ISSN: 0067-2904

Classifying Illegal Activities on Tor Network using Hybrid Technique

Mohammed Khalafallah Alshammery*[1], Abbas Fadhil Aljuboori [2]
[1]*College of Information Technology, University of Babylon, Babil, Iraq*
[2]*College of Engineering, University of Information Technology and Communications, Baghdad, Iraq*

Received: 23/9/2021 Accepted: 24/11/2021 Published: 30/9/2022

Abstract

With the freedom offered by the Deep Web, people have the opportunity to express themselves freely and discretely, and sadly, this is one of the reasons why people carry out illicit activities there. In this work, a novel dataset for Dark Web active domains known as crawler-DB is presented. To build the crawler-DB, the Onion Routing Network (Tor) was sampled, and then a web crawler capable of crawling into links was built. The link addresses that are gathered by the crawler are then classified automatically into five classes. The algorithm built in this study demonstrated good performance as it achieved an accuracy of 85%. A popular text representation method was used with the proposed crawler-DB crossed by two different supervised classifiers to facilitate the categorization of the Tor concealed services. The results of the experiments conducted in this study show that using the Term Frequency-Inverse Document Frequency (TF-IDF) word representation with a linear support vector classifier achieves 91% of 5 folds cross-validation accuracy when classifying a subset of illegal activities from crawler-DB, while the accuracy of Naïve Bayes was 80.6%. The good performance of the linear SVC might support potential tools to help the authorities in the detection of these activities. Moreover, outcomes are expected to be significant in both practical and theoretical aspects, and they may pave the way for further research.

Keywords: Dark Web, Crawling, Data Mining, Classification, TOR (The Onion Routing)

تصنيف الأنشطة غير القانونية على شبكة تور باستعمال التقنية الهجينة

محمد خلف الله الشمري[1]* وعباس فاضل الجبوري[2]
[1] قسم البرمجيات, كلية تكنولوجيا المعلومات, جامعة بابل, بابل, العراق
[2] قسم هندسة الحاسوب, كلية الهندسة, جامعة تكنولوجيا المعلومات والاتصالات, بغداد, العراق

الخلاصة

توفر حرية الوب العميق مكانًا آمنًا حيث يمكن للأشخاص التعبير عن أنفسهم دون الكشف عن هويتهم ولكن يمكنهم أيضًا القيام بأنشطة غير قانونية. في هذا البحث ، نقدم مجموعة بيانات جديدة لمجالات الوب المظلم النشطة، والتي نسميها قاعدة بيانات الزاحف. قمنا ببناء قاعدة بيانات الزاحف عن طريق أخذ عينات

*Email: mohammed.alshammery@student.uobabylon.edu.iq

لشبكة تور وصممنا زاحف ويب للزحف إلى الروابط التي تجمعها، وقمنا تلقائيًا بتصنيف كل عنوان إلى خمس فئات ، من خلال بناء خوارزمية لوضع العلامات التلقائية بدلاً وضع العلامات اليدوية على مجموعة البيانات. حققت الخوارزمية المقترحة دقة تصل الى 85٪. باستعمال قاعدة بيانات الزاحف. استعملنا أسلوبًا معروفًا لتمثيل النص عبر مصنفين مختلفين تحت الإشراف لتصنيف خدمات تور المخفية. وجدنا أن تمثيل كلمات باستعمال تقنية TF–IDF مع مصنف ناقل الدعم الخطي يحقق 91٪ من 5 أضعاف دقة التحقق من الصحة عند تصنيف مجموعة فرعية من الأنشطة غير القانونية من قاعدة بيانات الزاحف, بينما حثثت خوارزمية Naïve Bayes دقة تصل الى 80.6%. قد يدعم الأداء الجيد للمصنف الأدوات المحتملة لمساعدة السلطات في اكتشاف هذه الأنشطة. علاوة على ذلك ، من المتوقع أن تكون النتائج مهمة لكل من الجوانب العملية والنظرية وقد تمهد الطريق لمزيد من البحث.

1. Introduction

The Dark Web, which is also referred to as the "Darknet", refers to the content found on the internet, such as forums and webpages. These contents are normally encrypted in such a way that the accurate location of the servers hosting them cannot be traced [1]. It is also impossible for the content of the Dark Web to be indexed by traditional search engines like Yahoo, Google, and Bing, etc. [2].

It was in the early 2000s that the term "Dark Web" emerged, and since then, its use has been employed within academic and media contexts [3]. However, its popularity increased as a drugs' market known as "Silk Road" emerged in the year 2011; this market only operated for three years before shutting down in the year 2013 [4].

The Dark Web grants anonymity to the user, the website, and the servers. In other words, on the Dark Web, the user's identity, the owner of the website, and the location of the server remain anonymous. Presently, there are a number of open-source browsers like I2P, Freenet, Tor, etc. that can be used to freely access the content of the Dark Web [5][6]. As of January 2019, the Tor Metrics Project revealed that there were over 120K registered onion addresses in existence [5], and that number increased to 200K in May 2020, with an average of 2 million users using Tor daily [7].

