Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Babylon
College of Information Technology
Software Department

Design an Integrated System for Crawling and Classification of the Dark Web

A Thesis

Submitted to the Council of the College of Information Technology for Postgraduate Studies of University of Babylon in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology-Software

By

Mohammed Khalafallah Mohammed Nuzzal

Supervised by

Prof. Dr. Abbas Fadhil Aljuboori

2022 A.D. **1443 A.H.**

بِسم اللهِ الرَحمنِ الرّحيم

(رَبِّ أَوْزِعْنِي أَنْ أَشْكُرَ نِعْمَتَكَ الَّتِي
أَنْعَمْتَ عَلَيَّ وَعَلَىٰ وَالِدَيَّ وَأَنْ أَعْمَلَ

صَالِحًا تَرْضَاهُ)

صَدَقَ اللهُ العَليُ العَظيم
سورة الاحقاف (١٥)

Declaration

I hereby declare that this thesis entitled “**Design an integrated system for Crawling and Classification of the Dark Web**”, submitted to the University of Babylon in partial fulfillment of requirements for the degree of Master in Information Technology \ Software, has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose source is appropriately cited in the references.

Signature:
Name: **Mohammed Khalafallah Mohammed**
Date: / /2022

Supervisor Certification

I certify that the thesis entitled "**Design an integrated system for Crawling and Classification of the Dark Web**" was prepared under my supervision at the department of Software/ College of Information Technology/ University of Babylon as partial fulfillment of the requirements of the degree of Master in Information Technology-Software.

Signature:

Supervisor Name: **Prof. Dr. Abbas Fadhil Aljuboori**

Date: / /2022

The Head of the Department Certification

In view of the available recommendations, I forward the thesis entitled "**Design an integrated system for Crawling and Classification of the Dark Web**" for debate by the examination committee.

Signature:

Assistant Prof. Dr.Ahmed Saleem Abbas

Head of Software Department

Date: / /2022

Certification of the Examination Committee

We hereby certify that we have studied the dissertation entitled (**Design an integrated system for Crawling and Classification of the Dark Web**) presented by the student (**Mohammed Khalafallah Mohammed**) and examined him/her in its content and what is related to it, and that, in our opinion, it is adequate with (Viva Result) standing as a thesis for the degree of Master in Information Technology-Software.

Signature:
Name: Dr. Sattar B. Sadkhan
Title: Professor
Date: / / 2022
(**Chairman**)

Signature:
Name: Dr. Saif Ali Abd Al Radha Alsaidi
Title: Professor Assistant
Date: / / 2022
(**Member**)

Signature:
Name: Dr. Wadhah Razooqi Baiee
Title: Lecturer
Date: / / 2022
(**Member**)

Signature:
Name: Dr. Abbas Fadhil Aljuboori
Title: Professor
Date: / / 2021
(**Member and Supervisor**)

Approved by the Dean of the College of Information Technology, University of Babylon.

Signature:
Name: Dr. Hussein Atiya Lafta
Title: Professor
Date: / / 2022
(**Dean of Collage of Information Technology**)

Dedication

To those who sacrificed their lives so that we can live. To the martyrs of the Ministry of Interior who died defending our beloved Iraq.

To whom that God said in Quran "And do well to parents", my mother and my father, may God gives them long life.

To my sisters, brothers, and dear wife.

I dedicate this work.

Mohammed Khalafallah Alshammery

Acknowledgements

First and foremost, I would like to thank God for giving me the ability and strength to complete this thesis.

I would like to express my gratitude to my supervisor Prof. Dr. Abbas Fadhil Aljuboori for his guidance, encouragement, and advice throughout the work of this thesis.

I would like to thank everyone who helped me with a word of encouragement or scientific information that benefited me in this research: Asst. Prof. Abdul Hadi Mohammed Alaidi, and all my teachers who supported me throughout my academic study for a master's.

Last but not least, I would like to thank all my friends and especially the friends who supported me in any way during the creation period of this thesis.

Mohammed Khalafallah Alshammery

Abstract

Illegal activities are often shielded by the robust anonymity and hard-to-track mechanisms provided by the dark web. The dark web is made up of a variety of illicit contents that are often updated. In conventional classification of dark web, large-scale webpages are used for supervised training. Nevertheless, recent research has been faced with the challenges of collecting sufficient illegal dark web content and the time required for the manual labelling of webpages.

The proposed system in this research consists of four main stages. In the first stage, data is collected by our crawler system. In this research, a new dataset known as 'dark web-db' is introduced for active domains on the dark web. It was created using a smart crawler that has the ability to collect data from the dark web pages, explore and navigate through hyperlinks, simultaneously. To achieve this, The Onion Routing (TOR) network was sampled. The second stage involves pre-processing the data using a variety of pre-processing techniques. This stage comes immediately after each successful crawl, and afterwards, the results are stored in the database. The third stage involved the proposed of an algorithm known as automatic labelling algorithm, which was applied on the dataset, and each address was labeled automatically into five categories. The results showed that an accuracy rate of 85% was achieved by the algorithm. In the fourth stage, Linear Support Vector Machine (LSVM), Random Forest (RF), and Naive Bayes (NB) classifiers were used for the application of the classification method.

The use of several techniques was employed so as to determine the most accurate one in terms of classification of dark web data. In addition, the results of the performance of the classification algorithms were assessed using a number of performance parameters including accuracy, precision, recall, and F1-score. Based on the results, the LSVM yielded the highest accuracy. The accuracy is 91%, and f1-score are 88%. Given that, the classifier demonstrated superior performance, concerned authorities may find it relevant in terms of supporting potential tools for the detection of such activities. It is also expected that the results will be of practical and theoretical importance, and the results may serve as a direction for future work.

Declaration Associated with this Thesis

Some of the works presented in this thesis have been accepted as listed below.

- Crawling and Mining the Dark Web: A Survey on Existing and New Approaches.
- Classifying Illegal Activities on Tor Network using Hybrid Technique.

Table of Contents

Chapter One: Introduction

Chapter Two: Theoretical Background

Chapter Three: The Proposed System

Chapter Four: The Experimental Results and Disscussion

Chapter Five: Conclusions and Future Works

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
DM	Data Mining
FN	False Negative
FP	False Positive
HS	Hidden Services
I2P	Invisible Internet Project
LSVM	Linear Support Vector Machine
NB	Naive Bayes
NLP	Natural Language Processing
RF	Random Forest
SVM	Support Vector Machine
TM	Text Mining
TN	True Negative
TOR	The Onion Routing
TP	True Positive
WWW	World Wide Web

List of Algorithms

Chapter One

Introduction

Chapter One
Introduction

1.1 Overview

If the web is regarded as an ocean of data, then the Surface web is nothing short of the slight waves that stay afloat [1], whereas, the bottom contains a wealth of sunken information that is yet to be reached by the conventional search engines [2]. There are two major divisions of the web, including surface web and deep web [3]. The surface web refers to the part of the web which can be crawled and indexed by the typical search engines, like Bing or Google. Regardless of this, a huge portion of the web is yet to be indexed as a result of the large size and absence of hyperlinks, meaning that they not referenced by the other webpages. The part that is undetectable when a search engine is used is referred to as deep web [4]. More so, the content may be tamper-proof and can only be accessed through human interaction; either through the use of a log-in detail or by means of CAPTCHA [5]. Such pages are called "database-driven" websites. Also, the layers of the web that are beneath are not assessed by traditional search engines, and as such, are unable to reach the deep web. Within the deep web is a subset referred to as the dark web, which is cannot isolated and indexed [6]. The requirement for accessing the dark web is either a dedicated proxy server or specially designed software [1]. Virtual sub-network of the World Wide Web (WWW) is required for the functionality of the dark web, because it provides network users with an extra layer of anonymity [7]. Some of the commonly used dark web are Freenet, Invisible Internet Project (I2P), and The Onion Router, which is also referred to as TOR. In the TOR community, dark web websites are known as "Hidden Services" (HS) which can be accessed through the use of a special browser called TOR Browser [8].

There are several positive research reports about the deep web. For instance, just deep web alone contains over 550 billion individual documents as compared with the surface web which contains only 1 billion [9,10]. In addition, in other studies, they emphasized the immensity of the deep web which is 400 to 500 times larger than the surface web [11]. The concepts of deep web and dark web have been in existence since the WWW was established. However, it has recently gained popularity as a result of the FBI arrest of Dread Pirate Roberts, who owns the Silk Road black market in October 2013. Prior to his arrest, the sales on Silk Road were estimated by the FBI to be 1.2 Billion dollars, with a trading network of about 150,000 customers whose identities were anonymous. They also had about 4,000 vendors [12,13]. In the area of dark web, cryptocurrency has emerged as a hot subject due to the fact that it makes the identities of trade parties hidden, while financial transactions remain anonymous[14].

Usually, the dark web is characterized by illicit activities, amounting to about 57% [1].The authors in study [6] predicted that out of 1,000 samples of dark web Content, 68% of activities and content will be illicit. In another study, where by 5,000 Onion domains were analysed, it was found that TOR HS are commonly used for illegal and criminal purposes, like promotion of drug use, weapon use, and pornographic content [15].

Based on the statistics reported by TOR Metrics in January 2019, dark web domains is rapidly growing, with over 120,000 registered onion addresses, with about 2 million day-to-day use of TOR. Nevertheless, such domains that can be accessed publicly are not more than 10% of the total number of extant dark web Domains. This is because of the ambiguity that surrounds the dark web [6].

1.2 Problem Background

Building a very robust crawler is one of the ways through which data from the dark web can be extracted. The crawler task is to retrieval of the dark web pages, this process that begins with seeds URLs, then downloading of all pages under the addresses that have been selected, extraction of hyperlinks from the pages and adding them to the list of addresses, crawling for each links that have been explore, and so on [16].

The process of dark web crawling it is confronted by several limitations that exist because of the features of TOR network, especially uncorrelated websites, in which connections between sites are sparse, rendering it difficult for the crawler to follow [10].

Researches that have been carried out previously show that the link address (URL) lifespan of dark web sites hosted on private encrypted networks is shorter than those of websites on the surface web due to the fact that they are constantly transferred across a wide range of addresses. Also, for the avoidance of surveillance, websites are shifted by web administrators among a variety of web addresses, particularly, within dark web electronic markets are the most worrisome challenge.

Consequently, it is important to develop and implement algorithms and methods that can help in extracting, and classifying data. Given the above situations, this research focuses on introducing a techniques through which the successfully crawling, the effective classification of illegal activities on the dark web can be achieved.

1.3 Problem Statement

In this research, a number of problems were solved which faced by researchers in analyzing the dark web, as follows:

1. The problem of data collection: One of the most important problems that researchers face is the difficulty of collecting data from the dark web, due to the mechanism of designing those sites, the difficulty of tracking them, the strong anonymity that characterizes the dark web, as well as the lack of a dataset because the dark web content is diverse and constantly updated.

