Dark Web Illegal Activities Crawling and Classifying Using Data Mining Techniques

https://doi.org/10.3991/ijim.v16i10.30209

Abdul Hadi M. Alaidi[1(✉)], Roa'a M. Al_airaji[2], Haider TH. Salim ALRikabi[3], Ibtisam A. Aljazaery[4], Saif Hameed Abbood[5]
[1]College of Computer Science and Information Technology, Wasit University, Wasit, Iraq
[2]Information Technology College, Babylon University, Babylon, Iraq
[3]College of Engineering, Wasit University, Wasit, Iraq
[4]Engineering College, Babylon University, Babylon, Iraq
[5]Faculty of Engineering, University Technology Malaysia (UTM), Johor, Malaysia
alaidi@uowasit.edu.iq

Abstract—Dark web is a canopy concept that denotes any kind of illicit activities carried out by anonymous persons or organizations, thereby making it difficult to trace. The illicit content on the dark web is constantly updated and changed. The collection and classification of such illegal activities are challenging tasks, as they are difficult and time-consuming. This problem has in recent times emerged as an issue that requires quick attention from both the industry and academia. To this end, efforts have been made in this article a crawler that is capable of collecting dark web pages, cleaning them, and saving them in a document database, is proposed. The crawler carries out an automatic classification of the gathered web pages into five classes. The classifiers used in classifying the pages include Linear Support Vector Classifier (SVC), Naïve Bayes (NB), and Document Frequency (TF-IDF). The experimental results revealed that an accuracy rate of 92% and 81% were achieved by SVC and NB, respectively.

Keywords—Linear Support Vector Classifier, dark web, Naïve Bayes

1 Introduction

The web is divided into parts, which are the deep web and surface web [1, 2]. Classic search engines such as Google or Bing can crawl and index the surface web. Despite this, a large part of the web has not been indexed due to how large it is, and lack of hyperlinks. This means that other web pages cannot reference that part of the web. On the other hand, the deep web is described as the other portion of the web that cannot be detected by a search engine [3]. In addition, the content of the deep web is only assessable through human interaction, as it is sealed. To access the content of this part of the web, CAPTCHA or log-in details must be used [4–6]. These kinds of pages are known as "database-driven" websites. More so, traditional search engines cannot gain access to the web layers that are underneath, therefore making it impossible to assess the deep web. The deep web plays host to the dark web, which is the deep web's subset.

It is impossible to isolate and index the dark web [7]. However, two criteria must be met in order to assess the dark web including, software that is specifically designed for that purpose or a proxy server that is meant for assessing the dark web [8]. The functionality of the dark web is activated by the Virtual sub-network of the World Wide Web (WWW) because it gives the user of the network extra covering, making them more anonymous [9, 10]. Several dark webs are used more commonly than others, including Onion Router (also called TOR), Invisible Internet Project (I2P), and Freenet. Dark websites, in the TOR community, are referred to as "Hidden Services" (HS) that are only accessible when the TOR Browser is used; this browser is specially designed to enable the accessibility of the dark web pages [11, 12]. There is a huge amount of positive research articles on the deep web. For example, there are more than 550 billion individual documents on the deep web, which is far more than that of the surface web which has just 1 billion [13, 14]. Additionally, the enormity of the deep web has been highlighted in other studies, showing that it is 400 to 500 times larger than the surface web [15]. The existence of the dark web and deep web concepts goes as far back as the establishment of the WWW. Nevertheless, its popularity has increased in recent times due to the arrest of Dread Pirate Roberts by the FBI in October 2013, who is the owner of the Silk Road black market. Preceding his arrest, the FBI estimated the sales on Silk Road to be 1.2 billion Dollars, with a trading network of about 150,000 anonymous customers, and almost 4,000 vendors [16]. One of the issues that have emerged in the area of the dark web is a cryptocurrency which has become a much-discussed topic within the dark web arena, and this is because of its potential of making the identities of trading parties as well as their monetary dealings anonymous [17] A large part of the activities of the dark web is represented by 57% of illegal dealings amounting to about 57% [18]. Based on the prediction in the study [19] 68% of 1,000 dark web content will be illicit. The findings of a study in which 5,000 Onion domains were examined, showed that criminal and illegal activities like marketing and promotion of pornography, drugs, and weapons were mostly carried out using TOR HS [20–22]. The TOR metrics statistics of January 2019 showed that there is a rapid growth in the number of dark web domains, with 120,000 registered onion addresses with approximately 2 million daily usages of TOR. Yet, only about 10% of such domains can be accessed by the public. This is attribute to the anonymity which the dark web is characterized by [23]. The illegal dealings of the dark web remain concealed due to the fact that it is difficult to detect and it is robust. A wide range of illegal activities that are regularly updated, exist in the dark web . The extraction of data from the dark web can be achieved by constructing a robust crawler that will be able to retrieve dark web pages. The process of dark web pages retrieval starts with seeds URLs, then all dark web pages under the addresses are downloaded, followed by extracting hyperlinks from the downloaded pages and including them in the list of addresses, and lastly, each link is crawled into [24]. Dark web crawling is a process that is accompanied by many limitations due to the features of TOR network, particularly, unrelated websites, whereby there is weak connection between sites, thereby making it hard for the crawler to trail [25]. Previous studies have shown that dark websites that are hosted on private encrypted networks have link addresses with a shorter lifespan than those of websites found on the surface web [26]. This is because of constant transfer across different addresses. The addresses of the websites are constantly changed to different websites in the dark web electronic

market to make them undetectable; this changing of web addresses is carried out by the web administrators. This issue has become a major challenge. Traditionally, the classifiers used for the classification of the dark web are trained using supervised training which is applied on a huge number of web pages. Given the fact that webpages of the dark web are difficult to trace, conducting research in the field of the dark web is very challenging, because of the difficulty associated with gathering sufficient illegal dark web content for analysis. The second challenge associated with researching on dark webpages is the time require for labelling the webpages manually. In addition to the two aforementioned challenges, is the technical challenge encountered when working with platforms that exist on encrypted networks. Some such challenges include bandwidth limitations, which lessens the security of the connectivity, in comparison to websites that are hosted on the surface web. Also, TOR hosted websites need more loading time than those with direct connection, due to the tunnel-like mechanism of transportation across many nodes [27, 28]. Thus, the development and implementation of methods and algorithms that can support the extraction and classification of data. Considering the issues presented above, this paper introduces a method that can enable the successful crawling of the dark web and the efficient classification of illegal activities perpetrated on the dark web. The acquisition and preparation of a suitable dataset is the initial step of creating an intelligent system of classification. In several works that have been conducted previously, the researchers adopted a ready-to-use dataset. The dark web is home to numerous illicit activities due to the mechanism employed in the creation of websites on the dark web, and this activities are updated on a regular basis. In order to successfully analyze the dark web, a new database must be created, and the creation process may be accompanied by several challenges such as the ones mentioned earlier. This study focuses on addressing some of the challenges which researchers encounter during the analysis of the dark web. Those challenges are highlighted as follows:

1. A smart crawler is proposed to facilitate the extraction of data from the dark web, and this data is then used to create a dataset. The start point of the crawling process is the seed URL page, moving to the last link available.
2. In previous studies, the researchers implemented the process of crawling through two stages: the first involves the collection of tens of thousands of web links, and the second stage involves crawling the gathered links. Given that the short lifespan of the websites is a major challenge, the largest percentage of the collected links are broken. With the proposed system, only the relevant links are retrieved through the exploration and navigation of hyperlinks.
3. An algorithm known as an automatic labeling algorithm has been proposed to help in solving the problem of labeling the dataset manually.
4. Creation of models classifier that can efficiently perform the task of classifying without needing a large training set. In this study, the classification of pages has been achieved using Linear Support Vector Classifier and Naïve Bayes (SVC) alongside, Document Frequency (TF-IDF). Also, in this work, a system has been designed and developed to enable crawling into the DarkWeb to classify illegal dealings perpetrated in the DarkWeb. The use of a web crawler was employed in building the dataset; the crawler is made up of five categories of illegal activities gathered from the DarkWeb during the course of sampling. The aim of this research is to create a

system that can categorize the DarkWeb with precision based on the text content of the Hidden Service. The methodology introduced in this study could serve as a tool used by the authorities that are responsible for monitoring the abuse of the DarkWeb. The related works that have been previously carried out in this area is reviewed in Section 2. Section 3 introduces the technique that has been proposed for crawling and classifying the features of the dataset. In section 4, the implementation of the proposed technique as well the technical details of the experiment are described. The experimental results are discussed in Section 5. Lastly, the conclusion of the study as well as recommended directions for future work are presented.

2 Related work

In the work done by Kaur (2014) [20], an informal survey was proposed; the survey includes numerous methods of web content classification. In this study, emphasis was placed on the importance of the methods to data mining. In addition, the survey provided a pre-processing mechanism that might play a critical role in the discovery of relevant features. Among these methods are the removal of HTML tags punctuation marks, and stemming. A pipeline for the classification of Agora's goods was recommended by Graczyk et al. (2015) [29]. Agora is a common DarkWeb black market. With the recommended approach, the Agora goods can be divided into 12 groups with an accuracy rate of 79%. The extraction of features is achieved by using TF-IDF, whereas principal component analysis (PCA) is used for the collection of features, and SVM is used for feature classification in their proposed pipeline architecture. Moore et al. [30] conducted a study to analyze TOR secret services through the exploration and identification of the DarkWeb. The first step was the gathering of 5K TOR onion page samples, followed by their classification into 12 classes through the use of an SVM classifier. Attention has been paid to DarkWeb e-markets, particularly, Agora in the work done by Baravalle et al. (2016) [31]. Agora represents an electronic market where the buying and selling of fake identities and identities take place. Prior to the collection of data, the authors designed a spider having a few lines of code to simulate human verification in the market. The application used for collection has been built on a classic LAMP (Linux, Apache, MySQL, PHP) stack for data collection and a variety of languages for data analysis. Also, command-line PHP was used to design the miner The miner was developed using command line PHP the cURL library) as well as on object-oriented method, where the backend was MySQL as a backend. Numerous tools have been used to analyze the extracted data, including ad hoc Java, and python scripts. The TOR DarkWeb was crawled by Rahayuda and Santiari (2017) [32], who focused on nine different domains, and described the various kinds of activities and services found on the different domains. At the end of their study, they revealed how certain domains purposely isolate themselves from the other TOR. More so, the authors made use of fuzzy K-Nearest Neighbour for classification. The results obtained through the crawling system were saved in the database, and afterward, a fuzzy-KNN method was used in classifying the results. Consequently, the crawling framework produced data in form of URL

addresses and page information. The last involved comparing the crawling process and sample data process. In a recent publication by Riesco, Adrián, et al. (2019) [33], two text representation systems (BOW and TF-IDF) were used to classify TOR HS's illegal activities. Also, three classifiers including, NB, LR, and SVM were used in the study. A dataset known as DUTA was created by the authors. The dataset contains 7K samples that were manually labeled into 26 categories, including one extra class, referred to as "Others", which contains only illegal activities like child pornography and drug trafficking. In this study, it was observed that the combination of the Logistic Regression classifier and the TFIDF text representation can yield a precision rate of 96.6%, and a 93.7 macro F1 score over ten rounds of cross-validation. In the study by Marin et al. (2018) [34], an approach of malware mining communities was discussed and the vendors were exploited, with a special focus on gaining insight into vendors and the characteristics that are common among them. Relationships between the profiles of vendors, from several markets, common categories they promote, and the number of products they have in each category were determined using clustering methods as well as different measures of similarity. According to the researchers, the proposed method is a good method that enables security agents to gain insight into how hacking-related domains can be tracked. A method of classification was introduced by Siyu He et al. (2019) [6]. The data used in training their model is referred to as the 'Federal Code of United States of America'. Based on their experimental results, they concluded that an integration of the Naïve Bayes classifier and TF-DF feature extraction method yields and accuracy rate of 93% under an experimental condition. Efforts were made by Khare et al. (2020) [19] through their work to enhance the efficiency of searching the Deep-Web; they proposed a smart crawler, which starts the crawling process from the mid-page of the seed URL, and proceeding to the last link. With this crawler, inactive and active links can be sorted and separated according to the request to the webserver of sites. Furthermore, the crawler is designed with a fitted text-based site classifier; it classifies sites based on text. Two techniques were used for the classification of content of the deep web through the use of a neural network alongside supervised machine learning techniques. Their experimental results showed that their crawler achieved an accuracy rate of 95.4% in the classification of websites, whereas, an accuracy F1-score rate of 94% was achieved by the machine learning algorithm.