The TOR network emerged in the 1990s when it was developed by the United States Naval Research Laboratory for the purpose of transmitting encrypted military data anonymously. Subsequently, in the year 2004, the original 'The Onion Routing Project' was made freely available to the public under a free and open-source license known as the "Tor Project" [8].

Every computer has a specific address called "Internet Protocol" (IP), which is provided by a local Internet Service Provider (ISP), and there is a link between the IP address and the domain name. The routing of both domain names and IP addresses is done via the ISP's servers. One of the ways through which the location of a user can be identified is by tracing the IP address of the user's device, but TOR makes use of onion routers that bounce a connection via a wide network of relays globally [9]. This way, users as well as the webpage they are accessing remain anonymous. Presently, the majority of illegal activities like buying and selling of confidential records and data, weapons, and drugs are getting stronger within the Tor network due to the level of anonymity it offers [3]. We talked more about the dark web, its characteristics, and the mechanism for accessing it in our previous study (Crawling and Mining the Dark Web: A Survey on Existing and New Approaches).

Given the significance of the contents submerged in the Dark Web and the abuse of this content, this work is focused on designing and developing a system that has the capability of crawling into the Dark Web and classifying illegal activities carried out on the Dark Web. The dataset used in this work was built using a web crawler containing five categories of illicit activities monitored on the Dark Web at the time of sampling. This work is aimed at creating a system that is capable of precisely categorizing the Dark Web by means of the textual content of the Hidden Service (HS). The fixed methodology that is proposed may

significantly contribute to tools used by concerned authorities monitoring the abuse of the Dark Web.

The rest of the paper is organized as follows:

Section 2 presents a review of related works. In Section 3, the proposed technique for crawling and classification is introduced alongside the characteristics of the dataset. Section 4 provides a description of how the proposed technique is implemented and the technical information of the experiment carried out in the work. Section 5 provides a discussion of the results obtained from the experiments. Lastly, Section 6 presents the conclusion drawn from the study and highlights suggested future work.

2. Related Work

Introduced by Graczyk et al. [10], Agora is a popular Darknet black market. The technique introduced by these authors is capable of classifying the goods into 12 groups with an accuracy rate of 79%. Attributes are extracted using TF-IDF while features are collected using principal component analysis (PCA), and then classification of features within their pipeline architecture is performed through the use of SVM.

Baravalle et al. [3] focused their research on the Dark Web e-markets, with a particular emphasis on Agora, an electronic market where fake identities and drugs can be purchased and sold. Before the data was collected, they developed a spider with a few lines of code to simulate human authentication on the market. The crawler performs the simulation of the verification technique for user login with the aim of accessing 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. The miner was developed using command line PHP (and the cURL library) and an object-oriented approach, using MySQL as a backend. The analysis of the data has been carried out with several tools, including Weka and ad hoc Java and Python scripts.

Rahayuda and Santiari [11] crawled the TOR Dark Web, focusing on nine types of domains and defining the services or information hosted by the various domains. Their findings demonstrate how specific types of domains intentionally detach themselves from other TORs. In their work, they made use of fuzzy K-Nearest Neighbor for classification. The results obtained through the crawling system were saved in the database and classified through the use of the fuzzy-KNN method. As a result of this, data was produced by the crawling framework in the form of page information and URL addresses. Lastly, a comparison between the crawling and sample data processes was made by the authors.

Al Nabki et al. [12] in their recent study employed the use of LR, NB, and SVM with two types of text representation models: the TF-IDF and Bag of Words (BoW). The authors created a dataset known as DUTA for their experiments. The total number of samples contained in their dataset at the time it was used was 7K, and the categorization and labeling of the whole samples were done manually. Furthermore, the dataset was divided into 26 categories, including other categories of illicit activities such as child pornography and drug trafficking. The researchers reported that the combination of TFIDF text representation and Logistic Regression classifier made the proposed system achieve a macro F1 score of 93% and a precision rate of 96.6% over 10 folds of cross-validation.

In the work done by Siyu He et al. [5], a method of classification was introduced; the training data used for the model is the ‘Federal Code of the United States of America’. Their experimental results revealed that combining the Naïve Bayes Classifier and TF-DIF feature extraction yielded a 93 percent accuracy rate.

In their work, Khare et al. [13] introduced a smart crawler that can assist experts in searching the Deep Web efficiently. The proposed crawler begins the process of crawling from the center page of the seed URL, moving towards the last link. This crawler makes it possible to

sort active links and inactive links so that they can be separated based on the request to the webserver of sites. Furthermore, the crawler that they presented in their work is equipped with a text-based site classifier. They used two techniques to classify the deep web content: by using a neural network with supervised machine learning techniques, we achieved the best 95.46% of accuracy in classifying sites, while the machine learning algorithm Logistic Regression with TF-IDF has an accuracy F1-score of 94%.

