Bulletin of Computer Science and Electrical Engineering
Vol. 4, No. 1, June 2023, pp. 37~48
ISSN: 2722-7324, DOI: 10.25008/bcsee. v4i1.1179

Investigation the Dark Web Illegal Activities with Data Mining Approach

Agus Pamuji[1]
[1]Department of Islamic Counseling Guidance, Islamic State Religion Institute Sheikh Nurjati Cirebon, Indonesia

Article Info

Article history:

Received May 05, 2023
Revised May 23, 2023
Accepted Jun 12, 2023

Keywords:

Dark Web
Illegal Activities
Web Crawling
Data mining
Classification Classifier

ABSTRACT

The rapid advancements in internet technology have opened up various avenues for illicit activities targeting users. These nefarious activities are carried out by anonymous individuals or groups, making identification and tracking a challenging task. Periodical updates to the content of the dark web are common, with alterations to concealed data often escaping detection. Consequently, the primary arduous tasks concerned the data mining framework and its impact on the classification accuracy with regards to illegal activities. In contemporary times, the constraint of dealing with considerations pertaining to the academia and the business environment has emerged as a crucial phenomenon. This paper encompasses an analysis of a web crawler designed for the dark web. The crawler is proficient not only in data collection and cleansing but also in storage, making use of a data-driven approach. Data mining is a potent technique that enables thorough investigation through exhaustive exploration of data, often revealing evidence of illicit activity. Consequently, the crawler has emerged as a focal point for enforcing automated classification of the amassed web pages into five distinct categories. The classification process involved the utilization of classifiers, specifically the Linear Support Vector Classifier (SVC) and Naïve Bayes (NB) for the categorization of pages. Furthermore, as per the probationary findings, the Support Vector Classifier (SVC) and the Naive Bayes (NB) algorithm demonstrated precision rates of 91% and 84%, respectively, in the presented sequence.

Corresponding Author:

Agus Pamuji,
Department of Islamic Counseling Guidance,
Islamic State Institute Sheikh Nurjati Cirebon,
Jl Perjuangan Street, ByPass Sunyaragi, Kesambi Cirebon City 45132, Indonesia.
Email: agus.pamuji@syekhnurjati.ac.id

1. INTRODUCTION

The advent of internet technology has engendered unrestricted access to a vast array of resources available in the world of the virtual network, thereby facilitating the exploration of these resources, thus promoting a sense of unrestrained freedom in the online domain. The internet has undergone significant transformations over the past 30 years since it was first made publicly accessible. Its evolution has been marked by the emergence of the dark web, a contentious facet of this technology. The utilization of a sophisticated system by the dark web effectively obscures a user's genuine IP address, thereby rendering it daunting to deduce the webpages that a given device has accessed. The typical method of accessing it involves the utilization of specialized software commonly known as The Onion Router (TOR) [1]. In an alternative perspective, the TOR network is not commonly categorized as a component of the "dark web"; rather, it serves as a means to access the internet while maintaining user anonymity and preventing surveillance of user behavior [2].

The internet is commonly classified into two distinct categories, namely the surface web and the deep web. The surface web exhibits a proclivity to be targeted for crawling and indexing through the application of traditional search engines [3]. Additionally, the TOR provides anonymizing software as well as the option to download it as an extended feature. Search queries and messages transmitted through the TOR browser are not intended for direct delivery to their designated destinations. In lieu of direct transmission, information is conveyed through intermediary networks referred to as "nodes," which are managed by other Tor users [4]. One contributing factor to the lack of indexing of certain portions of the web is the absence of hyperlinks therein, combined with the relatively abundant size of the pages in question. As a result, other web pages are hindered in their ability to reference the unindexed areas. On the contrary, the search engine is incapable of detecting the deep web as it constitutes a distinct, unindexed segment of the internet [5]. Moreover, accessing the content present in the deep web poses a challenging constraint due to the requirement of log-in validation or log-in details to be furnished by the user. The types of webpages referred to as "database-driven" sites are recognized [6]. Consequently, traditional search engines are incapable of accessing the underlying layers of the internet. Additionally, the deep web is intricately linked to the dark web, which is classified as a subset of the deep web. Subsequently, gaining access to the deep web remains an elusive feat [7].