2.1 Crawling the dark web

The automatic collection of websites can be achieved by web crawlers once forums and markets on the dark web have been identified, and afterward, a customized crawler is used for the detection of a major seed [35]. The web crawler is responsible for collecting data from the internet and storing it in a database so that it can be further sorted and analyzed. This process involves the collection of pages from the web by using their extant hyperlinks to download them automatically. In this process, new web pages are captured continuously. After the process of downloading an illegal deal data is completed, the data is subjected to processing and categorization so that it can be stored for a longer period. The Figure 1 illustrates how a crawler works [36].

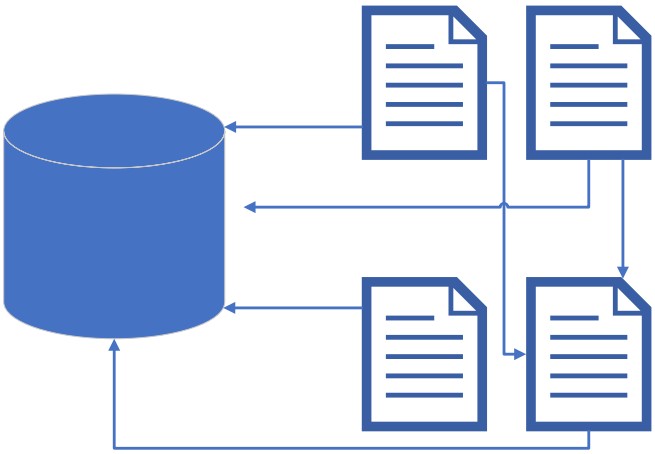

Fig. 1. Extraction and storage of data by a crawler

A web crawler, which is also known as a web spider, is an internet bot that has the ability to crawl across some HTML websites to collect information about a site. The crawler gathers information such as page titles, metatags, webpage contents, websites, and links found in the pages. Using the link that was been gathered from the first page, the crawler is able to visit the same data contained in the subsequent pages. The web crawler sends a source in form of robots.txt, and the source in turn ensures the delivery of the information to the server. In the past two decades, the development of crawling program has received great attention in dark web, however, it is accompanied by many challenges, some of which have been explained in the introduction. For such programs to be developed, more techniques will be required to enable the discovery of malevolent websites by the crawlers; the crawlers will also be able to store the websites' data for processing in the future, [37]. As noted earlier, the deep web is considerably large in size, and it is especially characterized by high quality data that is important in different semantic domains. To this end, the research area that deals with the designing of deep web crawlers that can automatically access such data is of great interest to researchers [38]. Crawlers can be used in different areas including:

- Copy creation in search engines, and the copy is afterwards used in processing all the previously visited pages in the future. This means that, search engines carry out the indexing of webpages for easy retrieval by a user when they search for particular subjects [39].
- A website is constantly secured by a crawler which cross-checks hyperlinks and authenticates HTML tags [40].
- The crawler is also responsible for the collection of particular kinds of data such email addresses, particularly malicious mails or spam [41].
- Recently, targeted crawling has been used to collect discussions from the dark [42].

2.2 Data mining

Data Mining (DM) is a process that involves the extraction of useful information or knowledge from a huge amount of data. There are different kinds of methods that can be used for data mining. Generally, there are two types of activities involved in the process of data mining, which is prediction and description [43]. Prediction is a process whereby supervised learning techniques are used to predict the value of a certain attribute based on the values of other known attributes. There are two categories of tasks that are performed in predictive modeling, including regression and classification. Meanwhile, the description tasks include mining associations, clustering, the discovery of sequence, and summarization. The techniques mentioned above, depending on unsupervised learning techniques to identify obvious patterns in data. In this work, classification methods like Naïve Bayes, Random Forest, and Support Vector Machines are used to discover categories of ambiguous data. Figure 2 below shows the key concepts associated with data mining.

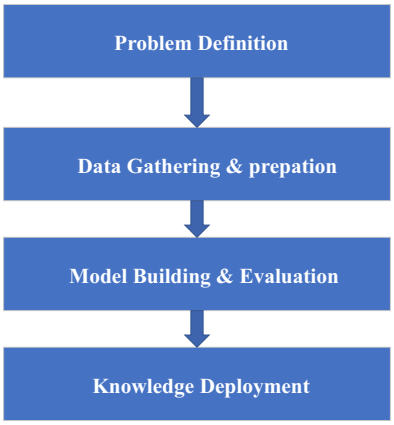

Fig. 2. Data mining concept

2.3 Linear support vector classifier (linear SVC)

Support Vector Machine (SVM) is saddled with the responsibility of finding a hyperplane or set of hyperplanes within an N-dimensional space, where N is the number of features, to divide the data points into distinct classes [44–47]. The distance between data points belonging to different classes is maximized by the optimal hyperplane. Previous studies have shown that in terms of data points that can be separated linearly, SVM has a good performance, but for non-linearity, kernel functions such as polynomial, Sigmoid, and Gaussian are required. The use of these kernels is employed in mapping non-linear separable data points into a feature space that has a higher dimension. In addition, SVC can be used when dealing with linearly separable data points of a two-class learning task due to its ability to separate two classes of a certain sample with a maximum margin. In some cases, the SVM is regarded as a maximal margin

classifier because of its ability to separate two classes. When this margin is available, the optimal generalization ability is provided [48–50]. The term generalization refers to a classifier's ability to predict with the highest rate of accuracy. In the earliest times, the linear SVC was intended for binary classification due to the fact that the problem of binary classification consists of N training instances. Each instance is represented by a tuple (*xi*, *yi*), where {*xi*, ...,} are a dataset and $yi \in \{1, -1\}$ represents its class label [51, 52]. A linear classifier's decision boundary can be represented by the following Equation 1 [53].