3. Methodology

In this work, the method proposed to facilitate the effective detection and classification of illicit Dark Web activities is introduced, and its framework can be seen in Figure 1. The development of the model is done in two stages; the first stage involves the creation of a dataset, and the second stage involves the application of the classification method. The main contribution of this work lies in the first and second stages, which are the application of the proposed method. In this regard, the contribution focused on creating a new dataset through designing a crawler system and introducing a proposed algorithm to label the dataset automatically and improving the technical information and enhancing the parameter adjustments to the basic model of classification so that better results can be achieved when the model is used; through using different types of algorithms that perform the tasks involved in data preprocessing, such as data cleaning, tokenization of words, stop word removal, and conversion of HTML pages to text documents. In this study, a range of machine learning classifiers and weighting methods were combined, and afterward, an evaluation of the results was carried out.

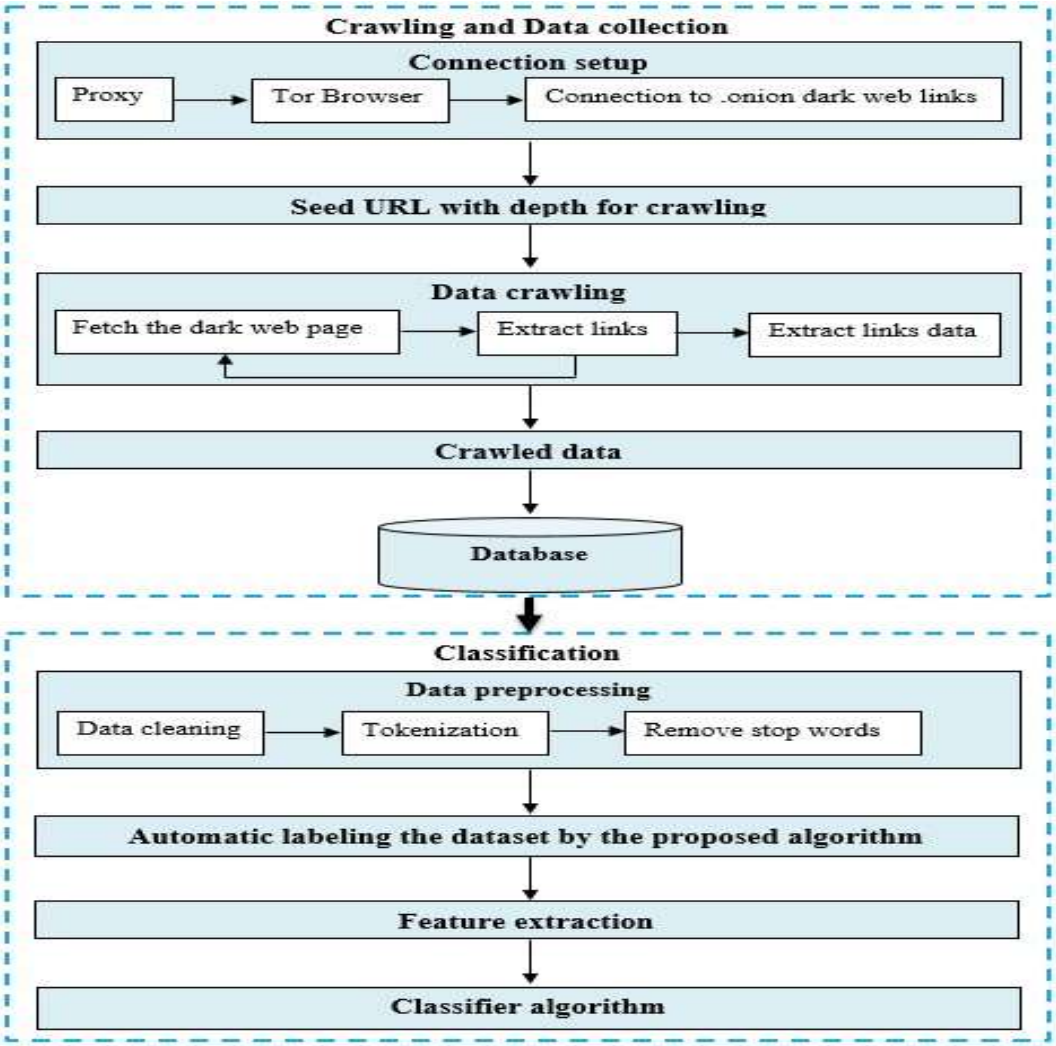


Figure 1 -Dark Web Crawling and Classification System framework

4. Experimental Setup

The setup for the experiment is described in this section. The experiment was carried out to evaluate the performance of the newly proposed crawling and classification technique in an experimental setting. More so, the section contains the exact technical details that were enhanced in the standard model. Furthermore, the performance of the model was improved so as to get optimal performance from the model by comparing the combination of a wide range of weighting techniques and machine learning classifiers. This was aimed at adjusting to the most appropriate parameters.