The software has been designed for that purpose or a proxy server in order to meet when assessing the dark web [8]. we have argued that it is impossible to isolate and index the dark web. As long as the user of the network has extra coverage and makes them more anonymous, the virtual sub-network of the world wide web activates the functionality of the dark web [9]. Onion Router (TOR) and Invisible Internet Project (I2P) are the several dark webs in use more commonly than others. The dark website is referred to as Hidden services when the TOR browser is used, and the website is designed to enable the accessibility of the dark web pages specifically. Although the illegal zeal of the dark web remains furtive, the left web is difficult to detect and it is robust increasingly while a wide range of illegal activities have been updated regularly and exist in the dark web. Generating a robust crawler will be capable to retrieve the dark web pages in order to the extraction of data from the web was considered illegal. The following steps of illegal web page retrieval start with the seed URL[10]. Next all illegal web pages under the address are downloaded, this stage is attended by extracting hyperlinks from the downloaded pages and including them in the list of addresses, and each link is crawled into lastly [11].

This phenomenon can be attributed to the recurrent relocation of data between various network locations. The web addresses of websites within the electronic marketplace of the dark web are subject to constant alteration with the aim of rendering them impervious to detection. This task of changing the website's URL is executed by the respective web administrators. This matter has become a significant obstacle. The TOR network's inclusion of non-associated web pages as part of the dark web crawling process has resulted in a tenuous linkage between said sites. Prior research has elucidated that the Uniform Resource Locator (URL) associated with private encrypted networks exhibits a decreased temporal distance as compared to the corresponding URL for the publicly accessible surface web. "The network is hosted on dark websites due to the frequent transfers between multiple addresses".

The aforementioned concern has emerged as a significant obstacle, prompting web administrators to frequently alter online addresses and redirect them towards diverse websites within the clandestine electronic commerce market, in order to circumvent detection. The classifier has traditionally employed supervised training, applied over a large corpus of web pages, to facilitate the categorization of the dark web. Undertaking research pertaining to the dark web poses a significant challenge, rendering the tracing of the dark web's webpages to be notably intricate. In light of the challenge of procuring adequate illicit content from the obscure realm of the dark web for analytical purposes. A crucial challenge pertinent to the investigation of dark webpages is the temporal constraints imposed, consequently necessitating manual labeling of the webpages. Moreover, in conjunction with the aforementioned issues, working with platforms present in encrypted networks presents a technical challenge. The constraining capacity of bandwidth, which reduces the level of security of connectivity in contrast to a website that is hosted on the visible

surface of the World Wide Web, has presented a number of notable challenges. Another point of consideration is the increased demand for loading time on websites hosted on TOR network as compared to those directly connected to the internet. This can be attributed to the use of the tunnel-like mechanism that requires transportation across multiple nodes. The support of data extraction and classification would be facilitated by the development and implementation of methods, including algorithms.

Given the aforementioned issues, an efficacious method for facilitating successful web crawling within the deep web has been proposed and implemented. This method has proven effective in systematically categorizing illicit activities within this domain [12]. At the outset, the procedure involved in the development of an intelligent classification system entails the collection or preparation of an appropriate dataset. Over the course of preceding investigations, researchers have employed a pre-existing dataset for their studies. The dark web is host to a multitude of illegal activities, primarily attributable to the techniques utilized in the development of websites on the dark web, which are frequently revised and replenished`.

The creation of a new database may be accompanied by various challenges, such as those identified during the preliminary stage. Therefore, it is imperative to consider the generation of a new database with caution and foresight. The researcher undertakes an analysis of the dark web while taking into account various constraints. The preliminary step towards ascertaining the endorsement of the final link furnished involves examining the digest URL page. The employment of the dark web for the purpose of extracting data has been investigated by means of implementing web crawlers. However, the collection of datasets continues to be a significant goal. The existing body of literature has expounded upon the approach of conducting web crawling in a sequential fashion consisting of two stages- firstly, the compilation of a corpus of ten thousand web links, followed by their acquisition through a sophisticated crawler. The truncated lifespan of the website represents a noteworthy constraint, since a substantial proportion of the aggregated hyperlinks have undergone degradation. The current methodology of the system involves the process of seeking and traversing hyperlinks with the aim of acquiring relevant information. The classification of datasets represents a complex task, which has been met with a solution in the form of an algorithm aimed at enhancing processing capabilities through automated means. The development of a classifier model allows for effective facilitation of grouping processes, with reduced dependence on extensive data training sets. In regards to this matter, the categorization of pages can be accomplished by employing a fusion of Naïve Bayes, Document Frequency (TF-IDF), and Linear Support Vector methodologies. Subsequently, a mechanism has been implemented to streamline the web crawling procedure, consequently allowing for the categorization of illicit practices transpiring within the abyssal layers of the dark web.

The implementation of a web crawler was instrumental in the creation of the dataset, comprising five distinct categories of illicit activities, culled from the dark web through a sampling methodology. The primary objective of the present study is to investigate a novel system that demonstrates the ability to effectively categorize the dark web with a high degree of accuracy based on the textual content of the concealed services.