$$\mathrm{wT}\ .x + b = 0 \tag{1}$$

Where *w* denotes the weight vector, and *b* represents the bias in the optimal hyperplane. The decision boundaries can be derived using Equations 2 and 3 [54].

$$\mathrm{wT}\ .xi + b = 1 \tag{2}$$

$$\mathrm{wT}\ .xi + b = -1 \tag{3}$$

SVM model learning involves selecting parameters *w* and *b* according to the two conditions in Equations 4 and 5 [54].

$$\mathrm{wT}\ .xi + b \geq 1 \text{ for } yi = +1 \tag{4}$$

$$\mathrm{wT}\ .xi + b < 1 \text{ for } yi = -1 \tag{5}$$

All points must be correctly classified. Equation 6 below summarizes the two differences [57].

$$\mathrm{y}(wT\ .xi + b) \geq 1,\ \mathrm{i} = 1,2,\ldots,\mathrm{N} \tag{6}$$

Joachims [59] demonstrated the efficiency of the SV algorithm as a good classifier of text. According to this author, the features which the SVM has, allow it to achieve this feat as compared to other algorithms: High-dimensional input space, sparse document vendors, few insignificant features, and the majority of text categorization problems can be separated linearly. With this theoretical evidence, it is expected that SVMs should be efficient in the classification of text.

2.4 Naïve Bayes classifier (NB)

Simply put, the Naïve-Bayes classifier is a simple classifier that is based on the Bayesian Theorem of conditional probability and strong independence assumptions. With this classifier, it can be determined if document A is part of class B or not. More so, the functionality of the classifier is dependent on the independent feature model, and it works based on the assumption a specific attribute's existence or that the existence or non-existence is correlated to the occurrence or non-occurrence of a given feature [55, 56]. One of the advantages of using a Bayesian classifier is that only a small dataset is needed for training, its implementation is easy, insensitive to irrelevant features,

rapid and efficient classification. The Naïve Bayes is a probabilistic classifier that works according to the Naïve Bayes rule, and the posterior possibility that a document "d" belongs to the class "c" is given as follows if Equations 7 and 8 are fulfilled [57].

$$P(c/d) = \frac{P(d/c)P(c)}{P(d)} \tag{7}$$

$$P(c \mid d) = \frac{P(w1.w2.\cdots.wn|c)p(c)}{P(d)} \tag{8}$$

Where the possibility of predicting a certain class is indicated by $P(d|c)$. Where $P(wi|c)$ is the conditional probability that the word wi will be found in document d of class c. $P(wi|c)$ represents how much wi shows that the appropriate class is c. ($w1$, $w2$, ..., wn) are tokens in document d that are part of the vocabulary used for classification, and n denotes the number of such tokens in document d. The parameter $P(c)$ is the prior probability of class approximated as follows if Equation 9 is met [58].

$$P(c) = \frac{NCi}{N} \tag{9}$$

Where NCi denotes the amount of documents found in class Ci, and N is representative of all the documents contained in the training set. For all classes, $P(d)$ represents the prior probability of predictor (d). The classification of documents is carried out with the main aim of finding the most correct class for the document. Consequently, the use of Naïve Bayes classifier has been employed in predicting the class that has the highest posterior probability as Equation 10 is met [59].

$$P(c \mid d) = \operatorname{argmax} \frac{P(d \mid c)P(c)}{P(d)} \qquad c \in C \tag{10}$$

3 The proposed system

The system which is proposed in this work has an architecture that consists of four phases (crawling and collection of data, pre-processing of data, auto-labelling of the dataset, classification, and evaluation. The first stage is concerned with crawling and collection of data, and this stage is made up of many sub-steps. The purpose of this stage is to acquire the data that is used on the proposed system. The second stage is concerned with preprocessing the data acquired in the first stage, and this stage involves several processes. The stage is aimed at ensuring that the data (which is the input) is properly prepared for the system to use. At the third stage, the algorithm of automatic labelling proposed in this study is employed in labelling the dataset automatically. At the fourth stage, which is the last stage, NB classifier, linear SVC, and RF are employed. Afterwards, the crawling and mining processes are implemented in the dark web, with the aim of identifying the most commonly perpetrated illegal activities in the dark web at the time of data collection. Crawling Data. The algorithm has been able to successfully collect many dark web links, out of which, ahmia was found to be the most popular one.

After the links have been collected, they are inputted into the web crawler, which in turn acquires any other link that is associated with the ones obtained initially. When the process of data collection was over, more than 3500 links together with their content were obtained, processed, and stored in the database that was used in this work. The popularity of web crawlers has increased tremendously in different projects concerned with the collection of data from the internet. Once the hyperlinks are traced by the crawler, they are used in downloading any webpage automatically, and in the course of this, new pages are discovered. Any data that is successfully downloaded by the crawler is processed and stored for future use. The implication of this is that the data can be preserved for a long time. The major steps involved in the data collection process from the dark web by the crawler system proposed in this study is illustrated in Algorithm 1:

Algorithm: 1 Crawling algorithm

```
Input: list of seeding URLS.
Begin
    explore seeding site.
    Collect onion links and store in the list.
    While the list of links is not empty do
       explore the links and add links to list
       save page content in DB
    End while
End
```

3.1 Pre-processing

Different preprocessing steps were used for the dataset in this study. The preprocessing step is aimed at ensuring that the suitability of the dataset for use with machine learning techniques is ensured. The dataset used in this study was subjected to three preprocessing steps which are briefly described: Firstly, data cleaning that it is a process whereby the different kinds of the same letter are unified by converting all the characters to upper case or lower case, while all symbols, numbers, and non-English words are deleted. The process of data cleaning involves the following:

1. Elimination of Tag: here, tags are eliminated from the content gotten from the dark-web pages.
2. Removal of Numbers: this step involves the elimination of numbers from the content of the dark web page.
3. Removal of Punctuations: the removal of punctuation marks is very critical to the process of data cleaning, because they do not offer any useful information about the data. Not only are punctuations removed at this step, but also, text is changed into lowercase and double space is removed from page texts. Examples of punctuations are:[‘!’, ‘””, ‘#’, ‘&’, ‘“”, ‘(‘, ‘)’, ‘.’, ‘/’, ‘:’, ‘;’, ‘<’, ‘>’, ‘?’, ‘@’, ‘[‘, ‘\\’, ‘]’, ‘^’, ‘_’, ‘`’, ‘{‘, ‘|’, ‘}’, ‘~’].
4. Deletion of special characters such as, [ϒ® € £ ƒ $ º ª ✓ × † ⅞].