4.1. The Dataset

The dataset used in this study was created by the manual collection of many Dark Web links from numerous websites, out of which 'ahmia' is the most popular. These links are crawled through by the proposed links, automatically collecting all other links related to them. With the proposed system, the contents of tens of thousands of Dark Web market sites can be collected. The system achieves this by moving every hyperlink that leads to other new sites. Afterwards, they are processed, and then the process of data extraction is implemented, followed by the storage of a database for further analysis. Over 2,300 active links as well as their contents were obtained. The Dark Web pages which were crawled by the crawler were saved using MongoDB.

The dark crawler has the ability to concurrently access Tor and the public internet. It also has the ability to carry out an automatic search of websites in the Tor network according to pre-defined links, and after the websites are found, they are saved in a database. The Tor network was infiltrated using third-party local HTTP proxy software, which is referred to as SOCKS5. This software establishes the connection between the Tor network and the dark crawler without difficulty. Figure 2 shows the crawler data flow chart.

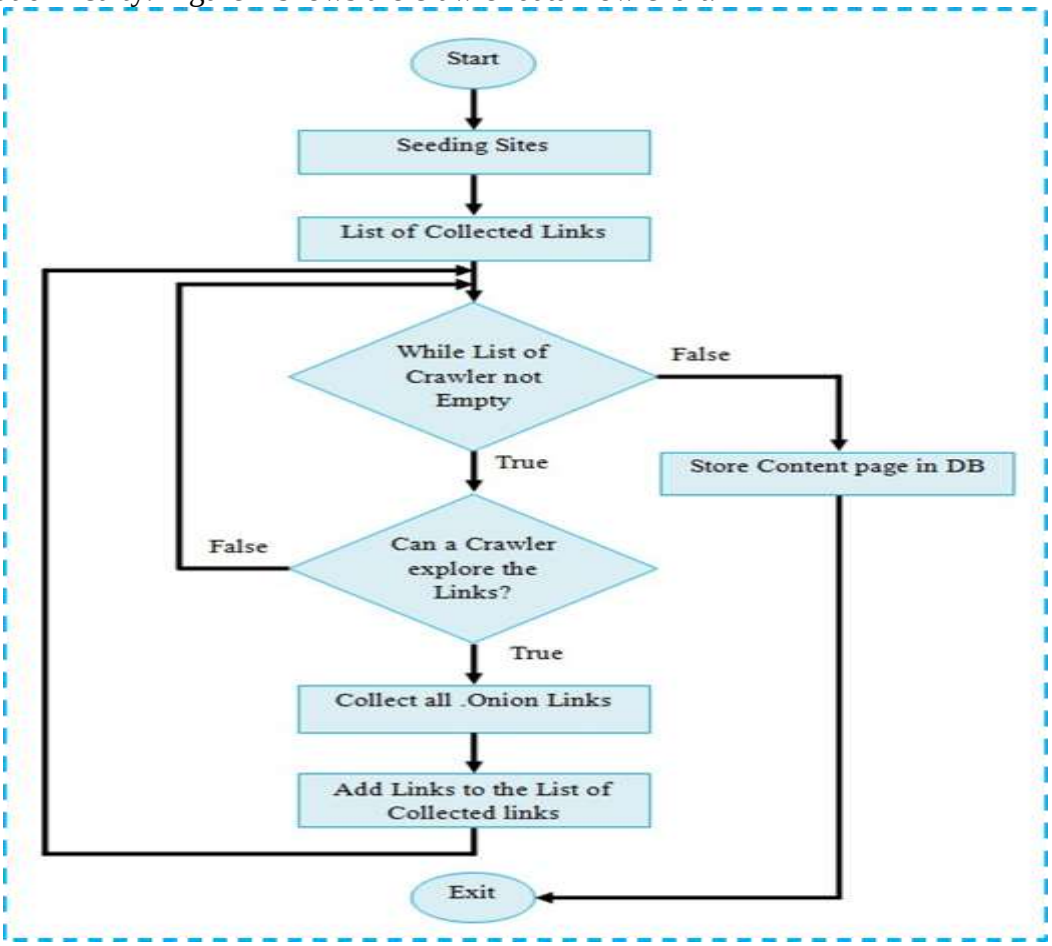


Figure 2-Crawler Architecture

4.2. Classification Method

Our classification implementation consists of four stages. This includes the preprocessing for crawled data, the proposed automatic labeling algorithm, the feature extraction, and machine learning classifiers.

4.2.1. Text Preprocessing

At this stage, the content of the Dark Web pages that have been retrieved by the Dark Crawler is subjected to preprocessing, because upon retrieval, they are not in the desired format. Therefore, subjecting the web pages to preprocessing is very critical to ensure that the contents are free of inappropriate details and noise.

In this work, a variety of preprocessing techniques were employed, including tag removal, removal of punctuation, elimination of numbers, tokenization, and elimination of stop words through the use of many libraries in the Python language.