2. Problem with the broken links: It has been observed in previous works that the crawling process takes place in two stages, the first is to collect tens of thousands of web links, and the other stage is to crawl those links. Since the most prominent challenge is the short lifespan of dark web sites, and their change constantly, the largest percentage of the collected links are broken. The proposed system ensures retrieval of the effective links only by exploring and navigating to hyperlinks.
3. Manual labelling problem: One of the important problems is the manual labelling of the dataset after the data collection process which requires a lot of time and effort. Proposed an algorithm to solve the problem of manual labelling the dataset.

1.4 Research Motivation

The summary of the major motivations of this research is given below in the following points:

1. It is crucial to analyse the dark web due to the continuously increasing illegal web activities and cybercrime. This will enable the monitoring and mitigation of criminal activities by law enforcement agencies.

2. Recently, the efforts made by government organizations to combat the activities of the dark web market have proven abortive, and this is one of the driving forces behind this research. This will help in providing greater insights on the services and products available within these dark web Markets.

3. Motivated by fact that there are relevant contents embedded within the dark web, and the rate at which such contents are abused, this research focuses on designing and building a system that is able to crawl and classify the illegal activities carried out on the dark web.

1.5 Research Objectives

This research is mainly aimed at proposing techniques and solutions that are based on machine learning (ML); this will help in monitoring suspicious content of online resources, especially the contents of the TOR network. For this reason, a wide range of methods and approaches have been explored and proposed using Machine Learning. The method that is proposed in this research is specially designed to be applied in real-time services and tools. Based on the main objective of this research, the following specific objectives serve as a guide for this research:

1. Build a smart crawler that can access the dark web, extract data, and then store it in the database used in this research.
2. Build a text classifier that is robust enough to detect and ascertain the category of illegal content that is present on the dark web.

1.6 Research Contributions

1. Building a dataset containing 2,363 onion domains that have been mined from the TOR. In the dataset introduced in this research, there were 5 categories covering a wide range of events monitored on the TOR dark web during the period of sampling.
2. In this research, an algorithm has been built and named automatic labeling algorithm, which is used in labeling the dataset automatically. It was observed that in previous related works, the authors manually labelled datasets. Consequently, this algorithm was designed to achieve the task of labelling automatically, which in turn saves time, and it is an addition to scientific research.

1.7 Related Works

The author, Kaur (2014) [17], introduced an informative survey which covers several techniques used for the classification of web content. The author placed emphasis on how crucial these techniques are to the mining of data. Additionally, techniques of pre-processing that could be useful in discover of features, were provided in the survey. Some of these include, stemming, punctuation marks, and elimination of HTML tags.

Graczyk et al. (2015) [18] suggested a pipeline in order to enable the classification of Agora's goods; 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 feature within their pipeline architecture is performed through the use of SVM.

Moore et al. (2016) [15] have proposed a recent analysis focused on TOR secret services to explore and identify the Darknet. Initially, they gathered 5K TOR onion page samples and classified them into 12 classes using an SVM classifier with an accuracy rate of 87%.

Baravalle et al. (2016) [19] the focus of their work was the DarkWeb e-markets, with special attention on Agora, which is an electronic market where fake identities and drugs can be purchased and sold as well. Before the data is collected, they developed a spider with a few lines of code to simulate human authentication on 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 with several tools including Weka and ad hoc Java, and Python scripts.

Rahayuda and Santiari (2017) [20] crawled the TOR Dark Web, focused on nine kinds of domain and defined the service or information hosted by the various domains. Their findings showed the manner in which particular kinds of domains intentionally detach themselves from the other TOR, alongside. In their work, they made use of fuzzy K-Nearest Neighbour for classification. The results obtained through the crawling system were saved in the database and classified through the use of a fuzzy-KNN method. As a result of this, data was produced by the crawling framework in form of page information and URL addresses. Lastly, a comparison between the crawling and sample data processes was done by the authors.

Al Nabki et al. (2017) [5] published a recent report that classified TOR HS's criminal activities using two text representation systems, TF-IDF and BOW, as well as three classifiers, SVM, LR, and NB. They created dataset DUTA, which contains 7K samples labeled manually into 26 categories, including the Others class, which is only concerned with illicit activities such as drug trafficking and child pornography. They discovered that integrating the TFIDF text representation with the Logistic Regression classifier would achieve 96.6 percent precision and 93.7 percent macro F1 score over ten folds of cross-validation.

Marin et al. (2018) [21] discussed an approach of mining communities of malware and exploited vendors, focusing on understanding the vendors and their common characteristics. They used clustering methods and different similarity measures to find correlations among vendors' profiles from several markets, common categories they promote, and the number of products they have in each category. They discussed how this method helps security agencies in understanding and tracking hacking-related communities.

Pannu et al. (2019) [22] presented a crawler approach that serves a search engine specialized for detecting information from suspicious and malicious websites. Its main process starts from downloading the HTML files of the pages from a previously instantiated list of seed URLs, following the links to other Tor websites for scrapping as well. After archiving the pages, the system scans them to fetch information.

Siyu He et al. (2019) [6] proposed a classification method that uses ‘Federal Code of United States of America’ as training data for their model. The results show that combined with TF-IDF feature extraction and Naive Bayes classifier, they're achieved an accuracy of 93 percent in the experimental environment.

Khare et al. (2020) [23] in their work, introduced a smart crawler that can assist experts in searching the DeepWeb 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. More so, the crawler which they presented in their work, is equipped with a text-based site classifier. They used two techniques to classifier the deep web content, by using a neural network with supervised machine learning techniques in which we achieved the best 95.46% of accuracy in classifying sites while the machine learning algorithm Logistic Regression with TF-IDF has got accuracy F1-score of 94%.

1.9 Research Outline

The rest of this research is made up of the following components:

- ✓ **Chapter Two**: contains a comprehensive review of the layers of the internet, the onion routing, crawling the dark web, mining of data from dark web, and some data mining algorithms that are also used in this research.
- ✓ **Chapter three**: entitled "The Proposed System". It covers the proposed system and its algorithms.

- ✓ **Chapter four**: in this chapter, the experiments that were carried out are presented alongside the experimental results, followed by the description of technical details of implementation, as well as how the classifier was successfully implemented in an application.

- ✓ **Chapter five:** is a summary of the entire work that has been carried out in this research, alongside the conclusions of the research contains, and recommendations for future work.

Chapter Two

Theoretical Background

Chapter Two
Theoretical Background

2.1 Overview

In this chapter, the major concepts related to the dark web are explained, describing how the dark web can be accessed for the extraction of data. Firstly, the theoretical background of the internet layers is presented, and subsequently, the main concepts associated with data mining are reviewed. Other concepts that are reviewed include, text preprocessing, and classification. The Chapter mainly focuses on the main methods and techniques adopted in this research.

2.2 Layers of the Internet

The internet is made up of three layers including, the surface web, deep web, and dark web. The dark web is not particularly a primary layer in itself; rather, it is a subset of the deep web. The deep web is a portion of the web which is unreachable through the conventional search engines, and is not indexed by web browsers. Nevertheless, there are certain software that can be used to access that part of the web; such software are capable of gaining access to anonymity networks [1]. This is further explained in Section 2.4 of this chapter. The aforementioned internet layers are schematically described in Figure 2.1 below, and described in the subsequent sections.

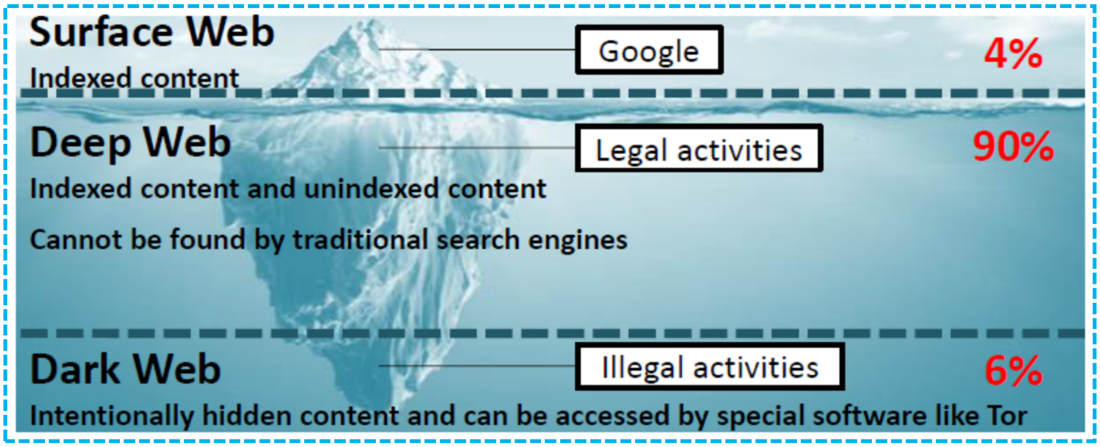

Surface Web
Indexed content
Google
4%
Deep Web
Indexed content and unindexed content
Cannot be found by traditional search engines
Legal activities
90%
Dark Web
Intentionally hidden content and can be accessed by special software like Tor
Illegal activities
6%

Figure (2.1): The Internet layers.

2.2.1 Surface Web

This part of the web is the part that can be indexed by search engines such as Google, and this part of the web is also referred to as Visible Web, Indexable Web, Clearnet or Lightnet [20]. More so, it is also regarded as the portion of the web that is liable and accessible by the public, without any restriction caused by verification or any form of payment. It is public because it is accessible by anyone at any time and not restricted by authentication or payment. This web is also liable because its users can be identified, and hence liable to law enforcement [4].

2.2.2 Deep Web

This part of the web has been in existence from as far back as 1994, and at that time, it was referred to as Hidden Webpages, and subsequently referred to as deep web in 2001 [24]. The Deep web is not indexed by search engines, and is also not linked to pages on the Surface web. This web is sometime referred to as Invisible web or Deepnet. Statistical evidence from previous studies has shown that 96% of the entire WWW is made up of the deep web, whereas the remaining 4% consists of the surface web [23]. The websites that are contained in the deep web are characterized by more complexities due to their content; the content of such websites include confidential information and research data [20]. It is impossible to quantify the capacity of the deep web because it is rapidly changing in terms of accessibility and presentation of information in deep web. This increases the difficulty of measuring this web [16].

There are many reasons as to why indexing of a website could be difficult. The difficulty associated with the indexing of a webpage could be ascribed to many reasons. First, is the fact that passwords are used so that crawlers can be denied access to the website. Additionally, the accessibility of the page is restricted when there is a limit to the number of times the website can be accessed. Lastly, if the entire URL is not common, then the Webpage cannot be reachable, and this is because it is either concealed or not connected to other pages on that site or other websites [16].