Secondly, Word Tokenization which is one of the critical processes of natural language. The tokenization process involves dividing texts obtained from the darkweb into

tokens, separating each word from another through the use of spacing. Tokens are of different kinds, including words, symbols, and numbers.

Finally, Remove stop words: Stop words refer to those words that occur frequently in a text, and they are words that add no meaning to the data, and as such are irrelevant to the classification process. It can also be any word with no clear relevance. Stop words are usually more than 400 words, thus, the dimensionality reduction can be achieved if stop words are deleted. The last stage of the preprocessing involves splitting the content of the dark web that has been crawled into URL, title keywords, title as a list of tokens, word frequency, and description. Figure 3 shows the pre-processing steps.

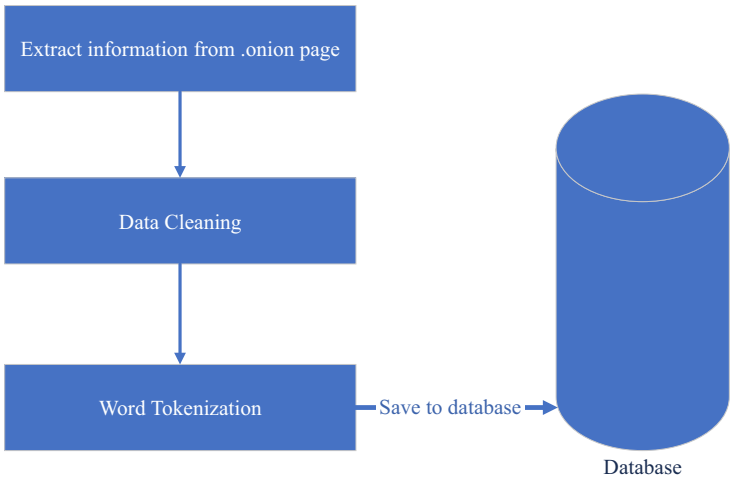

Fig. 3. Pre-processing steps

In the course of crawling and data collection, each link is pre-processed, and then tokens are extracted from all texts. After the extraction process was completed, the cleaning data started, and each token was assigned a weight according to the frequency of appearance on the webpage. Upon the completion of extraction of important features, the features were classified into different classes. For example, if a feature is related to drugs it will be placed in the list which is tagged "drug". The features that are the most important in each category are presented in Table 1 below.

Table 1. Important features for each category

No.	Category Name	Number of Features
1	Drug	104
2	Fake ID	25
3	Hacking	30
4	Weapon	94

3.2 Classification methods

Three classification systems that can perform automatic classification were used in this study, including, NB classifier, Linear SVC, and RF classifier. The automatic classification was performed according to what is traded in the market found on the darkweb. In this work, a page's features consist of the data collected from a given page as well as every word found on the page that has been crawled through. The process of classification was implemented after the implementation of several other operations. Upon completion of the classifiers training, the system was set for the next step which is testing. Testing was carried by inputting the test page into the system, and processing through the application of several steps until the system is able to make the correct prediction of the correct class for the testing page. It is noteworthy that the dataset was split into different sets (training and testing). The training set constituted 70% while the testing set was 30% of the dataset. Out of the three classifiers used, the best performance was achieved by linear SVC. Therefore, the classification of marketplaces on the darkweb was done using the linear SVC.

3.3 The validation of prediction model

This step involves verifying the model's accuracy of prediction according to the performance of the testing set. A 5-fold Cross-Validation has been used to validate the model's performance. By means of the 5-fold Cross-Validation, each instance is used four times for training and exactly once for testing.

4 Results and discussion

About 3,115 darkweb pages were obtained by the crawler proposed in this study, after each link has been crawled through, it is inputted into the crawler so that the data in that link can be extracted. Subsequently, the data is pre-processed directly through the use of data pre-processing methods and then saved in the database that has been tagged as MongoDB. With this algorithm, each class' scores can be computed using the features by measuring the similarity between document features and class and counting the number of features present in the document. Thus, it is based on the class with the greatest score, that the labelling of the document if is done. The several web markets are not restricted to just one trade but include so many like the buying and selling of illegal activities and products including, buying and selling of weapons and drugs. The darkweb also has other markets like Dark Fox and Dark Market. Some of the categories have almost the same scores, and so a fifth category was added to solve this problem. The results of this step are presented in Table 2.

Table 2. Labaling result

No.	Drugs Score	Fake ID Score	Hacking Score	Weapons Score	Class
1	18	13	12	15	Drugs
2	19	4	6	20	Others
3	36	11	11	90	Weapons
4	5	0	64	66	Other
5	4	10	0	5	Fake ID

After the dataset has been labeled automatically, some errors were identified, with several documents containing common words like hack, steroid, fake, gun in the description of the content. Using these words allows the direct labelling of the document and categorizing it into its class. The identified errors were rectified at this stage through a comparison of each word with the keywords, and then labelling based on them. The results of this step are presented in Table 3 below.

Table 3. Labelling fix

No.	Drugs Score	Fake ID Score	Hacking Score	Weapons Score	Class	Final class
1	18	13	12	15	Drugs	
2	19	4	6	20	Others	Weapons
3	36	11	11	90	Weapons	
4	5	0	64	66	Other	Hacking
5	4	10	0	5	Fake ID	

4.1 Evaluate labeling algorithm

As a way of ensuring the accuracy of the automatic labeling algorithm, the dataset was subjected to the process of manual labelling, and afterwards, a comparison of the class label result for the manual labelling was done against the dataset which was labelled automatically. Simply put, both datasets (manually labelled and automatically labelled) were compared to determine the accuracy of the methods. Table 4 shows the computed number of errors recorded for manual labelling. It is important to note that manual labelling produced almost zero percent error rate. The aim of computing the error rate was to ascertain the accuracy of the proposed algorithm. Based on the results, the proposed automatic labelling algorithm achieved an accuracy rate of 90%.