4.2.2. Automatic labeling algorithm

According to researchers, one of the main challenges facing the classification of dark web content is manual labeling of the dataset, which in turn requires significant time and effort. In this research, an algorithm has been built and named the "automatic labeling algorithm," which is used in labeling the dataset automatically. Consequently, this algorithm in turn saves time and effort. To the best of our knowledge, it is considered an addition to scientific research. The automatic labeling algorithm has classified the dataset into five classes, including fake ID, drugs, hacking, and weapons. The four categories were selected on the basis of security concerns and additions to other classes. In particular, out of all the activities on the dark web, these classes are a direct threat to societal security, so they were focused. Figure 3 presents the counts of illicit Dark Web activities categories in the dataset after it has been automatically labeled.

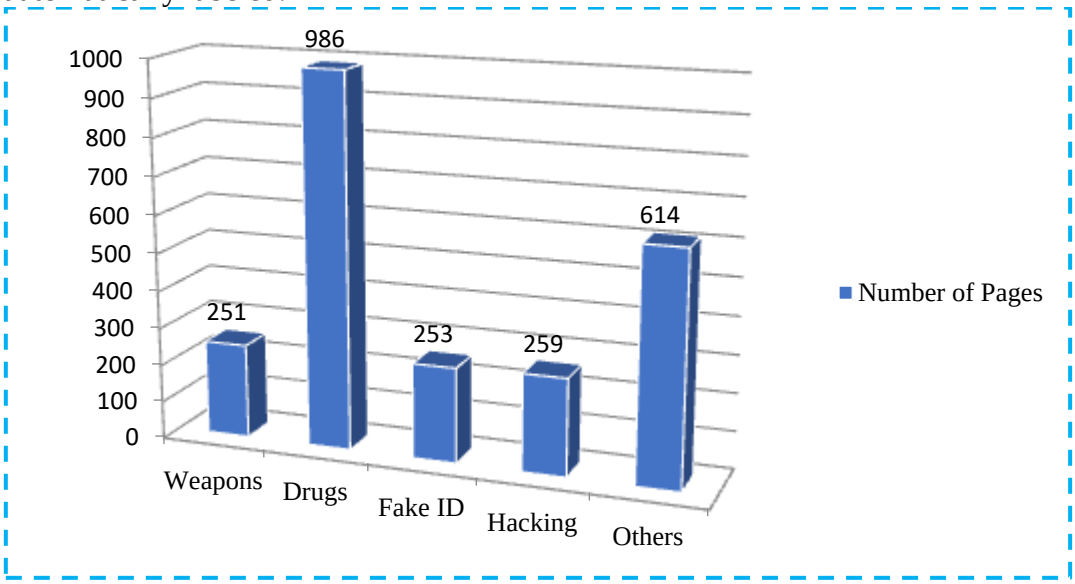


Figure 3- Number of pages of each class in the dataset

4.2.3. Features Extraction

Subsequent to the preprocessing and automatic labeling of the dataset, the vector space model was built using the widely used Term Frequency-Inverse Document Frequency model (TF-IDF); the TF-IDF is a model that is commonly applied in data mining and retrieval of information. The sequence of words in the document is not considered in the TF-IDF approach, but it is important since it indicates how relevant each word is, making it easier to compute. The basic notion is that if a phrase or word appears more frequently (TF) in one document and seldom in other documents in the dataset, then it is believed that the category

differentiation capability of the phrase or word is good and can be used for the purpose of classification.

4.2.4. Classifier Selection

The proposed classification model was implemented using two machine learning classifiers, including Naïve Bayes (NB) and Support Vector Machine (SVM). Afterwards, the efficiency of the proposed method was compared in different pipelines. The results obtained from the experiments are presented in Section 5 of this paper.

5. Results and Discussion

The results obtained from the experiments performed in this study are presented and discussed in this section.

5.1. Results of Features Extraction

The reliability of the training set introduced in this study was determined by analyzing the results of illegal documents weighted based on terms of TF-IDF. This step yielded its results in the form of a list of the features alongside the weight of individual features. Table 1 shows the results of samples of relevant features for every class after the application of the feature extraction process to the dataset. From the table, it can be clearly observed that the trained features are representative and can be used as the basis for classification.

Table 1-Sample the Features for Each Category

No.	Categories	Important Keywords
1	Drugs	'Drug', 'Heroin', 'Cocaine', 'LSD', 'Cannabis', 'Hush', 'Chemical', 'MDMA', 'Meth',..]
2	Weapons	['Weapon', 'CZ', 'Rifle', Smith, 'Ammunition', 'Gun', 'Manufacturer', 'Glock', 'Ammo', 'Beretta',...]
3	Hacking	['Hack', 'Attack', 'Malware', 'Raid', 'Spyware', 'Threat', 'Offense', 'Email', 'Spam',..]
4	Fake ID	['License', 'Card', 'Document', 'Fake', 'Passport', 'Identitycard', 'Residence', 'Money', 'Counterfit',...]