2. RESEARCH METHOD

2.1 Crawling the dark

A web crawlers have capability to collect a website automatically at the first take when the dark web had been recognized, besides the customization is implemented for detection of a major seed[13]. Moreover, during the collecting data from the long internet is referred to web crawler responsibility and perform storing in database, it seems can be shorted and analyzed. The next process is collection of pages which taken from the website through extant hyperlink to download automatically. The process of downloading an illegal deal data is completed before data become subjected to processing and categorization, it would be stored for a long period.

Thomaz declares that a web crawler is a web spider identically, an internet boat that can crawl across HTML websites to collect site information. For example, the page titles, metatags, web page contents, website and links are a link that is gathered from the first page and can be found in the pages. Another thing, the robot.txt is a form that was sent by a web crawler so the source in

turn ensures the delibery information to the server. The evolution of crawling activity underwent great potential in the dark web, it is accompanied by several challenges where some of which have been described in the preliminary stage. In fact, more techniques will be offered to enable the discovery of the malevolent website with the crawler and also be able to store the website data in order to process it in the future. Considering about the deep web is large in size and characterized using quality data is important in different semantic domains especially. To this end, the research area that deals with designing deep web crawlers can access automatically, for example, data is of great interest to researchers. Hence, the crawler in different areas including. The processing of all previously visited pages was adopting upon copy as the creation in search engines. This notice that the search engines carry out indexing f web pages for easy when the retrieval by user search for a particular subject. The crawler cross-checks, and authentical secured a website with hyperlinks and authenticates HTML tags. The crawler has supported the collection of particular kinds kinds of data including email addresses, and particularly malicious mail or spam.

2.2 Data Mining

Data mining is a process wherein useful information or knowledge is extracted from vast quantities of data. At present, various techniques are available for processing data in the field of data mining. In the data mining phase, two fundamental endeavors are conventionally undertaken, namely, prognostication and delineation. One aspect of data mining that pertains to predictive modeling involves the utilization of supervised learning methodologies for the purpose of forecasting the value of a specific attribute based on other attributes that have been previously identified [14].

Within the framework of data mining, there exist two fundamental modes of performance that are established through the use of predictive modelling: regression and classification. Whilst predictive modelling certainly plays a critical role in facilitating robust analysis and performance, the task of describing data involves a plethora of techniques, including association mining, clustering, sequence discovery, and summarisation. The identification of explicit patterns within data relies on unsupervised learning methodologies for the description tasks. The present study employs various machine learning techniques, namely, Naïve Bayes, Random Forest, and Support Vector Machines, to identify and resolve uncertain data instances. Figure 1 illustrates the fundamental concept that underlies the process of data mining [15].

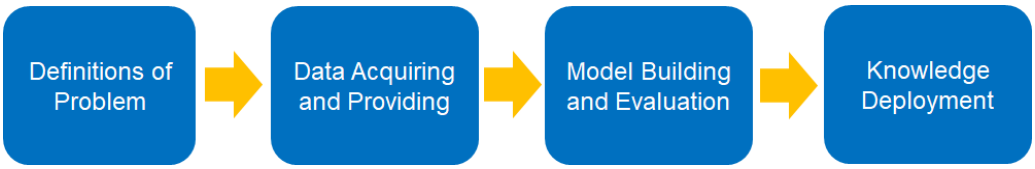

Figure 1. Data mining in illegal activities [5]

2.3 Linear Support Vector Machines (linear SVC)

Either the finding a hyperplane or set of hyperplanes within an N-dimensional space, where N is the numerous of the field as the class of classification, the distinction of the class was determined by dividing from the data points. The optimal hyperplane would be maximized through different classes when the distance between data points was identified. The term of data points could be separated linearly as consideration in the previous work, and Support Vector Machine (SVM) has a good performance while the kernel function such as polynomial, Sigmoid, and Gaussian are deserved. Moreover, the implementation of the kernel was signed in mapping non-linear separable data points that could be changed to a feature space that has a not low dimension. In addition, SVC could be utilized when attending to separable data points of a two-class learning task due to its ability to separate two classes of a determined sample from a maximum margin. In another case, the SCM was admired as a maximal margin, and the classifier capabilities separate two classes. The optimal generalisation ability is available when the margin was supported. Hence, the capability to predict with the highest rate of accuracy was called the term generalization. The constraint of binary classification includes N training instances because the linear SVC was

contemplated for binary classification, preliminary. Every instance in this context was symbolized by a tuple (xi, yi), where {xi, ...,} are a dataset and yi ∈ {1, −1} illustrates its class label [16]. A linear classifier's decision boundary could be characterized by the following Equation 1.

$$wT .x + b = 0 \quad (1)$$

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].

$$wT .xi + b = 1 \quad (2)$$

$$wT .xi + b = -1 \quad (3)$$

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

$$wT .xi + b \geq 1 \text{ for } yi = +1 \quad (4)$$
$$wT .xi + b < 1 \text{ for } yi = -1 \quad (5)$$

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

$$y(wT .xi + b) \geq 1, i = 1,2, \ldots, N \quad (6)$$

Chandrakala as an author for this case suggested the optimality of the SV algorithm is the best performance 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. In the working evidence, it is expected that SVMs should be more efficient in the classification of text.