The computed accuracy rate is given as follows:

Error rate = number of error / total number of documents
= 343/3595 = 0.0954103
Accuracy = 1 – Error rate
= 1 – 0. 0954103 = 0.90459

The dataset (crawler-db) created in this study, is made up of 3595 samples divided into five classes, i.e., four main classes plus an extra one which is meant to accommodate other activities that don't belong to any of the four classes created in this study.

The number of documents in each class is shown in Table 4 below. After the process of automatic labelling has been completed, the results revealed that drug trade constituted the greatest percentage of the dark web pages which the crawler crawled through.

Table 4. Documents for each class

No.	The Name of Class	The Number of Documents in Each Class
1	Drugs	1003
2	Fake ID	453
3	Hacking	359
4	Others	579
5	Weapons	1201

4.2 Results of classification

In total, 3595 pages found on the dark web were found and classified, and the dataset was divided into two groups (training and testing dataset), where, 70% constituted the training set, and the remaining 30% the testing dataset. The acquired pages were classified into five classes using three classification methods, including, Random Forest, Linear SVC, and Naïve Bayes classifier. However, the LSVC demonstrated superior performance by achieving an accuracy rate of 91%, then followed by Random Forest which achieved 89% accuracy, and Naïve Bayes (81%). This result is attributed to the fact that linear SVC functions better in the categorization of text of high-dimensional input space.

5 Conclusions

This study proposes a web crawler that is cable of crawling on darkweb links to find markets in the darkweb. A dataset was created for the sake of this study, and it was preprocessed through the use of a variety of techniques because of its relevance. It is critical to preprocess data as it helps in converting it to a format that is more suitable for machine learning algorithms. More so, it helps in cleaning the dataset, eliminating every noise contained in the dataset. Preprocessing through data cleaning is capable of enhancing the accuracy with which features are extracted. The performance of the Linear Support Vector Machine (LSVM) was better than those of Random Forest and Naïve Bayes. Given the superior performance demonstrated by the LSVM, it is concluded that out of the three, it is the best classification algorithm for classifying dark web pages. The proposed system was evaluated using four parameters, including, F1 score, accuracy, precision, and recall. Overall, the performance of the system is excellent, as it yielded an accuracy rate of 91%, precision rate of 89%, recall rate of 88%, and F1-score of 88%.

6 References

[1] W. Ma, X. Chen, and W. Shang, "Advanced deep web crawler based on Dom," in 2012 Fifth International Joint Conference on Computational Sciences and Optimization, 2012, pp. 605–609: IEEE. https://doi.org/10.1109/CSO.2012.138

[2] U. Noor, Z. Rashid, and A. Rauf, "A survey of automatic deep web classification techniques," International Journal of Computer Applications, vol. 19, no. 6, pp. 43–50, 2011. https://doi.org/10.5120/2362-3099

[3] D. Kavallieros, D. Myttas, E. Kermitsis, E. Lissaris, G. Giataganas, and E. Darra, "Understanding the dark web," in Dark Web Investigation, 2021, pp. 3–26: Springer. https://doi.org/10.1007/978-3-030-55343-2_1

[4] O. O. Oludayo, "Research trends on CAPTCHA: A systematic literature," International Journal of Electrical Computer Engineering, vol. 11, no. 5, 2021.

[5] M. Moradi and M. Keyvanpour, "CAPTCHA and its alternatives: A review," Security Communication Networks, vol. 8, no. 12, pp. 2135–2156, 2015.

[6] H. T. ALRikabi and H. T. Hazim, "Enhanced data security of communication system using combined encryption and steganography," International Journal of Interactive Mobile Technologies, vol. 15, no. 16, pp. 144–157, 2021. https://doi.org/10.3991/ijim.v15i16.24557

[7] C.-N. Huang, H. Chen, D. Denning, N. C. Roberts, C. Larson, and X. Yu, "The dark web forum portal: From multi-lingual to video," 2011.

[8] M. Chertoff, "A public policy perspective of the dark web," Journal of Cyber Policy, vol. 2, no. 1, pp. 26–38, 2017. https://doi.org/10.1080/23738871.2017.1298643

[9] N. Weaver, V. Paxson, S. Staniford, and R. Cunningham, "Large scale malicious code: A research agenda," ed: Citeseer, 2003.

[10] R. A. Azeez, M. K. Abdul-Hussein, and M. S. Mahdi, "Design a system for an approved video copyright over cloud based on biometric iris and random walk generator using watermark technique," Periodicals of Engineering Natural Sciences, vol. 10, no. 1, pp. 178–187, 2022. https://doi.org/10.21533/pen.v10i1.2577

[11] N. Negi, "Comparison of anonymous communication networks-tor, I2P, Freenet," International Research Journal of Engineering Technology, vol. 4, no. 7, pp. 2542–2544, 2017.

[12] G. Kalpakis, T. Tsikrika, N. Cunningham, C. Iliou, S. Vrochidis, J. Middleton, and I. Kompatsiaris, "OSINT and the dark web," in Open Source Intelligence Investigation, 2016, pp. 111–132: Springer. https://doi.org/10.1007/978-3-319-47671-1_8

[13] D. Lewandowski and P. Mayr, "Exploring the academic invisible web," Library Hi Tech, vol. 24, no. 4, pp. 529–539, 2006. https://doi.org/10.1108/07378830610715392

[14] H.-J. Oh, D.-H. Won, C. Kim, S.-H. Park, and Y. Kim, "Design and implementation of crawling algorithm to collect deep web information for web archiving," Data Technologies Applications, vol. 52, no. 2, pp. 266–277, 2018. https://doi.org/10.1108/DTA-07-2017-0053

[15] M. K. Bergman, "White paper: The deep web: Surfacing hidden value," Journal of Electronic Publishing, vol. 7, no. 1, 2001.