5.2. Result of evaluating the automatic labeling algorithm

In order to ensure that the automatic labeling algorithm proposed in this work works accurately, the labeling of the dataset was done manually, and a comparison of the result of the class label with both of them was carried out by calculating the error rate between manually and automatically labeling the dataset by using the equations 1, 2 below.

$$\text{Error rate} = \text{number of error} / \text{total number of documents} \quad (1)$$

$$\text{Accuracy} = 1 - \text{Error rate} \quad (2)$$

Error rate = 343/2363 = 0.145

Accuracy = 1 – 0.145 = 0.854

Based on the experimental results of this study, the algorithm proposed in this paper achieved an accuracy rate of 85 percent.

The dataset created in this work (crawler-BD) is made up of 2363 samples that were divided into five classes (the initial four classes and the others). The number of samples of Dark Web pages found in each class can be seen in Table 2 below. Upon the completion of the automatic labeling task, it was observed that the pages of the drug trade had the highest percent of all the Dark Web pages that were crawled.

Table 2-The Crawler-DB Dataset Classes

No.	The name of class	The number of pages in each class
1	Drugs	986
2	Others	614
3	Hacking	259
4	Fake ID	253
5	Weapons	251

5.3. Results of Classification Methods

Table 3 shows in detail the analysis of the experimental results. TF-IDF term weighting technique and two classifiers: 1) Gaussian Naive Bayes Classifier; 2) Linear Support Vector Classifier were used. Using the linear SVC, the proposed technique demonstrated superior performance, with an accuracy rate of 91%; in several rounds of verification trials, it also proved to be stable. Nevertheless, the performance of the model when used with the Gaussian NB classifier was found to be lower, achieving an accuracy rate of 80.6%.

The experimental results of the performance of the proposed technique are presented in Table 4. The performance was determined using recall, accuracy, F1-Score, and precision as the parameters for performance evaluation. The aforementioned parameters are commonly used within the field of information retrieval.

Table 3-The Performance Metric of the Two Algorithms

Classifiers	Accuracy	Precision	Recall	F1-score
SVM	91%	89%	88%	88%
Gaussian NB	80.6%	78%	77%	77%

Figure 4 shows the Accuracy, Precision, Recall, and F1-score of the NB Algorithm over the five classes.

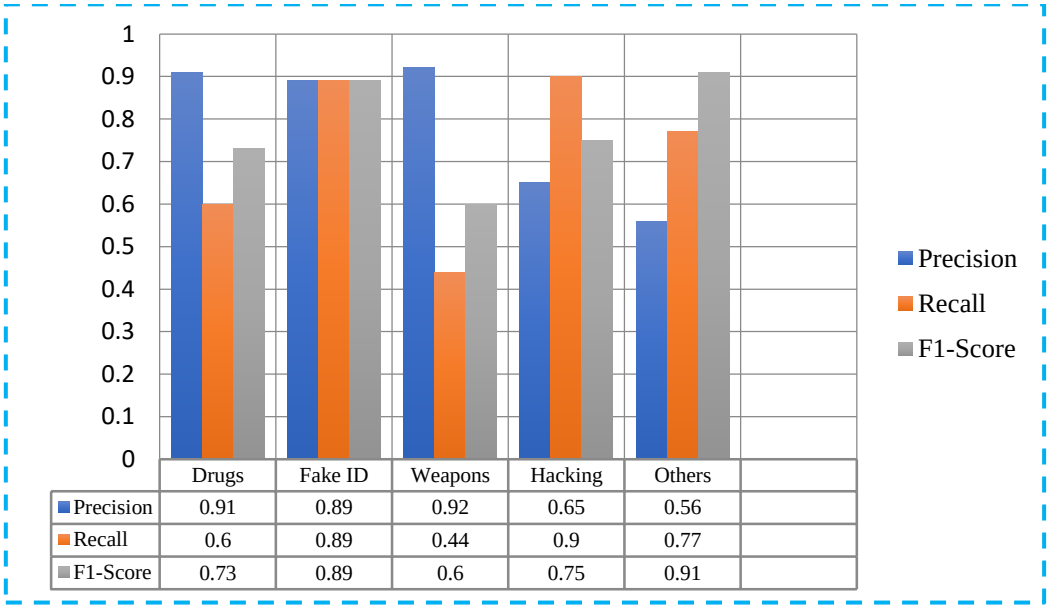
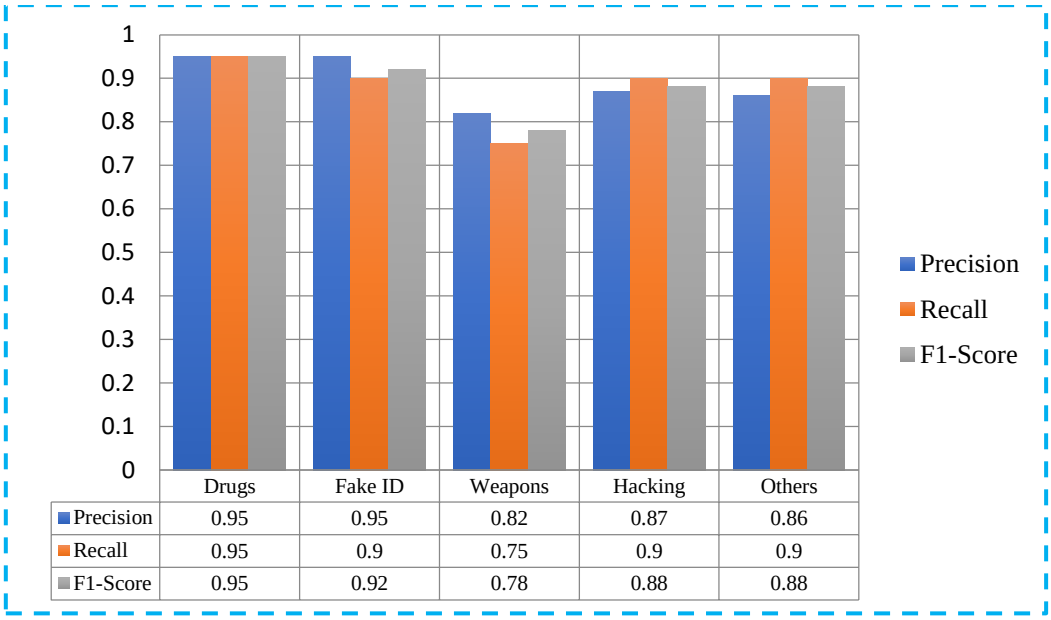