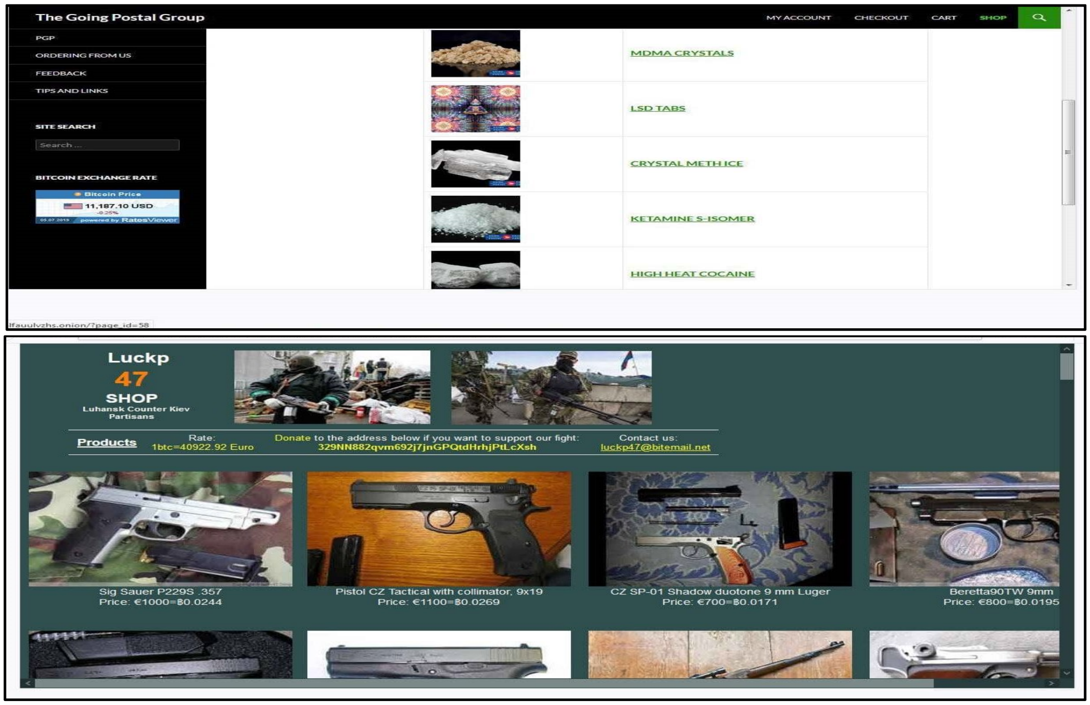
Within the Deep web, there are two kinds of activities that are carried out by users, either legal or illegal activities. The legal activities that are carried out on the deep web include academic activities that involves accessing and storage of databases, libraries containing research materials or even just anonymous surfing of the web when user do not want to be tracked [23]. Due to the fact that the deep web offers users the opportunity to use the web discretely so that their privacy can be maintained. Examples of some users who use the deep web for the sake of privacy include, military forces, security agents, the press, and many other agents [25]. In contrast to the legal activities that are carried out on the deep web, there are illegal activities that are also perpetrated therein, and such kinds of activities are referred to as criminal; the part of the deep web where such crimes are committed is referred to as the dark web, which is explained in details in the subsequent section.

2.2.3 Dark Web

Majority of the illegal activities are carried out on the dark web, which is commonly known as the WWW content. However, due to the fact that the dark web cannot be accessed by browsers that are often used to gain access to the surface web. The advancement of the dark web started at the time the United States military began using it for the purpose of intelligence sharing; this was used to establish communication with intelligence assets that were remotely located without being noticed [26,27].

Some of the illegal transactions performed on the dark web include, child abuse, buying and selling of ammunitions, trafficking of human organs, drug buying and selling, Botnet renting (a network that is well-equipped and connected to the internet and can be operated by hackers, performing various illegal activities), sharing of exploits information that have been hacked from computer systems by hackers. Also, apart from illegal activities involving the trading of documents such as medical information of patients, selling of pseudo identities, stolen banking cards, and other means of personal identification [28]. Furthermore, asides the financially fraudulent activities that criminal carry out in dark web, criminal philosophies and ideologies are publicized, and hitmen are also recruited through the dark web. The illicit activities that are carried out within the dark web are schematically represented in figure 2.2.

Hidden Services of the dark web are rendered in the dark web, and there are special services that promote security breach activities and serve as hosting avenues for malware [19]. The “Silk Road” is one of the well-known e-marketplaces that is particularly dedicated to the sales of drugs and electronic products like passports, hacked multimedia, malware, piracy services, and social card fraud. The founder of this market is Ross William Ulbricht who found it in the 2011 [12].



Donate to the address below if you want to support our fight:
329NN882qvm692j7jnGPQtdHrhjPtLcXsh

Figure (2.2): Samples of drugs and weapons markets in the Tor network.
Source: the Tor Network Retrieved by the author in March 2021.

2.3 Accessing the Dark Web

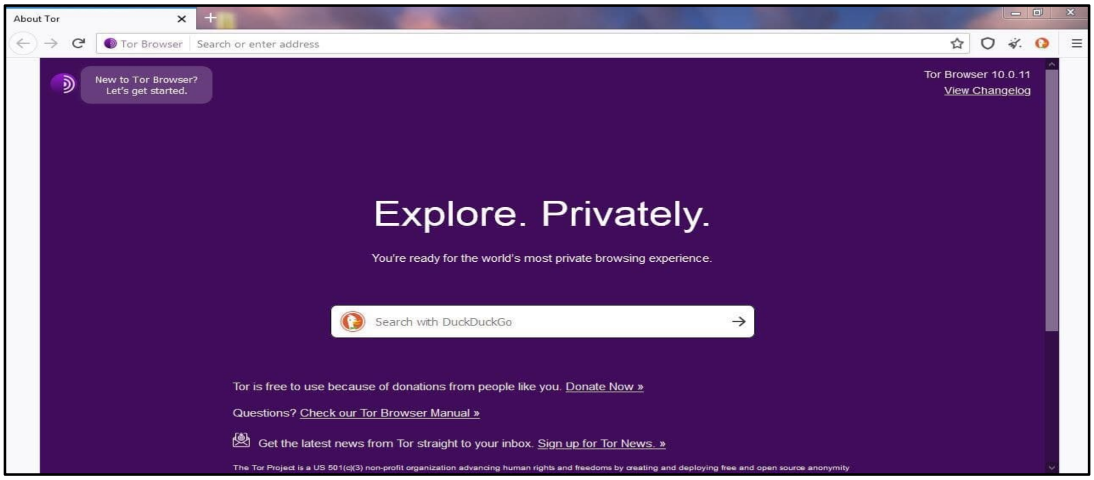
Usually, there is a special browser through which the services and websites in the dark web can be accessed. Nevertheless, there are many other websites within the dark web that are intentionally concealed, meaning that they are yet to indexed by the search engine in the traditional manner. This means that, if one does not have the URL of such websites, it is impossible to access such websites [28].

Accessing the dark web can only be possible through the use of some tools like Freenet, I2P, and TOR which is widely known. More so, entry into certain levels require password authentication and permission [29]. By means of the programs and tools used in accessing the dark web, the sources of data remain private, while the identity of users accessing the target data is preserved. Given this feature, users have shown preference for the dark web in terms of hiding data [30].

In the case of TOR, computers that have been uniquely configured are used in passing requests cross network of connected nodes, thereby providing a certain level of anonymity and privacy. Messages are transferred in encrypted form, such that the only thing that is known to each relay is the computer which the request originated from [23][28].

Since the deep web is made up of a wide range of websites, networks, and databases, certain protocols must be followed before it can be accessed, and such protocols are not easily accessible by every person using the regular web browsers. TOR network is one that has gained so much popularity as a technology of deep web [31]. The name TOR results from its onion-like nature, as it contains several layers [32]. At present, the TOR project is an open-source and non-profit organization with a large society. Given that TOR is an overly network, it employs the use of extant TCP/IP infrastructure [13].

Regardless, of the key factor influencing the establishment of such web, TOR remains the best tool employed in carrying out illicit activities as it provides covering for the clients and operators, thereby making their operations anonymous [16]. Some of these sites can be readily found on the network by accessing some pages that work as references of lists of links to this website like Hidden Wiki, or through the use of special search engines that can be found on TOR networks such as DuckDuckGo, TOR Search, and Grams as contained in Figures 2.3, and 2.4.



Explore. Privately.

You're ready for the world's most private browsing experience.

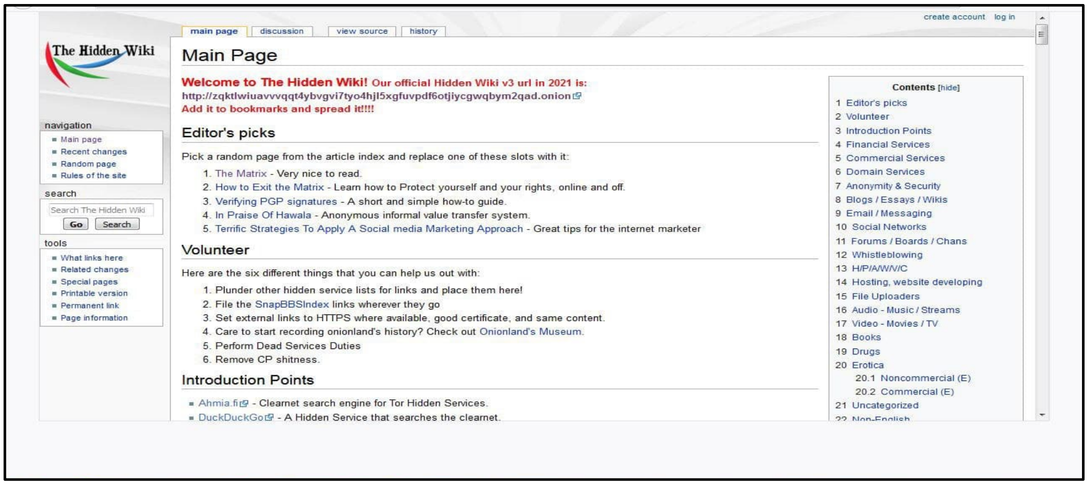
Tor is free to use because of donations from people like you. Donate Now »

Questions? Check our Tor Browser Manual »

Get the latest news from Tor straight to your inbox. Sign up for Tor News. »

The Tor Project is a US 501(c)(3) non-profit organization advancing human rights and freedoms by creating and deploying free and open source anonymity

Figure (2.3): Search engine on Tor networks. Source: the Tor Network Retrieved by the author in March 2021.



Main Page

Welcome to The Hidden Wiki! Our official Hidden Wiki v3 url in 2021 is: http://zqktlwiuavvvqqt4ybvgvi7tyo4hjl5xgfuvpdf6otjiycgwqbym2qad.onion
Add it to bookmarks and spread it!!!!