[16] D. S. Rudesill, J. Caverlee, and D. Sui, "The deep web and the darknet: A look inside the internet's massive black box," Woodrow Wilson International Center for Scholars, STIP, vol. 3, 2015.

[17] B. J. Holland, "Transnational cybercrime: The dark web," Encyclopedia of Criminal Activities the Deep Web, pp. 108–128, 2020.

[18] G. Weimann, "Terrorist migration to the dark web," Perspectives on Terrorism, vol. 10, no. 3, pp. 40–44, 2016.

[19] D. Kavallieros, D. Myttas, E. Kermitsis, E. Lissaris, G. Giataganas, and E. Darra, "Using the dark web," in Dark Web Investigation, 2021, pp. 27–48: Springer. https://doi.org/10.1007/978-3-030-55343-2_2

[20] A. Kulm, "A framework for identifying host-based artifacts in dark web investigations," 2020.
[21] R. Liggett, J. R. Lee, A. L. Roddy, and M. A. Wallin, "The dark web as a platform for crime: An exploration of illicit drug, firearm, CSAM, and cybercrime markets," The Palgrave Handbook of International Cybercrime Cyberdeviance, pp. 91–116, 2020.
[22] L. A. Schintler and C. L. McNeely, Encyclopedia of big data. Springer, 2019. https://doi.org/10.1007/978-3-319-32001-4
[23] S. He, Y. He, and M. Li, "Classification of illegal activities on the dark web," in Proceedings of the 2019 2nd International Conference on Information Science and Systems, 2019, pp. 73–78. https://doi.org/10.1145/3322645.3322691
[24] S. Raghavan and H. Garcia-Molina, "Crawling the hidden web," Stanford 2000.
[25] V. G. Li, G. Akiwate, K. Levchenko, G. M. Voelker, and S. Savage, "Clairvoyance: Inferring blocklist use on the internet," in International Conference on Passive and Active Network Measurement, 2021, pp. 57–75: Springer. https://doi.org/10.1007/978-3-030-72582-2_4
[26] A. I. Aljazaery, H. T. Alrikabi, and H. M. A. Alaidi, "Encryption of color image based on DNA strand and exponential factor," International Journal of Online and Biomedical Engineering (iJOE), vol. 18, no. 3, pp. 101–113, 2022.
[27] B. AlKhatib and R. Basheer, "Crawling the dark web: A conceptual perspective, challenges and implementation," Journal of Digital Information Management, vol. 17, no. 2, p. 51, 2019. https://doi.org/10.6025/jdim/2019/17/2/51-60
[28] A. F. Al-zubidi, N. F. AL-Bakri, R. K. Hasoun, and S. H. Hashim, "Mobile application to detect covid-19 pandemic by using classification techniques: Proposed system," International Journal of Interactive Mobile Technologies, vol. 15, no. 16, pp. 34–51, 2021. https://doi.org/10.3991/ijim.v15i16.24195
[29] M. Graczyk and K. Kinningham, "Automatic product categorization for anonymous marketplaces," Date Unknown, 2015.
[30] R. M. Marra, J. L. Moore, A. K. J. E. T. R. Klimczak, and Development, "Content analysis of online discussion forums: A comparative analysis of protocols," vol. 52, no. 2, pp. 23–40, 2004. https://doi.org/10.1007/BF02504837
[31] A. Baravalle, M. S. Lopez, and S. W. Lee, "Mining the dark web: Drugs and fake ids," in 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016, pp. 350–356: IEEE. https://doi.org/10.1109/ICDMW.2016.0056
[32] I. G. S. Rahayuda and N. P. L. Santiari, "Crawling and cluster hidden web using crawler framework and fuzzy-KNN," in 2017 5th International Conference on Cyber and IT Service Management (CITSM), 2017, pp. 1–7: IEEE. https://doi.org/10.1109/CITSM.2017.8089225
[33] A. Riesco, E. Fidalgo, M. W. Al-Nabki, F. Jáñez-Martino, and E. Alegre, "Classifying Pastebin content through the generation of PasteCC labeled dataset," in International Conference on Hybrid Artificial Intelligence Systems, 2019, pp. 456–467: Springer. https://doi.org/10.1007/978-3-030-29859-3_39
[34] E. Marin, M. Almukaynizi, E. Nunes, and P. Shakarian, "Community finding of malware and exploit vendors on darkweb marketplaces," in 2018 1st International Conference on Data Intelligence and Security (ICDIS), 2018, pp. 81–84: IEEE. https://doi.org/10.1109/ICDIS.2018.00019
[35] H. Chen, Dark web: Exploring and data mining the dark side of the web. Springer Science & Business Media, 2011.
[36] R. Rawat, A. S. Rajawat, V. Mahor, R. N. Shaw, and A. Ghosh, "Dark web—onion hidden service discovery and crawling for profiling morphing, unstructured crime and vulnerabilities prediction," in Innovations in Electrical and Electronic Engineering, 2021, pp. 717–734: Springer. https://doi.org/10.1007/978-981-16-0749-3_57
[37] R. Campbell and M. Storr, "Challenging the kerb crawler rehabilitation programme," Feminist Review, vol. 67, no. 1, pp. 94–108, 2001. https://doi.org/10.1080/01417780122701