Figure 4-The Performance Metrics of the NB Algorithm

Figure 5 shows the Accuracy, Precision, Recall, and F1-score of the SVM Algorithm over the five classes.



	Drugs	Fake ID	Weapons	Hacking	Others
Precision	0.95	0.95	0.82	0.87	0.86
Recall	0.95	0.9	0.75	0.9	0.9
F1-Score	0.95	0.92	0.78	0.88	0.88

Figure 5-The Performance Metrics of the SVM Algorithm

It can be seen from the learning curve for the TF-IDF pipeline presented in Figure 6 that the algorithm is learning correctly, as there is a rise in the validation accuracy curve and an increase in the classification accuracy, which is projected through the increase in the number of samples. More so, the curve of the training accuracy is stable. With this high accuracy rate, a standard model that has the capability of detecting illicit Dark Web activities can be developed.

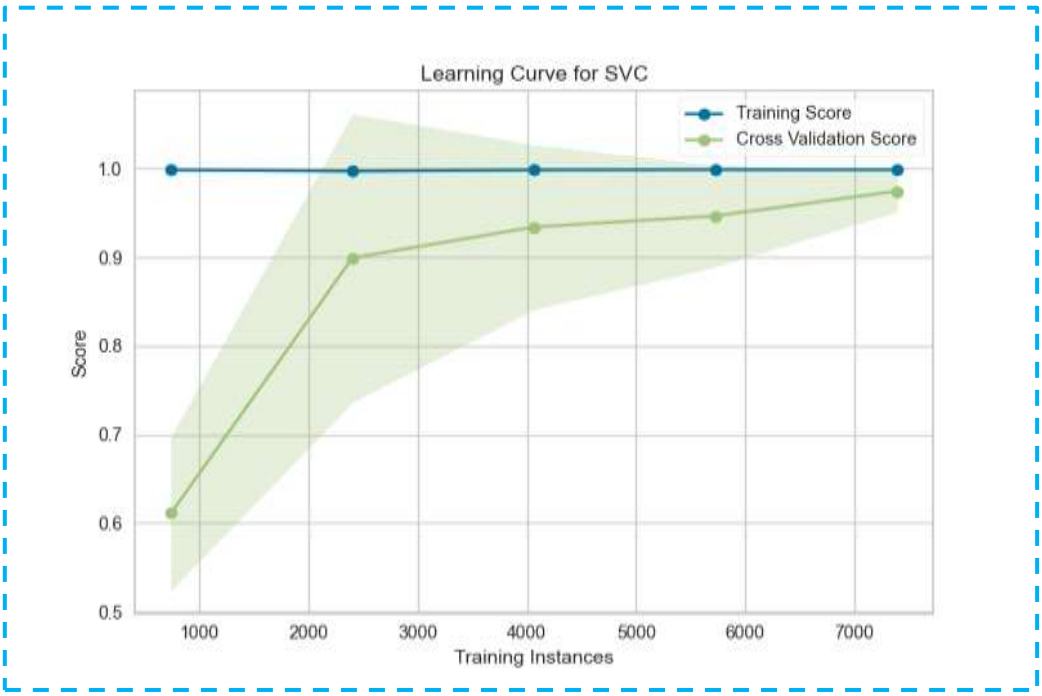


Figure 6-Shows the Learning Curve of the linear SVC

As mentioned in previous sections of this research, the problem of manually labeling the dataset has been solved in this study, as an automatic labeling algorithm was proposed. To the best knowledge of the researcher, this is the first of its kind. In other words, in this study, an algorithm was designed to label the dataset automatically. Additionally, a web crawler that is

capable of extracting an unlimited amount of data from Dark Web pages has been proposed. The crawler achieves this by exploring the hyperlinks of the pages, navigating around them, extracting their data, and producing a new dataset that can be analyzed for purposes of scientific research. Furthermore, a prediction model was built to classify illegal activities that are carried out on the Dark Web, and the proposed model achieved high accuracy.