2.4 Naïve Bayes

The Naïve Bayes technique is one of the classification algorithms which is underlying the Bayesian Theorem of conditional probability besides it has strong independence assumptions. The fundamental notes, a data whether document A is part of class B considering the probability. Moreover, the functionality of the kind of classifier is dependent on the feature model independent conceptually [17]. Another thing, the Naïve Bayes tools have been working through the assumption a specific attribute exists while that either existence or non-existence is correlated to the occurrence of a given feature. The literature claimed that one of the beneficial implementing a Bayesian classifier is only a not overall dataset requirement. The first is the training process, flexibility aspect, and incentive to irrelevant features and also efficient classification as the evaluation in the analysis of context. For example, a document "d" belonging to class "c" was given as the following if the equations 7 and xxx were fulfilled. Eventually, the Bayesian-based algorithm could be worked according to the not only Naïve Bayes rule, and posterior possibility but also the probabilistic classifier[18][19].

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

$$P(c/d) = \frac{P(w1.w2 \ldots \ldots wn|c)p(c)}{P(d)} \quad (8)$$

The possibility of predicting a determined class was designated by P(d|c). Where P(wi|c) is the conditional probability that the word w1 will be identified in document d of class c. P(wi|c) denotes how sizable wi displays that the suitable class is c. (w1, w2, ..., wn) are tokens in document d that are part of the data item employed 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 meet.

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

Where NCi denotes the numerous of documents identified in the class Ci, and N is representative of all the documents composed in the training set. For all classes, P(d) illustrates 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.

$$P(c) = argmax\ \frac{P(d|c)P(c)}{P(d)}\ c\ \epsilon C \quad (10)$$

2.5 The proposed system

As previously indicated, the methodologies involved include crawling and data collection, pre-processing of data, automated labelling of the dataset, classification, and evaluation. Initially, the phase involves the acquisition and consolidation of information, whereby the various constituent sub-stages are integrated within the corresponding ongoing phase. The primary objective of this stage is to procure the requisite data, which is subsequently utilized in the development of the integrated system. The second phase of the process pertains to the preprocessing of the data obtained during the preliminary stage, in addition to the stages that assimilate the progression sequence. Furthermore, this phase is intended to ensure that the data, serving as the input, were appropriately provided during the implementation of the system. In the third stage of the experiment, an algorithm for non-traditional labelling was implemented in order to automate the labelling process of the dataset. In the fourth stage, the selection of Naïve Bayes, Linear SVC, and RF classification methods represents both the primary and final stage of the analysis. On a separate occasion, the procedures of mining and crawling were executed within the confines of the dark web in order to ascertain prevalent instances of anomalous activity during the data acquisition process. Furthermore, the presented algorithm has successfully demonstrated its capacity to aggregate numerous distributions of dark web links.

Upon completion of the data collection process, a total of over two thousand links, along with their corresponding content, were obtained and subsequently accumulated in the database. This approach formed an integral component of the methodology employed in the study. Consequently, there has been a notable escalation in the number of projects focused on amassing data from online resources for the purpose of enhancing the standing of web crawlers. When employing crawler tools to automatically download webpages, hyperlinks are traced by the tools leading to the identification of subsequent pages. The crawler has been engaged in the processing and storage of downloaded data, with particular attention to ensuring its longevity for future use. The primary stages encompassed in acquiring data from the dark web through the usage of a crawler system created in this research are outlined in Algorithm 1:

Algorithm: 1 Crawling algorithm

```
Input: list of digests URLS.
Begin
      explore digest site.
      Collect onion links and store in the draft.
      While the list of links is not empty do
            explore the links and add links to draft
      record page content in checklist
      End while
End
```

2.6 Pre-processing

In this work, the dataset related to the case employed different pre-processing rules. The pre-processing will be able to determine that the profitability of the dataset with the assist of the machine learning techniques was secured [20]. The dataset employed in this work was subjected to three pre-processing sequences which are explained briefly. Firstly, a process where the different kinds of the same letter are unified by converting all the characters to either upper or lower case, while all symbols, numbers, and any words are removed the process of data cleaning is intended as the following:

1. Elimination of Tag:the tags are deleted from the content that obtained from the considered darkweb pages.
2. Removal of Numbers: its involves the removing of numbers from the content
3. of the dark web page.
4. Removal of Punctuations: the deletion of punctuation marks is realy critical to the process of data cleaning considerably, due to the fact that they do not provide any useful information about the data. Not only are punctuations removed at this step, but also, text is changed into lowercase and double space was decreased from page texts. Examples of punctuations are :['!', '"', '#', '&', '"', '(', ')', '.', '/', ':', ';', '<', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~'].
5. Deletion of special characters such as, [® € £ ƒ $⁒].

The second stage of the analysis involves the incorporation of natural language, which encompasses a fundamental process known as word tokenization [21]. The objective of the tokenization exercise is to obtain a distribution of text sourced from the dark web by converting it into individual tokens. This process involves isolating each word from its immediate neighbor by utilizing white space as a delimiter. Consequently, various types of tokens exist, comprising lexemes, ideograms, and numerals. Ultimately, the elimination of stop words does not contribute valuable meaning towards the data. Stop words are recognized as the most commonly occurring words within a textual resource, whereby they are essentially deemed irrelevant in the context of the classification process. Furthermore, it is imperative to note that stop words within a given context may refer to any word whose relevance remains unclear. The frequency of stop words typically exceeds 300 words, thus a reduction in dimensionality may be necessary to ascertain the impact of elimination of these stop words. A further stage in the pre-processing of the dark web entails the partitioning of its content following its retrieval from diverse URLs, accompanied by consideration of keywords, titles, token lists, frequency of words, and descriptions. Figure 2 is capable of presenting the pre-processing stages.[21].

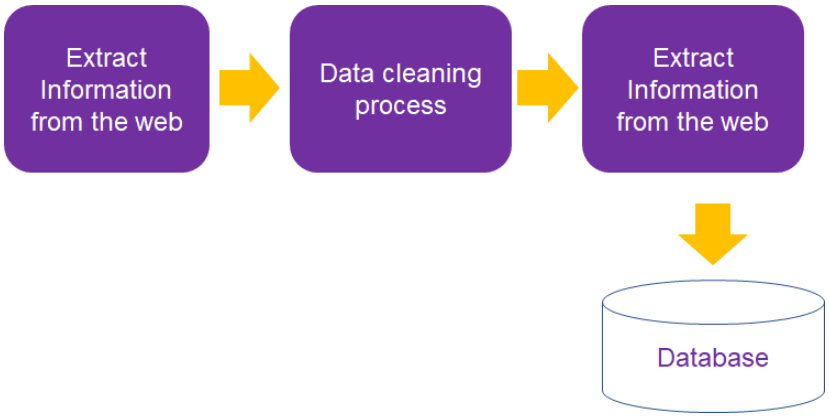

Figure 2. Pre-processing in web crawling data mining step [17]

Each hyperlink originating from the websites was subjected to preprocessing procedures. Furthermore, tokens were extracted from all textual content in both the crawling process and data gathering phase. Upon completion of the extraction process, the cleaning of the data ensued,

followed by the allocation of weights to each token based on its frequency of occurrence on the webpage. For instance, a characteristic was given preference over pharmaceutical substances if it could be categorically arranged under the label of "drug. " Table 1 below enumerates the features deemed most indispensable in certain categories.

Table 1. Cleaning Data Parameter of Feature

No	Category name	Number of features
1	Theft	56
2	Hacking	12
3	Fault	23
4	Fake Account	10
5	Drug	31

2.7 The method of classifications

In this work, there are three ways of classification system which could perform automatic classification were employed among other things Naïve Bayes, Linear SVC, and RF classifier. The traded in the market could be identified on the dark web that considered as bottom-line when an automatic classification had been conducted[22]. In this scenario, a page attribute comprises of the information accumulated from an exhibited webpage, encompassing all the words identified during the crawling process. This is presented in an academic style of writing. Subsequent to this, the procedure of categorization shall have been implemented subsequent to the execution of other sequential activities. The system was configured and prepared for the subsequent stage, namely the testing phase. However, prior to its implementation, the classifier data training was successfully accomplished. In the process of testing, the page is inputted into a system and undergoes a series of sequential steps aimed at enabling the system to render accurate predictions of the correct class during the test. It is noteworthy that the dataset was partitioned into cohesive subsets, encompassing both the training and testing phases. The training set was comprised of approximately 70% of the dataset, whereas the testing set constituted just below 50%. Of the three classifiers utilized, it was determined that linear support vector classification (SVC) achieved the most optimal performance. Hence, the categorization of marketplaces on the darkweb was implemented utilizing the linear Support Vector Machine (SVM) algorithm.