Editor's picks

Pick a random page from the article index and replace one of these slots with it:

1. The Matrix - Very nice to read.
2. How to Exit the Matrix - Learn how to Protect yourself and your rights, online and off.
3. Verifying PGP signatures - A short and simple how-to guide.
4. In Praise Of Hawala - Anonymous informal value transfer system.
5. Terrific Strategies To Apply A Social media Marketing Approach - Great tips for the internet marketer

Volunteer

Here are the six different things that you can help us out with:

1. Plunder other hidden service lists for links and place them here!
2. File the SnapBBSIndex links wherever they go
3. Set external links to HTTPS where available, good certificate, and same content.
4. Care to start recording onionland's history? Check out Onionland's Museum.
5. Perform Dead Services Duties
6. Remove CP shitness.

Introduction Points

- Ahmia.fi - Clearnet search engine for Tor Hidden Services.
- DuckDuckGo - A Hidden Service that searches the clearnet.

Contents [hide]

1 Editor's picks
2 Volunteer
3 Introduction Points
4 Financial Services
5 Commercial Services
6 Domain Services
7 Anonymity & Security
8 Blogs / Essays / Wikis
9 Email / Messaging
10 Social Networks
11 Forums / Boards / Chans
12 Whistleblowing
13 H/P/A/W/V/C
14 Hosting, website developing
15 File Uploaders
16 Audio - Music / Streams
17 Video - Movies / TV
18 Books
19 Drugs
20 Erotica
20.1 Noncommercial (E)
20.2 Commercial (E)
21 Uncategorized
22 Non-English

Figure (2.4): The Hidden Wiki site. Source: the Tor Network Retrieved by the author in March 2021.

If a user of TOR, who is also referred to as a source joins the network through the TOR browser, he/she will be connected to a virtual circuit of TOR nodes that have been selected randomly. Normally, three computers running the TOR Browser will be selected. The virtual circuit is usually replaced with a new one after about ten minutes [33]. The following three types of nodes are the components of the virtual circuit:

- **Entry Node**: arriving traffic is received by the entry node.
- **Intermediate Node**: this node is responsible for transferring information between nodes.
- **Exit Node**: this is the last node, which is responsible for conveying traffic to the open Web, being the final destination of the traffic.

Requests made by hundreds of users seeking anonymity when using the web are put forward by the exit relays on behalf of the users. The exit relay to be used for the execution of this task is determined through the randomization of algorithms. Globally, there are approximately seven thousand computers that serve as relays. Consequently, anonymity is granted to users, as every user is embedded within the numerous layers of the onion [13]. The elements that make up the TOR Network are demonstrated in Figure 2.5 below.

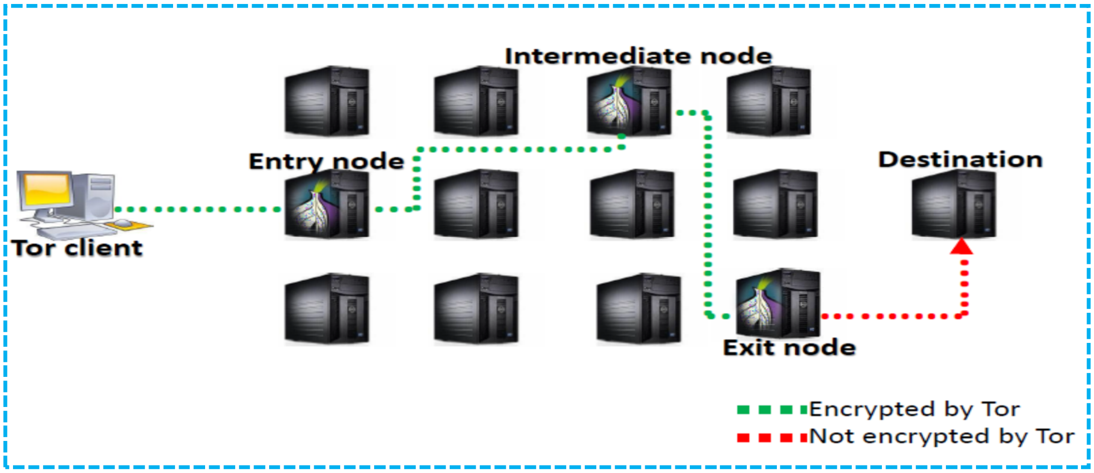


Figure (2.5): Components of TOR network.

2.4 Crawling the Dark Web

Websites can be collected automatically by web crawlers once dark web markets and forums have been determined, and subsequently, a major seed site is detected using a custom web crawler [34]. The operation of the web crawler involves the collection of data from the internet and its storage in a database for further sorting and analysis. In this operation, pages are collected from the net, and then downloaded automatically while following the existing hyperlinks. Fresh webpages are continuously captured [35]. After an illegal deal data has been well-downloaded, it is then processed and categorized for longer-term storage [34]. Figure 2.6 illustrates how a crawler works.

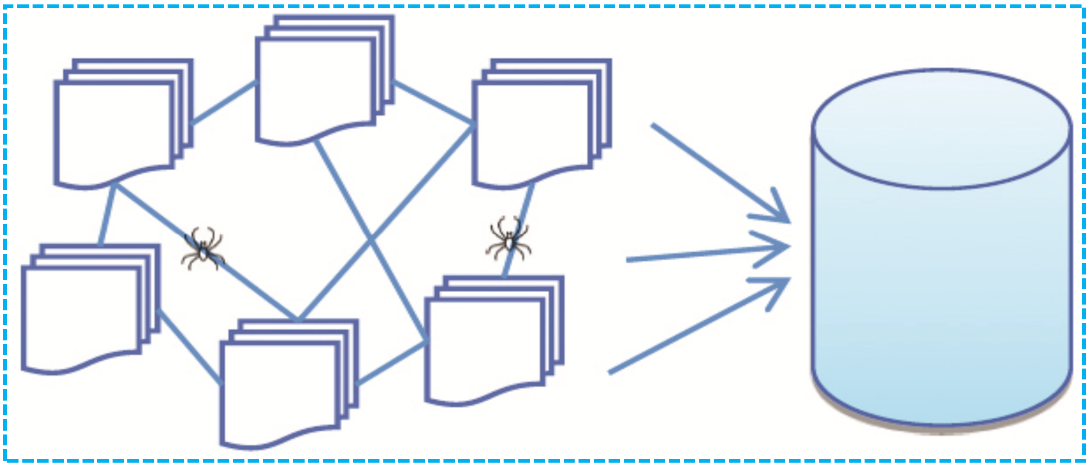


Figure (2.6): Data extraction and storage operation by a crawler.

A web crawler, which is also referred to as a web spider, is an internet bot that crawls through a HTML website for the purpose of collecting information about a site. The kind of information gathered by the crawler include, contents of a Webpage, metatags, websites' URL, page titles, and most importantly, links that are contained in the page. The crawler uses the link it has collected from the first page to visit and store the same data of the next pages. A source is sent in form of robots.txt by the web crawler, and the source in turn ensures that all the information is delivered to the server [20][23].

In the last twenty years, crawling program development have seen increasing concern in dark web, but with the multiple challenges involved (which have previously mentioned in chapter one), developing such programs requires additional techniques so that the crawlers would be capable of discovering malevolent websites, accessing them, and storing their data for future processing [16].

It has been mentioned earlier that the deep web is significantly large in size, and it is particularly characterized by data that is of superior quality and great relevance within a wide range of semantic domains. Consequently, the area of research that focuses on the design of Deep web crawlers that are capable of automatically accessing such data is of great interest [3].

The use of crawlers is employed in a wide range of applications as given below:

- In search engines, crawlers create a copy for future processing of all pages they visit, i.e., search engines index web pages to retrieve them when a user searches for certain subjects easily [23].
- The crawler provides constant security for a website by cross-checking hyperlinks and authenticating HTML tags [31].
- It also collects specific kinds of data like email addresses, especially spam or malicious emails [16].
- In recent times, discussions from dark web forums have been collected through targeted crawling [31].

2.5 Web Mining

Web mining is a process whereby information from the web data is obtained using data mining approaches. Such web data include hyperlinks, web records, web use logs, etc. WWW is a collection documents consisting of images, texts, video and audio data. Due to the existence of this kind of huge amount of unclassified data, the interaction between users and the data becomes quiet challenging, thereby creating difficulties for the users in terms of obtaining the right data they are searching for [36].

There are three broad categories of web mining research, including web structure mining, web content mining, and web usage mining [37]. This research used web content mining, specifically, textual content. Figure 2.7 shows web data mining classification.

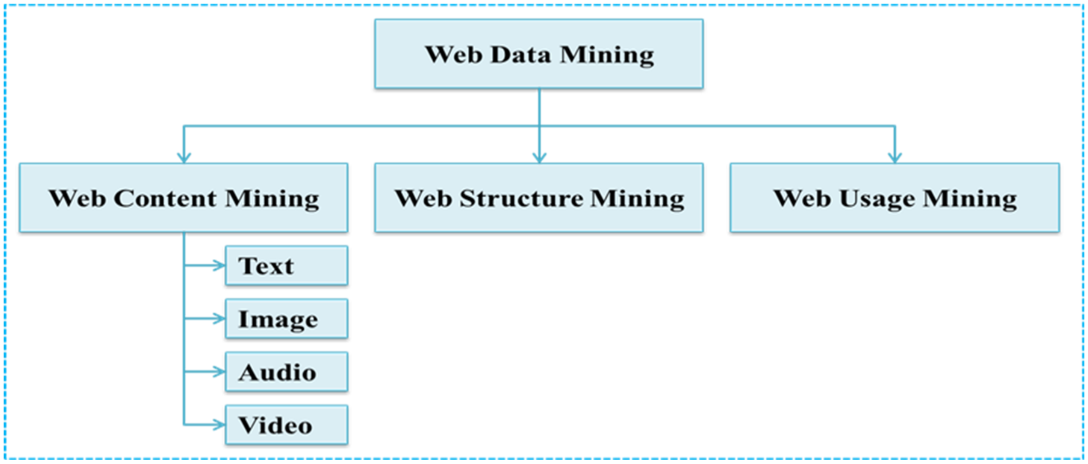


Figure (2.7): Classification of the web data mining [37].

2.5.1 Web Content Mining

Web Content Mining is the process which is concerned with the use of data mining approaches to extract contents that are available on the internet, normally XML, HTML, or plaintext [38]. In this process, relevant data is discovered from the web; contents like videos, texts, audios, and images, etc. [37].

2.5.2 Web Structure Mining

This kind of mining is normally used to obtain the hyperlink organization of a website, as well as for the extraction of structural information that aids users in getting better search results. Through this kind of mining the correlations between the user and the web can be identified, while the link of structure of hyperlinks in the documents can be ascertained [36][39].

2.5.3 Web Usage mining

This type of web mining involves the use of data mining methods to scrutinize activity or search logs with the aim of discovering relevant patterns of interest [37]. Also, it is a method used in discovering what users are searching for on the internet [36]. When a user surfs a website, a log of the user's activity is created, and can be referred to much later [39].

