic science international* **2016**, *264*, 7–14.
12. Oad, A.; Razaque, A.; Tolemyssov, A.; Alotaibi, M.; Alotaibi, B.; CHENGLIN, Z. Blockchain-Enabled Transaction Scanning Method for Money Laundering Detection. *Electronics* **2021**, *10*(15), 1766.
13. Razaque, A.; Al Ajlan, A.; Melaoune, N.; Alotaibi, M.; Alotaibi, B.; Dias, I.; ... ; Zhao, C. Avoidance of Cybersecurity Threats with the Deployment of a Web-Based Blockchain-Enabled Cybersecurity Awareness System. *Applied Sciences* **2021**, *11*(17), 7880.

14. Avarikioti, G.; Brunner, R.; Kiayias, A.; Wattenhofer, R.; Zindros, D. Structure and content of the visible Darknet. *arXiv preprint arXiv:1811.01348* **2018**.
15. Dolliver, D. S.; Kenney, J. L. Characteristics of drug vendors on the Tor network: a cryptomarket comparison. *Victims & Offenders* **2016**, *11*, 600–620.
16. Al Nabki, M. W.; Fidalgo, E.; Alegre, E.; de Paz, I. Classifying illegal activities on TOR network based on web textual contents. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* **2017**, *1*, 35–43.
17. Branwen, G. Darknet market archives (2013-2015). **2013**.
18. Demant, J.; Munksgaard, R.; Houborg, E. Personal use, social supply or redistribution? Cryptomarket demand on Silk Road 2 and Agora. *Trends in Organized Crime* **2018**, *21*, 42–61.
19. Munksgaard, R.; Demant, J.; Branwen, G. A replication and methodological critique of the study "Evaluating drug trafficking on the Tor Network". *International Journal of Drug Policy* **2016**, *35*, 92–96.
20. Kalpakis, G.; Tsikrika, T.; Iliou, C.; Mironidis, T.; Vrochidis, S.; Middleton, J. Interactive discovery and retrieval of web resources containing home made explosive recipes. *In International Conference on Human Aspects of Information Security, Privacy, and Trust* **2016**, 221–233, Springer, Cham.
21. Pannu, M.; Kay, I.;Harris, D. Using dark web crawler to uncover suspicious and malicious websites. *In International Conference on Applied Human Factors and Ergonomics* **2018**, 108–115, Springer, Cham.
22. Fidalgo, E.; Alegre, E.; González-Castro, V.; Fernández-Robles, L. Illegal activity categorisation in DarkNet based on image classification using CREIC method. *In International Joint Conference SOCO'17-CISIS'17-ICEUTE'17 León, Spain, September 6–8, 2017, Proceeding* **2017**, 600–609, Springer, Cham.
23. Chhabra, P.; Garg, N. K.; Kumar, M. Content-based image retrieval system using ORB and SIFT features. *Neural Computing and Applications* **2020**, *32*, 2725–2733.
24. Yakovleva, O. V.; Nikolaieva, K. Research of descriptor based image normalization and comparative analysis of SURF, SIFT, BRISK, ORB, KAZE, AKAZE. *Advanced Information Systems* **2020**, *4*, 89–101.
25. Razaque, A.; Ben Haj Frej, M.; Almiani, M.; Alotaibi, M.; Alotaibi, B. Improved Support Vector Machine Enabled Radial Basis Function and Linear Variants for Remote Sensing Image Classification. *Sensors* **2021**, *21*(13), 4431.
26. Hiramoto, N.; Tsuchiya, Y. Measuring dark web marketplaces via Bitcoin transactions: From birth to independence. *Forensic Science International: Digital Investigation* **2020**, *35*, 301086.