2.8 The validation of prediction model

In the prediction model, sequence steps involve confirming the model's accuracy of prediction according to the performance of the testing set[23]. The validation of model performance was accomplished through the practical implementation of 10-fold Cross-Validation with 10 iterations of testing. In order to produce a dependable model and meet the criteria for in-depth analysis, the proposal advocates for the implementation of a predictive model in the analysis of web crawling. The utilization of the 10-fold Cross-Validation method involves the repeated implementation of data training nine times, with each instance serving as a treatment, followed by one instance of testing at an early stage in the process.

3. RESULTS AND DISCUSSION

As a consequence, the current study involved the acquisition of slightly over 3500 darkweb locations via a step-by-step web crawling approach. Subsequently, each link enlisted in the aforementioned list undergoes a comprehensive crawling process to ensure its inclusion in the crawler and to obtain the data encapsulated within said link. The process of recording data in a database layer can be facilitated through the utilization of tagging, specifically within the MariaDB system. The algorithmic performance can facilitate computation of the total score of a class via data manipulation utilizing features. Additionally, it is imperative that this particular functionality effectively evaluates the resemblance among document characteristics, classifications, and the frequency of feature occurrence in the given data. As such, the attainment of high scores appears to be associated with the focus on particular classes, while the labeling of documents is found to be a

crucial factor in ensuring the validity and reliability of the analysis. The plethora of online marketplaces is not limited to a solitary commercial enterprise but encompasses a wide array of illicit activities, such as the exchange of contraband goods, including firearms and narcotics. Nevertheless, the web market is not restricted solely to conventional activity behaviors, as the dark web encompasses alternative markets such as the dark Fox. As such, it has been observed that certain categories exhibit a high degree of similarity in terms of the value of their constituent items. In order to effectively address this issue, a fifth category was introduced. The outcome of this step is evidenced in Table 2.

Table 2. Identified class in the classification

No	Theft scores	Hacking score	Fault Score	Drug score	Fake Acc Score	Class
1	19	21	7	18	16	Hacking
2	3	23	11	15	12	Hacking
3	5	10	10	17	7	Drug
4	9	8	7	5	15	Fake acc
5	4	10	0	5	16	Fake acc

After the automatic labeling of the dataset, certain errors were detected through a thorough examination of several documents. These errors stemmed from the prevalence of commonly used words, such as "hack," "steroid," "fake," and "gun," found within the content descriptors. These terms enable the grant of labeling documents and their categorization into respective classes. The error in question was rectified at this stage by means of performing a meticulous analysis of individual words using the associated keywords and their linkage to the overall process. The outcome of this stage is exhibited in Table 3 presented below.

Table 3. Final class in the classification

No	Theft scores	Hacking score	Fault Score	Drug score	Fake Acc Score	Class	Final class
1	19	21	7	18	16	Hacking	Hacking
2	3	23	11	15	12	Hacking	Weapons
3	5	10	10	17	7	Drug	Drug
4	9	8	7	5	15	Fake acc	Hacking
5	4	10	0	5	16	Fake acc	Hacking

3.1 Evaluate labeling algorithm

To ascertain the precision of the automatic labelling algorithm generated by the method, the conventional labelling process was retained from the dataset upon its verification. In this instance, a comparison was made by applying a class label result to traditional labelling and evaluating its effectiveness in relation to a non-conventional dataset. In a modest manner, the dataset comprising both traditionally labelled and automatically labelled data was subjected to validation through the reporting of results in order to assess the accuracy of the proposed method. Table 4 showcases the computation report denoting the quantity of errors documented in the conventional labeling procedure. Additionally, the objective of computing the error rate through calculation is to determine the level of precision achievable when the algorithm is designed and applied. Consequently, the automatic labelling algorithm that was proposed demonstrated a rate of accuracy of roughly 91%. The accuracy rate has been calculated along with an associated error rate. The error rate is computed by dividing the total number of errors by the total number of documents. In this instance, the error rate is approximately 0. 08974 Conversely, the accuracy rate is 0. 91524.

An additional noteworthy aspect is the creation of a dataset covering Crawler-db, which was generated as a result of data collection carried out specifically for extraction purposes in the dark web. The procedure yields a total of 4750 distinct samples, which are partitioned into five distinct categories. The present study established four major categories, with an additional class

designed to encompass any activities that do not pertain to the established categories. The assortment of documents assessed for each category can be depicted in Table 4 as exemplified. During the subsequent phase, once the automated labelling process had been completed, the findings revealed that drug trafficking accounted for the highest proportion among the percentage of dark web pages that were crawled.