2.6 Text Pre-processing

Data preprocessing methods are important to prepare the dataset. In general, all of these methods fall into two categories: selecting data objects and attributes for the analysis or creating/changing the attributes [40]. These methods include several strategies for handling dataset issues such as noise, missing values, and inconsistent data.

Text pre-processing is the method that converts texts into an appropriate form to be classified [41]. Text pre-processing is very important because it can be used to decrease the feature space and computational process, and this, in turn, can positively affect the classification accuracy [42].

Texts are normally prepared for classification using numerous pre-processing steps. This is a process that involves the removal of irrelevant data like, numbers, symbols, special characters, and punctuations. Apart from all the aforementioned, any other irrelevant information that hampers on the accuracy and efficiency of classification algorithm should also be eliminated from the data during the pre-processing stage [43].

In this section, a number of the steps involved in preprocessing are discussed. One of the steps is word tokenization, and it is a process whereby sentences are segmented into their component words [44]. The aim of this process is to reconstruct words sentence. The step of tokenization is crucial in Natural Language Processing (NLP), because it is critical to several steps of processing, given that it helps in splitting the text document into tokens and grouping according to periods, semicolons, space, and quotes between words [45]. The tokens that are split could be numbers, words, or symbols. Typically, a wide range of interfaces can be accessed from the python library for the purpose of tokenization. The output obtained from the process of tokenization, is then use as input in the subsequent stage.

Another step involved in the pre-processing stage is the elimination of irrelevant words like Stop Words, and missing values.

Stop words are referred to as words that are of no relevance, and contain no useful information in the text. Stop words include, conjunctions, pronouns, and prepositions [46]. Normally, they are eliminated from the text in the course of processing so that the size of the text that is to be processed by the classification algorithm can be minimized; the remaining words that are retained in the process will be words with great context and significance [47]. Often times, stop words are words with frequent occurrence in the vocabulary of a language, and every language has its particular stop words. Hence, there is no detailed list of stop words.

If the problem of missing values is not well addressed the performance of the classifiers models could be greatly affected. The problem of missing values is usually addressed using the following methods [39]:

- Elimination of data objects
- Neglecting missing values during analysis.
- Estimation of missing values.

In an event that the missing values contained in the attribute are small and extensively scattered, then the use of estimation method can be employed. More so, the use of residual values can be employed in the evaluation of missing values. If the predictor is categorical, the predictor value with the highest occurring frequency can be taken. Conversely, if the predictor value is continuous, then the average predictor value of the nearest neighbor can be used [48].

2.7 Data Mining

Data Mining (DM) is the process through which relevant knowledge or information is extracted from a large collection of data [49]. This process can be carried out using a wide variety of techniques that are specially designed for this purpose.

In general, there are two kinds of tasks that are involved in the process of data mining: prediction and description [50]. Prediction involves the use of supervised learning techniques for the prediction of the value of a given attribute depending on values of other known attributes. There are two categories of tasks that are performed in predictive modeling, including regression and classification. On the other hand, the tasks that are associated with description include, clustering, mining associations, sequence discovery, and summarization. The aforementioned approaches depend on unsupervised learning techniques for the identification of obvious patterns in data [51,52]. In this research, the use of classification techniques is employed in the discovery of categories of ambiguous data. The classification techniques used in this research include, Support Vector Machines, Random Forest, and Naïve Bayes.

2.8 Text Mining (TM)

Text Mining (TM) is a multidisciplinary field that seeks for extracting significant information from unstructured data. TM is based on data mining, retrieval of information, computational linguistic, machine learning, and statistics [53]. The presence of a huge amount of information like books, Webpages, blogs, digital libraries, and news articles, is one of the reason why research in the area of text mining is still vibrant. Thus, one of the major aims of text mining is to extract high-quality information from text [54].

There are similarities between text mining and data mining, except that text mining is used alongside datasets that are either semi-structured or unstructured, such as texts, HTML files, and emails. On the other hand, data mining tools are employed in manipulating the structured data from the databases [53].

The operation of text mining is performed with natural language text which is often in a format that is either unstructured or semi-structured [54]. There are numerous methods such as clustering, classification, and summarization that can be used for the extraction of knowledge.

2.9 Machine Learning Algorithms

One of the applications of artificial intelligence is machine learning which provides the systems with the capability to automatically learn and get better from experience. Machine learning focuses on developing computer applications that can access data and use it to learn themselves.

This research employed the use of three machine learning algorithms, which are Linear Support Vector Machine (LSVM), Fandom Forest (RF), and Naïve Bayes algorithm (NB).

2.9.1 Linear Support Vector Machine (LSVM)

Support Vector Machine (SVM) tries to find a hyperplane or set of hyperplanes in an N-dimensional space, whereas *N* refers to the number of features, to divide the data points into distinct classes [5]. The best hyperplane is the one that can maximize the distance between data points of different classes.

The performance of SVM has been good in terms of data points that can be separated linearly, for non-linearity; kernel functions like Gaussian, Sigmoid, and polynomial are required. Essentially, such kernels are used in mapping the non-linear separable data points into a feature space of higher dimension [11].

More so, the use of LSVM can be employed in the case of a linearly separable of a two-class learning task, because it helps in finding a hyperplane that is capable of separating two classes of a given sample with a maximum margin. Due to this capability possessed by SVM, it is sometimes referred to as maximal margin classifier. With this margin, the optimal generalization ability can be provided [55]. Generalization is described as ability of the classifier to perform efficiently in terms of high prediction and accuracy rates. Figure 2.8 depicts the optimum LSVM hyperplane for a linear case.

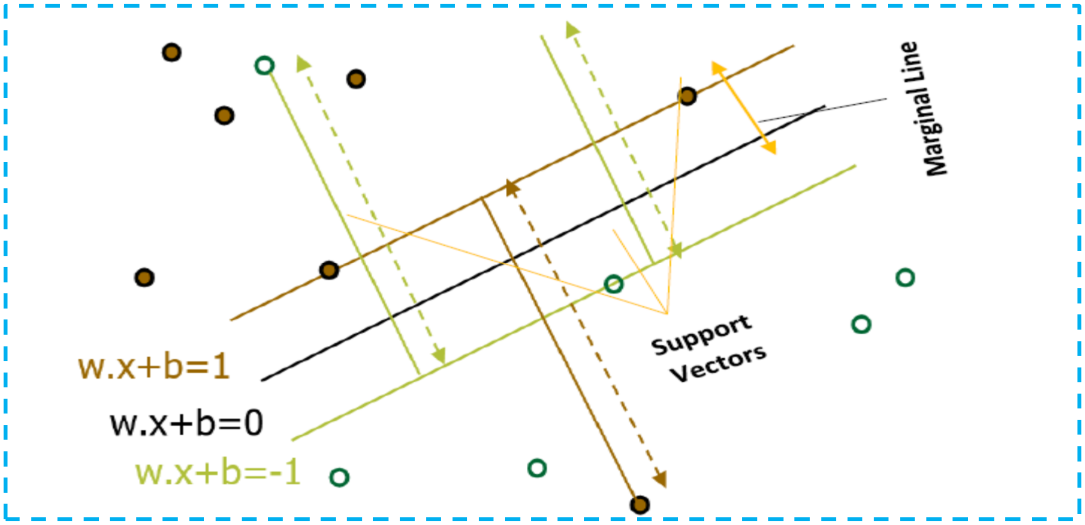


Figure (2.8): The LSVM algorithm concept.

Originally, linear SVC was specially designed to carry out the task of binary classification, because the problem of binary classification is made up of N training instances. Each instance is indicated by a tuple (x_i, y_i), where {x_i, ...,} are a dataset and y_i∈ {1, -1} represents its class label [56]. There are many mathematical equations, according to Figure 2.8 above:

The following Equation (2.1) can be used in representing the decision boundary of a linear classifier [57].

$$w^T .x+b=0 \tag{2.1}$$

Where **w** represents the weight vector, and **b** refers to the bias in the optimal hyperplane. Equations (2.2 and 2.3) denote decision boundaries [56].

$$w^T .x_i + b= 1 \qquad (2.2)$$

$$w^T .x_i+ b= - 1 \qquad (2.3)$$

To learn the SVM model, the selection of parameters w and b must done based on the two conditions in Equations (2.4 and 2.5) [56].

$$w^T .x_i+b \geq 1 \text{ for } y_i= +1 \qquad (2.4)$$

$$w^T .x_i+b < 1 \text{ for } y_i= -1 \qquad (2.5)$$

The good performance of SVM algorithm in terms of classifying text, has been highlighted in the study of Joachims [58]. The author noted that this ability of the SVM can be attributed to the features possessed by this algorithm as compared to other algorithms: High-dimensional input space, few irrelevant features, document vectors are sparse, and most text categorization problems are linearly separable. These arguments give theoretical evidence that SVMs should perform well for text classification.

2.9.2 Random Forest Classifier (RF)

Random forest (RF) is a regression and classification technique that fits a large number of decision tree classifiers on various sub-samples of a dataset. RF is built on the foundation of the decision tree [59]. A large number of trees are created from a training set and validated to predict future observations. This method can have both a categorical and a continuous value output. There are many benefits of using the Random Forest algorithm such as: It can be used for both classification and regression task, it can handle missing values and maintain accuracy for missing data, and it has the ability to deal with a large dataset of a higher dimensionality [60,61]. Figure 2.9 illustrates the main notion behind this algorithm.

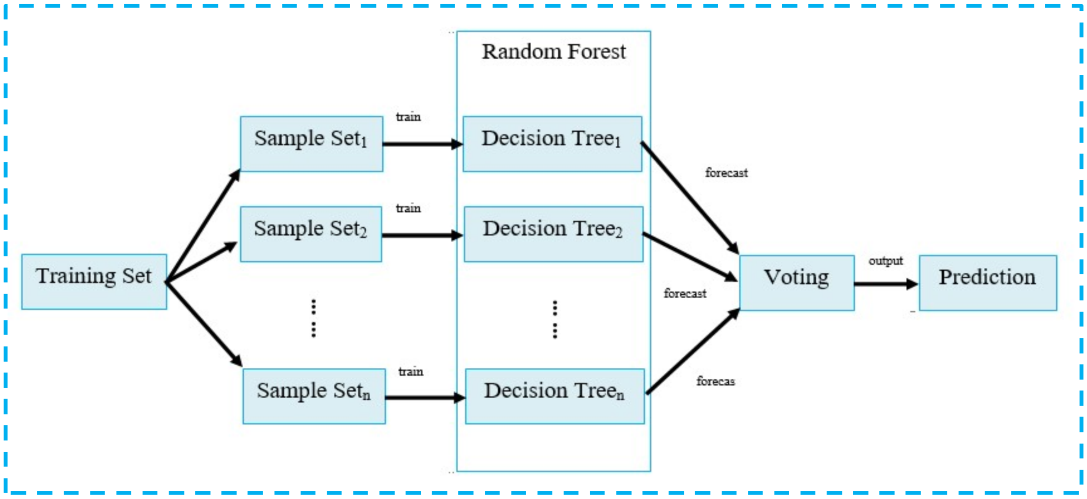


Figure (2.9): The basic flow chart of the Random Forest algorithm [59].