[38] A. T. Zulkarnine, R. Frank, B. Monk, J. Mitchell, and G. Davies, "Surfacing collaborated networks in dark web to find illicit and criminal content," in 2016 IEEE Conference on Intelligence and Security Informatics (ISI), 2016, pp. 109–114: IEEE.
[39] K. Desai, V. Devulapalli, S. Agrawal, P. Kathiria, and A. Patel, "Web crawler: Review of different types of web crawler, its issues, applications and research opportunities," International Journal of Advanced Research in Computer Science, vol. 8, no. 3, 2017.
[40] E. Ofuonye and J. Miller, "Securing web-clients with instrumented code and dynamic runtime monitoring," Journal of Systems Software, vol. 86, no. 6, pp. 1689–1711, 2013. https://doi.org/10.1016/j.jss.2013.02.047
[41] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time url spam filtering service," in 2011 IEEE Symposium on Security and Privacy, 2011, pp. 447–462: IEEE. https://doi.org/10.1109/SP.2011.25
[42] P. Koloveas, T. Chantzios, C. Tryfonopoulos, and S. Skiadopoulos, "A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence," in 2019 IEEE World Congress on Services (SERVICES), 2019, vol. 2642, pp. 3–8: IEEE. https://doi.org/10.1109/SERVICES.2019.00016
[43] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," Expert Systems with Applications, vol. 166, p. 114060, 2021. https://doi.org/10.1016/j.eswa.2020.114060
[44] C. Zhang, X. Shao, and D. Li, "Knowledge-based support vector classification based on c-svc," Procedia Computer Science, vol. 17, pp. 1083–1090, 2013. https://doi.org/10.1016/j.procs.2013.05.137
[45] S. Amarappa and S. Sathyanarayana, "Data classification using support vector machine (SVM), a simplified approach," International Journal of Electronics and Computer Science Engineering, vol. 3, pp. 435–445, 2014.
[46] A. H. M. Alaidi, A. A. Alsaidi, and O. H. Yahya, "Plate detection and recognition of Iraqi license plate using KNN algorithm," Journal of Education College Wasit University, vol. 1, no. 26, pp. 449–460, 2017. https://doi.org/10.31185/eduj.Vol1.Iss26.102
[47] N. Alseelawi, H. Tuama Hazim, and H. T. Alrikabi, "A Novel Method of Multimodal Medical Image Fusion Based on Hybrid Approach of NSCT and DTCWT," International Journal of Online and Biomedical Engineering, vol. 18, no. 3, 2022. https://doi.org/10.3991/ijoe.v18i03.28011
[48] M. B. De Almeida, A. de Pádua Braga, and J. P. Braga, "SVM-KM: speeding SVMs learning with a priori cluster selection and k-means," in Proceedings. Vol. 1. Sixth Brazilian Symposium on Neural Networks, 2000, pp. 162–167: IEEE.
[49] A. H. Alaidi, C. S. Der, and Y. W. Leong, "Systematic review of enhancement of artificial bee colony algorithm using ant colony pheromone," International Journal of Interactive Mobile Technologies, vol. 15, no. 16, p. 173, 2021. https://doi.org/10.3991/ijim.v15i16.24171
[50] O. H. Yahya, H. Alrikabi, and I. Aljazaery, "Reducing the data rate in internet of things applications by using wireless sensor network," International Journal of Online and Biomedical Engineering (iJOE), vol. 16, no. 3, pp. 107–116, 2020. https://doi.org/10.3991/ijoe.v16i03.13021
[51] G. Madzarov, D. Gjorgjevikj, and I. Chorbev, "A multi-class SVM classifier utilizing binary decision tree," Informatica, vol. 33, no. 2, 2009.
[52] A. H. M. Alaidi, "Enhanced a TCP security protocol by using optional fields in TCP header," Journal of Education College Wasit University, vol. 1, no. 24, pp. 485–502, 2016.
[53] F. Colas and P. Brazdil, "On the behavior of SVM and some older algorithms in binary text classification tasks," in International Conference on Text, Speech and Dialogue, 2006, pp. 45–52: Springer. https://doi.org/10.1007/11846406_6

[54] Y. Lin and J. Wang, "Research on text classification based on SVM-KNN," in 2014 IEEE 5th International Conference on Software Engineering and Service Science, 2014, pp. 842–844: IEEE. https://doi.org/10.1109/ICSESS.2014.6933697
[55] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," Journal of Information Science, vol. 44, no. 1, pp. 48–59, 2018. https://doi.org/10.1177/0165551516677946
[56] P. Joe Dhanith, B. Surendiran, and S. Raja, "A word embedding based approach for focused web crawling using the recurrent neural network," International Journal of Interactive Multimedia Artificial Intelligence, vol. 6, no. 6, 2021.
[57] W. Ding, S. Yu, Q. Wang, J. Yu, and Q. Guo, "A novel naive bayesian text classifier," in 2008 International Symposiums on Information Processing, 2008, pp. 78–82: IEEE.
[58] J. Kazmierska and J. Malicki, "Application of the Naïve Bayesian Classifier to optimize treatment decisions," Radiotherapy Oncology, vol. 86, no. 2, pp. 211–216, 2008. https://doi.org/10.1016/j.radonc.2007.10.019
[59] H. Adel and M. A. Bayati, "Building bi-lingual anti-spam SMS filter," International Journal of New Technology Research, vol. 4, no. 1, p. 263147.

7 Authors

Abdul Hadi M. Alaidi is a Asst. Prof. in the Engineering College, at the Wasit University, Iraq. His area of research focuses on algorithm and image processing.

Roa'a M. Al_airaji received the Bachelor degree from Department of Computer, College of Science, University of Babylon, in 2012. Received the Master degree from Department of software, College of Information Technology, University of Babylon, in 2018. Currently work as assistant teacher in College of Science, University of Babylon.

Haider Th. Salim ALRikabi is presently Asst. Prof. and one of the Faculty College of Engineering, Electrical Engineering Department, Wasit University in Al Kut, Wasit, Iraq. He received his B.Sc. degree in Electrical Engineering in 2006 from the Al Mustansiriya University in Baghdad, Iraq. His M.Sc. degree in Electrical Engineering focusing on Communications Systems from California State University/Fullerton/USA in 2014. His current research interests include Communications systems with the mobile generation, Control systems, intelligent technologies, smart cities, and the Internet of Things (IoT). Al Kut City-Hay ALRabee, Wasit, Iraq. E-mail: hdhiyab@uowasit.edu.iq. The number of articles in national databases – 15, The number of articles in international databases – 45.

Ibtisam A. Aljazaery is presently Asst. Prof. and on the faculty of Electrical Engineering Department, College of Engineering, University of Babylon. Babylon, Iraq. E-mail: ibtisamalasady@gmail.com. The number of articles in national databases – 10, The number of articles in international databases – 5.

Saif Hameed Abbood is member of School of Computing, Faculty of Engineering, University Technology Malaysia (UTM), Johor 81310, Malaysia.

Article submitted 2022-02-14. Resubmitted 2022-03-17. Final acceptance 2022-03-17. Final version published as submitted by the authors.