Table 4. Data Extraction

No	The name of class	The number of documents in each class
1	Theft	203
2	Hacking	107
3	Fault	132
4	Drug	5301
5	Fake account	1309

3.2 Results of classification

In this instance, the ensemble of data entities was configured as a dataset that was bifurcated into two segments, namely the training dataset and the testing dataset. The page in question underwent classification via three distinct methods, namely Random Forest, Linear SVC, and Naïve Bayes classifier, resulting in its categorization into five classes. The present study yielded notable findings regarding the performance of three machine learning algorithms, namely LSV, Random Forest, and Naïve Bayes. The LSV algorithm exhibited remarkable superiority over the other two algorithms, achieving an impressive accuracy rate of 94%. In contrast, the Random Forest algorithm achieved a lower accuracy rate of 88%, while Naïve Bayes yielded moderate success with a rate of 92%. The aforementioned outcome is ascribed to the observation that linear support vector machines (SVMs) exhibit improved performance in the classification of high-dimensional textual data.

However, it is essential to critically evaluate the approach used in this scenario with respect to the employed data handling techniques. The lack of performance in the three primary methods under consideration has hindered the attainment of optimization. In spite of the fact that the outcome of performance has been characterized as the manifestation of information inaccuracy, new identifications have surfaced in this segment, indicative of the insufficiency of training data to account for unrecognized data. The dataset implies invalidity when the data remains either inconsistent in training or testing. Additionally, limitations arise in comprehending the classification of dark web. This information presents an opportunity for further exploration and analysis within a scholarly context.

4. CONCLUSION

The challenge in this study involves the utilization of a web crawler to process information, accounting for the laborious effort of crawling on the dark web to gather additional knowledge regarding the object of interest. For the purpose of conducting data mining analysis, a dataset was generated specifically for this study and subsequently subjected to preprocessing using various techniques to ensure data suitability. Furthermore, it can be deemed precarious to engage in the preparation phase of data analysis, as this process facilitates the transformation of data into a format that is better suited for utilization in machine learning algorithms such as recommendation systems. Optimizing the data preparation process by means of purifying and eliminating extraneous components within our dataset can significantly augment the efficacy of web crawler analysis. Moreover, the process of pre-processing data, particularly in the context of data cleaning, was found to enhance the accuracy of feature extraction. The Linear Support Vector Machine (LSVM) exhibited superior performance compared to Random Forest and Naïve Bayes, as determined by selection criteria.

It is inferred that the LSVM exhibits superior performance, thus rendering it the optimal classification algorithm among the three for the purpose of categorizing dark web pages. The

system's performance was deemed notably outstanding, accompanied by an accuracy rate of 92%, a precision rate of 88%, a recall rate of 85%, and an F1-score of 89%.

REFERENCES

[1] N. Tavabi, N. Bartley, A. Abeliuk, S. Soni, E. Ferrara, and K. Lerman, "Characterizing activity on the deep and dark web," *Web Conf. 2019 - Companion World Wide Web Conf. WWW 2019*, vol. 3, no. 2, pp. 206–213, 2019, doi: 10.1145/3308560.3316502.

[2] A. Alharbi *et al.*, "Exploring the Topological Properties of the Tor Dark Web," *IEEE Access*, vol. 9, no. 1, pp. 21746–21758, 2021, doi: 10.1109/ACCESS.2021.3055532.

[3] C. Wang, Q. Xu, X. Lin, and S. Liu, "Research on data mining of permissions mode for Android malware detection," *Cluster Comput.*, vol. 22, pp. 13337–13350, 2019, doi: 10.1007/s10586-018-1904-x.

[4] A. Souri and R. Hosseini, "A state-of-the-art survey of malware detection approaches using data mining techniques," *Human-centric Comput. Inf. Sci.*, vol. 8, no. 1, 2018, doi: 10.1186/s13673-018-0125-x.

[5] S. He, Y. He, and M. Li, "Classification of illegal activities on the dark web," *ACM Int. Conf. Proceeding Ser.*, vol. Part F1483, no. 5, pp. 73–78, 2019, doi: 10.1145/3322645.3322691.

[6] J. H. Park, S. M. Yoo, I. S. Kim, and D. H. Lee, "Security Architecture for a Secure Database on Android," *IEEE Access*, vol. 6, pp. 11482–11501, 2018, doi: 10.1109/ACCESS.2018.2799384.