2.9.3 Naïve Bayes Classifier (NB)

In simple terms, Naïve-Bayes classifier can be referred to as a simple classifier that is based on the Bayesian Theorem of conditional probability, and strong independence assumptions [62]. The classifier can be used to determine if document A is part of class B or not [63]. In addition, its functionality is based on the independent feature model, and its working is also based on the hypothesis that the existence or non-existence of a particular attribute has a relationship with the occurrence or non-occurrence of a particular feature.

Using the Bayesian classifier is advantageous in the sense that it only requires a small dataset for training, it is not sensitive to insignificant features, easy implementation, fast and efficient classification [64].

As a probabilistic classifier, the Naïve Bayes functions based on Naïve Bayes rule, and the posterior probability of a document “d” being in class “c” is given as Equations (2.6 and 2.7) are met [63].

$$P(c|d) = \frac{P(d|c)P(c)}{p(d)} \qquad (2.6)$$

$$P(c|d) = \frac{P(w1.w2.\cdots.wn\ |c)p(c)}{p(d)} \qquad (2.7)$$

Where P(d|c) indicates the likelihood of predicting given class. Where $P(w_i \mid c)$ is the conditional probability of finding word w_i in a document d of class c. $P(w_i \mid c)$ is a measure of how much w_i demonstrates that c is the proper class. $(w_1, w_2, \ldots, w_n)$ are tokens in document d that are part of the vocabulary used for classification, and n is the number of such tokens in document d.

The parameter P(C) is prior probability of class estimated as, Equation (2.8) is met [62].

$$P(C) = \frac{NCi}{N} \tag{2.8}$$

Where NC_i represents the number of documents in class C_i and N represents the total number of documents in the set of training.

For all classes, P(d) is the prior probability of predictor (d).

In documents classification, the goal is to find the best class for the document. The Naive Bayes classifier predicts the class with the maximum posterior probability as, Equation (2.9) is met [65].

$$P(c|d) = \underset{c \in C}{\operatorname{argmax}} \frac{P(d|c)P(c)}{p(d)} \tag{2.9}$$

2.10 Performance Evaluation of Classification Algorithms

For the generalization power of the trained model to be evaluated, the use of the performance metrics is employed. It is also used in assessing the quality of the trained model when the data is unseen. There are several performance metrics that can be employed for the evaluation of the efficiency of a given classification model. The performance metrics include, but not limited to, accuracy, precision, recall, and f1-score.

Despite the fact that the performance of the models can be evaluated using a variety of metrics, accuracy remains one of the commonly used methods of evaluating generalization power of algorithms [66]. When using accuracy, the evaluation of the trained model is done based on the total instances that are correctly predicted by the trained model when it is tested with the unseen data. The use of precision, recall, or f1- score metrics is more suitable when dealing with problems of unequal classes, so that the performance in terms of accuracy can be optimized [66]. These measures are calculated in accordance with the computation of confusion matrix. A summary of the number of instances that have been correctly or wrongly predicted by a classification algorithm is provided through this matrix (see Table 2.1) [67]:

Table 2.1: Two dimensional confusion matrix.

Predicted **Actual**	Positive	Negative
Positive	TP	FN
Negative	FP	TN

1. True Positive (TP): The positive examples that have been classified correctly.
2. False Negative (FN): The positive examples that have been classified wrongly.
3. False Positive (FP): The negative examples that have been classified wrongly.
4. True Negative (TN): The negative examples that have been correctly classified.

Below is a brief description of the accuracy, precision, recall, and f1-score, metrics [67]:

1. Accuracy refers to the number of predictions that have been made correctly, and is divided by the total number of predictions. The accuracy can be computed based on Equation (2.10).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Number of Predictions}} \qquad (2.10)$$

2. Precision is described as the number of TP divided by the number of TP and FP. The precision can be computed based on Equation (2.11).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (2.11)$$

3. A recall is the number of TP divided by the number of TP and the number of FN. This metric can be computed based on Equation (2.12).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (2.12)$$

4. F1-score is the 2*((precision*recall)/ (precision + recall)). It is also called the f1- score or the f1- measure. An equation of this metric can be computed based on Equation (2.13).

$$\text{F1} - \text{score} = \frac{2*\text{TP}}{2*\text{TP}+ \text{FN} + \text{FP}} \qquad (2.13)$$

Chapter Three

The Proposed System

Chapter Three
The Proposed System

3.1 Overview

This chapter provides a description of the practical aspect of this research, starting with the presentation of the proposed systems' architecture, followed by the discussion on the crawling and data collection procedure. Also, the process of algorithms' pre-processing is explained in this chapter, while the automatic labeling algorithm which was used to label the dataset used in this research. Finally, the adopted classifiers and their functions in the proposed system are presented, and a discussion on the evaluation approaches used, is also presented.

3.2 The Proposed System Architecture

The architecture of the proposed system is made up of four stages, with each of the stages having sub-steps. With the sub-steps, the main objective of the research as well as the specific objectives are achieved. The four stages include, crawling and collection of data, data pre-processing, dataset auto-labeling, classification, and evaluation. Figure 3.1 shows the four stages.

The first stage, which is primarily focused on crawling and collection of data, consists of several sub-steps. This stage was aimed at acquiring data used on the proposed system. In the second stage, the data obtained in the first stage is pre-processed, and there are several processes involved in this stage. The aim of this stage is to ensure that the input, i.e., the data is well prepared for use on the system. The third stage involves applying the automatic labelling algorithm proposed in this research to the dataset, so that the dataset can be labeled automatically. Lastly, the fourth stage is basically focused on using LSVM, RF, and NB classifier. Subsequently, the crawling and mining operations are implemented within the dark web, so as to enable the identification of the most illicit activities within the period of data collection.

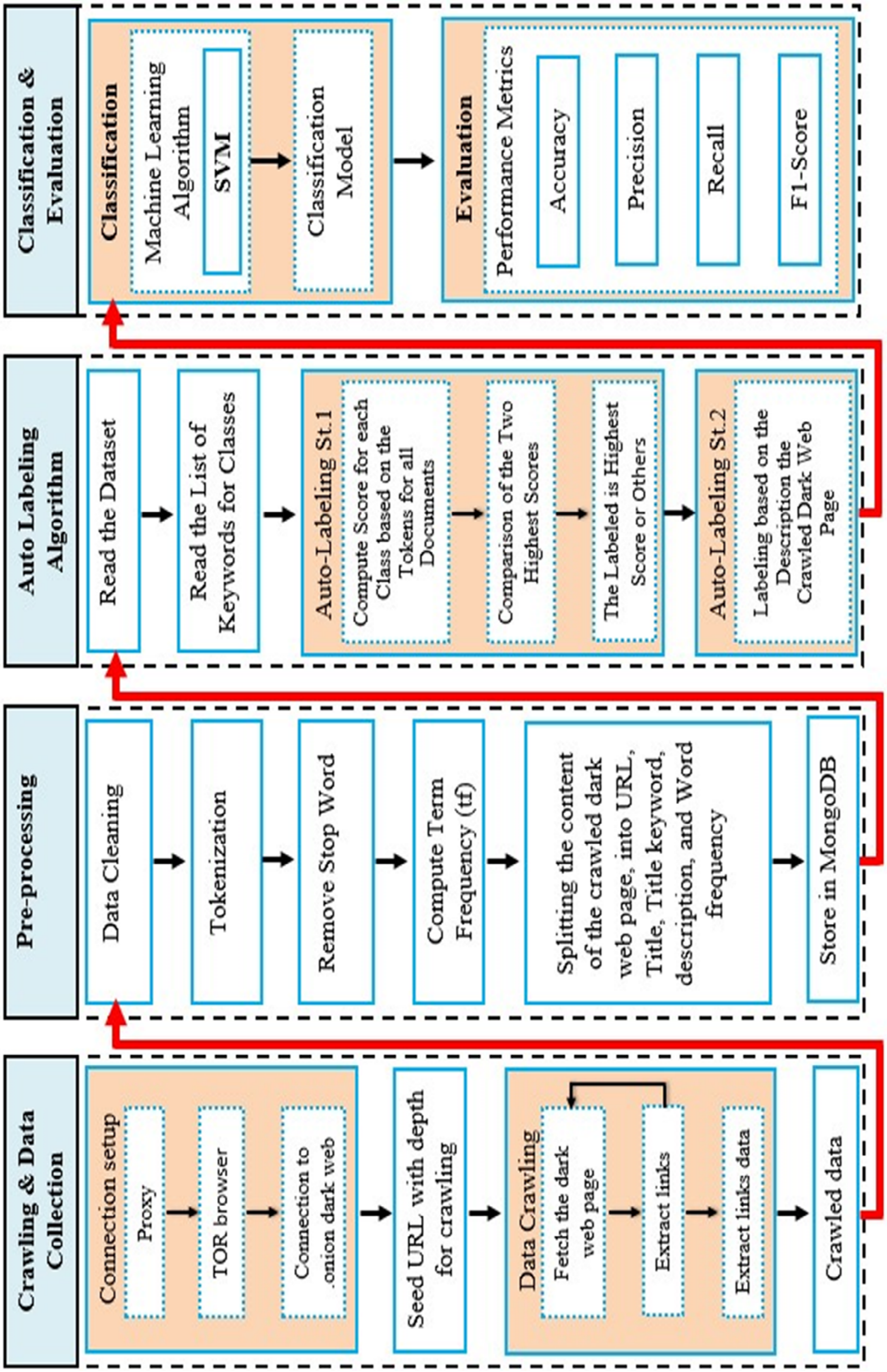

Crawling & Data Collection
Connection setup
Proxy
TOR browser
Connection to .onion dark web
Seed URL with depth for crawling
Data Crawling
Fetch the dark web page
Extract links
Extract links data
Crawled data
Pre-processing
Data Cleaning
Tokenization
Remove Stop Word
Compute Term Frequency (tf)
Splitting the content of the crawled dark web page, into URL, Title, Title keyword, description, and Word frequency
Store in MongoDB
Auto Labeling Algorithm
Read the Dataset
Read the List of Keywords for Classes
Auto-Labeling St.1
Compute Score for each Class based on the Tokens for all Documents
Comparison of the Two Highest Scores
The Labeled is Highest Score or Others
Auto-Labeling St.2
Labeling based on the Description the Crawled Dark Web Page
Classification & Evaluation
Classification
Machine Learning Algorithm
SVM
Classification Model
Evaluation
Performance Metrics
Accuracy
Precision
Recall
F1-Score

Figure (3.1): The proposed system.

3.3 Crawling and Data Collection

Several dark web links have been collected from several sites, and out of all the links obtained, ahmia is the most commonly used one. After the collection of the links, the web crawler is fed with the links as input, and the web crawler in turn obtains every other link that is related to the ones obtained initially. At the end of the data collection, over 2,300 links alongside their contents were acquired, processed and stored in the database used in this research.