6. Conclusion and Future Work

In this article, illicit activities of Tor's hidden service have been classified through the use of the TF-IDF technique, which was applied together with two classifiers (NB and SVM). The dataset (crawler-DB) was built so that it could support classification pipelines; the crawler-DB consisted of 2363 samples that were automatically labeled into five categories. Rather than performing the task of labeling through manual means, an algorithm was built in this work for automatic labeling, and superior performance was demonstrated by the algorithm in terms of accuracy and effectiveness. These five classes, including the others that are related only to illegal activities (drugs, weapons, hacking, fake ID, and others), were selected and used for training the proposed model. The results of the experiments showed that the linear SVC with TF-IDF demonstrated superior performance as an algorithm for the classification of Dark Web pages. Overall, the proposed system yielded good results in terms of recall, precision, accuracy, and F1-Score. The accuracy is 91 %, precision is 89%, recall is 88%, and the F1-Score is 88%.

For future work, the enlargement of the dataset is proposed. This is to be achieved by carrying out a deeper search of the Dark Web through the addition of more HS sources. Furthermore, in this study, the best accuracy rate was recorded for the Support Vector Machine. Nevertheless, adaptations can be made to the SVM so that the possibility of inaccurate classification can be avoided while accuracy is improved. The validity and accuracy of the proposed automatic labeling algorithm can be further evaluated through its application to a different dataset. Furthermore, improvements can be made to the dataset so that its accuracy can be improved.

References

[1] S. Ghosh, P. Porras, V. Yegneswaran, K. Nitz, and A. Das, "ATOL: A framework for automated analysis and categorization of the dark web ecosystem," *AAAI Work. - Tech. Rep.*, vol. WS-17-01-, pp. 170–178, 2017.

[2] X. Zhang and K. P. Chow, "A framework for dark web threat intelligence analysis," *Int. J. Digit. Crime Forensics*, vol. 10, no. 4, pp. 108–117, 2018, doi: 10.4018/IJDCF.2018100108.

[3] A. Baravalle, M. S. Lopez, and S. W. Lee, "Mining the Dark Web: Drugs and Fake Ids," *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 0, pp. 350–356, 2016, doi: 10.1109/ICDMW.2016.0056.

[4] S. Nazah, S. Huda, J. Abawajy, and M. M. Hassan, "Evolution of Dark Web Threat Analysis and Detection: A Systematic Approach," *IEEE Access*, vol. 8, pp. 171796–171819, 2020, doi: 10.1109/access.2020.3024198.

[5] S. He, Y. He, and M. Li, "Classification of illegal activities on the dark web," *ACM Int. Conf. Proceeding Ser.*, vol. Part F1483, pp. 73–78, 2019, doi: 10.1145/3322645.3322691.

[6] A. Montieri, D. Ciuonzo, G. Aceto, and A. Pescape, "Anonymity Services Tor, I2P, JonDonym: Classifying in the Dark (Web)," *IEEE Trans. Dependable Secur. Comput.*, vol. 17, no. 3, pp. 662–675, 2020, doi: 10.1109/TDSC.2018.2804394.

[7] H. Thorat, S. Thakur, and A. Yadav, "Categorization of Illegal Activities on Dark Web using Classification," no. May, pp. 1230–1234, 2020.

[8] D. R. Hayes, F. Cappa, and J. Cardon, "A framework for more effective dark web marketplace investigations," *Inf.*, vol. 9, no. 8, pp. 1–17, 2018, doi: 10.3390/info9080186.

[9] M. Chertoff, "A public policy perspective of the Dark Web," *J. Cyber Policy*, vol. 2, no. 1, pp. 26–38, 2017, doi: 10.1080/23738871.2017.1298643.

[10] M. Graczyk and K. Kinningham, "Automatic Product Categorization for Anonymous

Marketplaces," *Comput. Sci.*, pp. 1–6, 2015.

[11] I. G. S. Rahayuda and N. P. L. Santiari, "Crawling and cluster hidden web using crawler framework and fuzzy-KNN," *2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017*, 2017, doi: 10.1109/CITSM.2017.8089225.

[12] M. W. Al Nabki, E. Fidalgo, E. Alegre, and I. De Paz, "Classifying illegal activities on tor network based on web textual contents," *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 1, pp. 35–43, 2017, doi: 10.18653/v1/e17-1004.

[13] A. Khare, A. Dalvi, and F. Kazi, "Smart Crawler for Harvesting Deep web with Multi-Classification," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, 2020, doi: 10.1109/ICCCNT49239.2020.9225369.