[7] A. E. H. Hor, G. Sohn, P. Claudio, M. Jadidi, and A. Afnan, "A semantic graph database for BIM-GIS integrated information model for an intelligent urban mobility web application," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. 4, no. 4, pp. 89–96, 2018, doi: 10.5194/isprs-annals-IV-4-89-2018.

[8] A. Kumar, "Improving Database Security in Cloud Computing," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 8, no. 8, pp. 607–611, 2020, doi: 10.22214/ijraset.2020.30962.

[9] F. Thomaz, C. Salge, E. Karahanna, and J. Hulland, "Learning from the Dark Web: leveraging conversational agents in the era of hyper-privacy to enhance marketing," *J. Acad. Mark. Sci.*, vol. 48, no. 1, pp. 43–63, 2020, doi: 10.1007/s11747-019-00704-3.

[10] R. J. S. Raj, M. V. Prakash, T. Prince, K. Shankar, V. Varadarajan, and F. Nonyelu, "Web Based Database Security in Internet of Things Using Fully Homomorphic Encryption and Discrete Bee Colony Optimization," *Malaysian J. Comput. Sci.*, vol. 2020, no. Special Issue 1, pp. 1–14, 2020, doi: 10.22452/mjcs.sp2020no1.1.

[11] M. Y. Alyousef and N. T. Abdelmajeed, "Dynamically detecting security threats and updating a signature-based intrusion detection system's database," *Procedia Comput. Sci.*, vol. 159, pp. 1507–1516, 2019, doi: 10.1016/j.procs.2019.09.321.

[12] E. Kokolaki, E. Daskalaki, K. Psaroudaki, M. Christodoulaki, and P. Fragopoulou, "Investigating the dynamics of illegal online activity: The power of reporting, dark web, and related legislation," *Comput. Law Secur. Rev.*, vol. 38, p. 105440, 2020, doi: 10.1016/j.clsr.2020.105440.

[13] K. Zhang, "Research on Data Mining Security under the Background of Big Data Era," in *International Conference on Management and Computer Science*, 2018, vol. 77, no. Icmcs, pp. 236–239. doi: 10.2991/icmcs-18.2018.48.

[14] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," *Int. J. Inf. Technol.*, vol. 12, no. 4, pp. 1243–1257, 2020, doi: 10.1007/s41870-020-00427-7.

[15] T. Chandrakala, S. N. S. Rajini, K. Selvam, and K. Dharmarajan, "Implementation Of Data Mining And Maching Learing In The Concept Of Cybersecurity To Overcome Cyber Attack Turkish Journal of Computer and Mathematics Education," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 12, pp. 4561–4571, 2021.

[16] C. Tex, M. Schaler, and K. Bohm, "DISTANCE-BASED DATA MINING over ENCRYPTED DATA," *Proc. - IEEE 34th Int. Conf. Data Eng. ICDE 2018*, vol. 1, pp. 1268–1271, 2018, doi: 10.1109/ICDE.2018.00126.

[17] T. Javid, M. K. Gupta, and A. Gupta, "A hybrid-security model for privacy-enhanced distributed data mining," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2020, doi: 10.1016/j.jksuci.2020.06.010.

[18] F. Paquin, J. Rivnay, A. Salleo, N. Stingelin, and C. Silva, "Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic semiconductors," *J. Mater. Chem. C*, vol. 3, no. 2, pp. 10715–10722, 2015, doi: 10.1039/b000000x.

[19] Z. Sun, K. D. Strang, and F. Pambel, "Privacy and security in the big data paradigm," *J. Comput. Inf. Syst.*, vol. 60, no. 2, pp. 146–155, 2020, doi: 10.1080/08874417.2017.1418631.

[20] S. Al-Darraji, D. G. Honi, F. Fallucchi, A. I. Abdulsada, R. Giuliano, and H. A. Abdulmalik, "Employee attrition prediction using deep neural networks," *Computers*, vol. 10, no. 11, pp. 1–11,

2021, doi: 10.3390/computers10110141.
[21] T. Siddiqui, A. Y. A. Amer, and N. A. Khan, "Criminal Activity Detection in Social Network by Text Mining: Comprehensive Analysis," *2019 4th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2019*, pp. 224–229, 2019, doi: 10.1109/ISCON47742.2019.9036157.
[22] T. Imandasari, E. Irawan, A. P. Windarto, and A. Wanto, "Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 750, 2019, doi: 10.30645/senaris.v1i0.81.
[23] M. K. Hossain, M. M. Haque, and M. A. A. Dewan, "A comparative analysis of semi-supervised learning in detecting burst header packet flooding attack in optical burst switching network," *Computers*, vol. 10, no. 8, 2021, doi: 10.3390/computers10080095.