Web crawlers have gained popularity in several projects involving the collection of data from the internet. As the crawler tracks the hyperlinks, it downloads any webpage it comes across automatically, and during the process, it keeps discovering new webpages. Any data that is downloaded successfully can be processed and stored for future use. This means that, the data can be stored for a long period of time. The crawler data flow chart is shown in Figure 3.2.

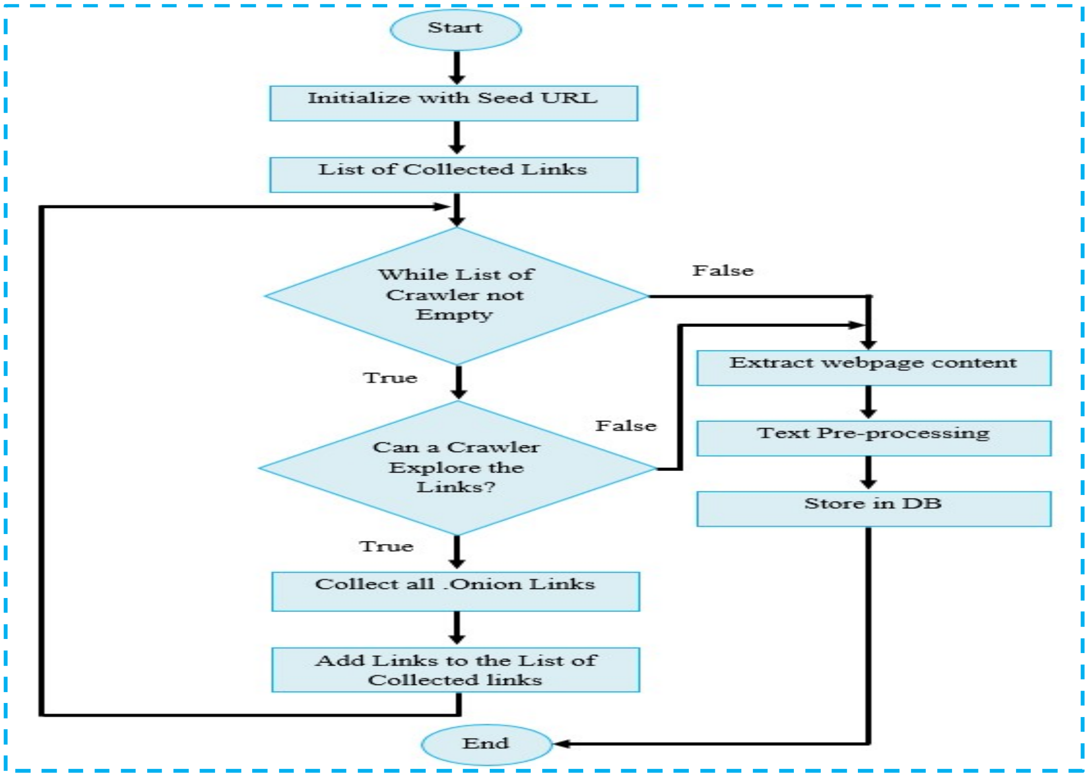


Figure (3.2): Crawler data flow chart.

The Algorithm 3.1 shows the key steps involved in the process of collecting data from the dark web by the proposed crawler system.

Algorithm 3.1: Crawling into the Dark Web

```
Input: Dark web seeding site
Output: Set dark web pages that crawled (S)
Begin
Step1: enter seed site to crawler
Step2: collect the hyperlinks and storing in the list (L)
Step3: // explore links and extract webpages content
        i=1
        N= number of links in list
        While list of links not empty and i<=N do
        Begin
            If crawler explores the links then
              add links to list of links
              N= N+links
              step3.1: extract webpage content
              step3.2: text pre-processing
              step3.3: store in DB
            i=i+1
            End If
        End while
End
```

3.3.1 Dark Web Crawler

The crawler designed in this research has demonstrated the ability to access TOR and the public internet, simultaneously. More so, it carries out automatic search of websites in TOR network on the basis of pre-defined links, so that the content of dark web pages can be extracted and archived in a database. In this work, the TOR network was infiltrated using local HTTP proxy software, which is referred to as SOCK5h. This software is an Internet protocol used by TOR. When this software is used, the traffic is sent via the TOR network, rather than from the IP address into the open network.

In this research, TOR was employed as a proxy so that the crawler can be enabled to connect to the internet via the TOR network. While the TOR was connected and running, the SOCKS5h software was configured in the crawler, and in the SOCKS5h settings, 127.0.0.1 port 9150 was typed. TOR listens for SOCKS requests on the special loopback address 127.0.0.1 which other devices within the network cannot access. It is very essential for HTTP SOCKS to be used so that the connection can remain encrypted throughout.

At the time the crawling process begins, the proposed system explores the site to determine if there are hyperlinks leading to other sites. In an event, that such hyperlinks are found, those links are acquired by the crawler to be crawled on sequentially basis so that data can be collected, processed and stored. This process is carried out by the crawler continuously, until all links have been explored. This is beneficial to the proposed system, where the function was called for an unlimited number of times till it was completed; this was achieved using a recursive algorithm.

The main components of the dark web crawling system are summarized as follows:

1. Crawling Space: the starting point of the crawler is a list of websites where illicit activities are carried out. In this work, different sources were explored including, security sources that are accessible, non-governmental organizations, and government sources, as well as electronic sources such as indexes of TOR network. The acquisition of links to dark web sites can be achieved by exploring just the search engines like Google, or the surface

web. The addition of new links led to expansion of the crawling space, and this links were found by the crawler on the retrieved webpages. The proposed system has been programmed to carry out the extraction of website data within a time frame of 10 seconds, and afterwards, the data is forwarded to the succeeding link.

2. Storage: this involves the storage of the webpages that have been harvested by the crawler. For the crawler to be able to save the harvested data, it must have access to a storage space where the data can be stored permanently, and there are two techniques that can be used in achieving that, depending on the capabilities possessed by the used equipment. Connecting a database, or using a simple file storage system. Thus, in this work, the use of MongoDB was employed for data storage.

3.3.2 Proposed Crawling System

The dark web crawler presented in this research was designed in a python programming language, making it unique and different from what has been done previously by other researchers. Its uniqueness also lies in the fact that the crawler is connected to the dark web sites on TOR network through TOR software. The TOR network is used together with proxy so that optimal security and anonymity of the crawler can be guaranteed; to achieve this, the IP address is relocated. Upon the establishment of the connection between the TOR network and proxy, the operation of the crawler is launched, beginning with the first seed URL so as to retrieve contents of the page, explore hyperlinks, and crawl through the hyperlinks. This process is implemented repeatedly until the crawler finishes the search and completely, and afterwards moves to the next hyperlink.

The process of crawling the dark web involves reading webpages, extraction of hyperlinks alongside their contents. The reason for the use of Python programming language in the crawler is that, it provides a wide variety of libraries that can aid the process of crawling and hyperlinks extraction from webpages.

3.4 Pre-processing

The dataset used in this research was preprocessed using a range of preprocessing steps. The purpose of preprocessing is to ensure that the dataset is prepared in a manner that is very suitable for machine learning techniques, like prediction models. There are three steps that are involved in the preprocessing of the dataset, and they are briefly described as follows:

3.4.1 Data Cleaning

The process of data cleaning involves unifying various forms of the same letter through the conversion of all characters into upper case or lower case, as well as all numbers, symbols, and all Non-English words are eliminated. The steps involved in the cleaning process are described as follows:

- **Tag Removal**

In the first step, tags are eliminated from content obtained from the dark web page through the process of crawling. The subsequent steps.

- **Removing Numbers**

The second step in data cleaning is focused on the removal of numbers from the content of the dark web page.

- **Removal of Punctuations**

It is very important to remove the punctuation marks as they do not provide any relevant information. This step focuses on the removal of punctuations, and conversion of text into lowercase, and removal of double space from page texts. Examples of punctuations are:['!', '"', '#', '&', "'", '(', ')', '.', '/', ':', ';', '<', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~'].

- **Removing Special Characters**

Examples of special characters include [♈® € £ ƒ $ ° ª‽ ✔ ✖ † ⁒].

3.4.2 Word Tokenization

The process of tokenization is one of the most crucial process required in natural language processing. In the process of tokenization, texts from dark web page is divided into tokens, using the space to isolate each word from another. The tokens come in different forms including numbers, symbols, and words.

3.4.3 Remove stop words

Stop words are described as those words with frequent occurrence. They can also be referred to as words that have no relevance in the process of classification, or any word without a clear significance. Stop words is greater than 400 words, hence deleting stop words leads to dimensionality reduction. Table (3.1) contains some of the stop words used in the proposed system.

Table (3.1): Sample of the English Stop Words.

No.	Stop word	No.	Stop word	No.	Stop word
1	up	11	no	21	both
2	wouldn't	12	mustn't	22	any
3	can	13	he	23	as
4	too	14	this	24	or
5	below	15	weren't	25	further
6	down	16	same	26	which
7	should	17	Each	27	such
8	and	18	re	28	haven
9	not	19	needn't	29	my
10	myself	20	Wouldn’t	30	into

In the last step of the preprocessing stage, the content of dark web page that has been crawled through is split into URL, title, title keywords as a list of tokens, description, and word frequency. Figure 3.3 shows the block diagram of the preprocessing steps.

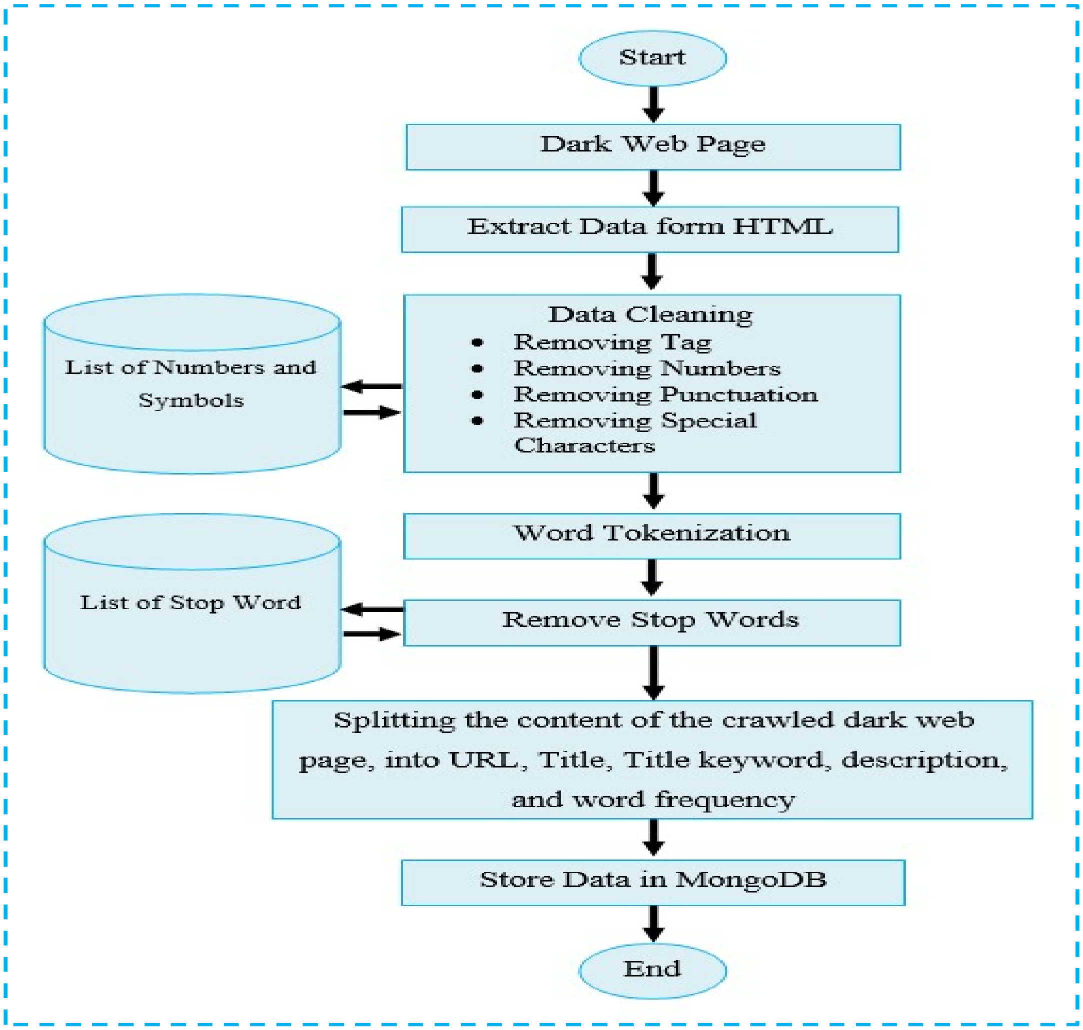


Figure (3.3): Block diagram of the Pre-processing steps.

Algorithm 3.2 depicts the steps used in the pre-processing of the dark web pages that crawled.

Algorithm 3.2: Pre-processing of dark web pages contents

Input: set of dark web pages that crawled (S), numbers list (NL), punctuation list (PL), special characters list (SL), and stop word list (SWL)

Output: set of dark web pages splitting into URL, title, description, title keywords, word frequency

Begin

```
Step1: read the dark web page
Step2: remove tag
Step3: // remove numbers
    i=1
    N= number of all tokens
    While i<= N do
    Begin
        If token [i] is found in NL then
            remove token [i]
        i=i+1
    End
Step4: // remove punctuation
    i=1
    N= number of all tokens
    While i<= N do
    Begin
        If token [i] is found in PL then
            remove token [i]
        i=i+1
    End
Step5: // remove special characters
    i=1
    N= number of all tokens
    While i<= N do
    Begin
        If token [i] is found in SL then
            remove token [i]
        i=i+1
    End
Step6: convert all remaining tokens into lower case
Step7: separate each word from other depending on space to obtain tokens
Step8: // remove stop words
    i=1
    N=number of all tokens
    While i<= N do
    Begin
```

```
            If token [i] is found in SWL then
                remove token [i]
            i=i+1
        End
Step9: splitting the page content into URL, title, title keywords as a list
       of tokens, description, and word frequency.
End
```

3.5 MongoDB

Upon the completion of each crawl and pre-processing stage, the data is stored in MongoDB, which is an open source database system that is publicly available. The name MongoDB has its roots in the word "hu**mongo**us", and it was developed and by 10gen. It is part of the NoSQL family of database systems. Unlike in linked bases where data is stored in tabular form, MongoDB organizes data as JSON-like documents with dynamic schemas (MongoDB called BSON format). This way, data can be easily and rapidly integrated into certain kinds of applications. Table 3.2 shows a sample of the dataset created in this work; it is called "crawler-db".

Table (3.2): Sample the crawler-db.

Crawler-db		**Details**
Dark web site 1	URL	http://luckp47s6xhz26rn.onion/
	Title	Luckp 47 SHOP
	Title keywords	Object (luckup: 1 , shop: 1)
	description	The biggest selection from guns in the Europe.
	Links	Array(hyperlinks in this URL)
	Word frequency	Object (glock: 13, beretta: 2, weapon: 4, sigsauer: 3, luhansk: , luckp : 1, address: 2, contact: 1, tactical: 4, ...)
	URL	http://weapon5cd6o72mny.onion/
	Title	Black Market - Guns Arms Ammo Drugs for Bitcoin - Supplier since 2001

Dark web site 2	Title keywords	Object (black: 1, market: 1, gun: 1, arm: 1, ammo:1, drug: 1, bitcoin: 1, supplier: 1)
	description	Choose right now what do you need, we are shipping worldwide and guarantee the successful shipping of all our physical good.
	Links	Array(hyperlinks in this URL)
	Word frequency	Object (gun: 1, arm: 1, ammo: 1, guarantee: 1, weapon: 1, parcel: 1, ...)
Dark web site 3	URL	http://dwayoz4ucb4nszmi.onion/
	Title	Drugs and Pills SA - DarkWay
	Title keywords	Object (drug: 1, pill: 1, darkway: 1)
	description	-
	Links	Array(hyperlinks in this URL)
	Word frequency	Object (drug: 3, pill: 4, cocaine: 2, weed: 1, weapon: 1, fake: 1, shop: 1, sale: 1, codein: 2, colombia: 1, ...)
Dark web site 4	URL	http://fzqnrlcvhkgbdwx5.onion/
	Title	CannabisUK - UK Wholesale Cannabis Supplier - order weed online in the UK for bitcoin - marijuana for bitcoin - cannabis for bitcoin
	Title keywords	Object (cannabisuk: 1, uk: 2, wholesale: 1, cannabis: 2, supplier: 1, weed: 1, ...)
	description	CannabisUK - buy weed in the UK with Bitcoins - we are UKs #1 marijuana vendor on the deep web
	Links	Array(hyperlinks in this URL)
	Word frequency	Object (cannabisuk: 2, kushbeautiful: 1, crystal: 1, kush: 5, afghani: 3, aroma: 1, ...)

3.6 Automatic labelling the dataset

The initial step involves the selection of four categories of illegal activities on the dark web including, (drugs, weapons, fake ID, and hacking) where many types of illegal trade are active in the dark web, as mentioned in Chapter Two, Section 2.2.3. It is based on a security perspective that the categories used in this research were selected. Particularly, these types of trafficking pose a direct security threat to the society. The aforementioned categories selected in this research are offences that are punishable under the laws of Iraq such as, the Anti-Terrorism Law within the provisions of Article 4, the Iraqi Penal Code No. 111 of 1969 as amended, the Iraqi Weapons Law No. 51 of 2017 within the provisions of Article 24, and the Iraqi Narcotics and Psychotropic Substances Law No. 50 of 2017 within the provisions of Article 28.

In this research, an algorithm has been built and named automatic labelling algorithm, which is used in labelling the dataset automatically. It was observed that in previous related works, the authors manually labelled datasets. Consequently, this algorithm was designed to achieve the task of labelling automatically, which in turn saves time. The automatic labelling algorithm has classified the dataset into five classes. The proposed algorithm consists of the following steps:

3.6.1 Loading of Dataset

This stage involves loading the algorithm with the dataset, which is to be read and divided into two parts. The first part is the columns in which numeric values like word frequency and title keywords are contained, whereas, the second part contains texts such as title and description. More so, in the first part, there are words that are representative of the keywords, with each keyword having a given weight which is the number of times a given keyword occurs in a document.

3.6.2 Handling the keyword weights

This step involves processing the keyword weights in the dataset. The existence of blank cells is attributed to the fact that the number of columns in the dataset represents the number of keywords. Therefore, there are so many keywords that are found in the document, but are absent in other documents, and as such, these keywords take weight in the documents they contain, while the values remain empty in the documents which they are not found in. this issue was resolved in the first part of the dataset, where there were columns for word frequency and title keywords. In order to solve the problem, a weight of "zero" was assigned. The second part of the dataset which contained the title and description. In some cases, a description is absent from a document, so the word "nothing" was used.

3.6.3 Categories Lists

During the process of crawling and collection of data, each crawled link was subjected to the process of pre-processing, and afterwards, the tokens were extracted from all texts. Having completed the process of extraction, the process of data cleaning was applied on the data and assigning weight to each token based on how frequent it appears in the page by calculating term frequency (tf).

After the relevant keyword were extracted, they were placed in the different categories used in this research. For instance, any keyword associated with drugs was placed in a list labeled drugs. Table 3.3 shows the number of the most important keywords for each category.

Table (3.3): The number of important keywords for each category.

No.	Category name	Number of keywords
1	Weapon	104
2	Drug	90
3	Fake ID	19
4	Hacking	13

3.6.4 Labeling the Dataset

Prior to the implementation of this process, modifications were made to the weights of the keywords extracted from the title, and this was done by multiplying the weight by ten, because they are critical keywords that enable the algorithm to determine the page a category belongs to.

The automatic labelling algorithm consists of two phase:

- **The First phase**

The first step in this phase is the creation of a data frame by the proposed algorithm, comparing all the lists of keywords for all classes with the tokens in each page. More so, a score is assigned to a column that represents the name of class for each keyword that exits in the page (similarity compute). This means that, when the process of comparison is completed, all the keywords listed in a given category are calculated and their presence in a particular page is confirmed through verification, which begins from the first row to the last row in the dataset, and placing this value in a column indicating the name of the category.

Using the scores obtained from this process as seen in Table 3.4 below, the process of automatic labelling is implemented based on the highest score recorded for each document, and the page label is determined by the highest score for any of the categories.

Table (3.4): Score for each class based on keywords in document.

No. of webpages in dataset	Drugs Score	Fake ID Score	Weapons Score	Hacking Score
1	30	1	93	4
2	28	0	37	0
3	36	11	90	11
4	5	0	66	0
5	4	0	61	0
...	...	...	...	...
2359	18	13	15	12

2360	18	16	13	14
2361	19	4	20	6
2362	5	3	20	5
2363	21	7	17	8

If it happens that there are close scores for a number of categories, a condition has been set to solve this problem. However, if the condition is not met, the algorithm labels the page as others.

- **The second phase**

Within the pages of the dataset used in this research, there is a column that is tagged as “description” and contains only texts. Several words found in this text are very essential keywords, and as such a condition was set to enable the specification of the basic keywords for each category (Steriod, Fake, Gun, and Hack); one keyword is assigned to each category. If the presence of a keyword is observed in the description, then the labelling is done based on the found keyword, thereby helping to correct some labelling errors that could have occurred at the first phase. Therefore, the proposed algorithm worked as a supervised machine learning.

3.6.5 Evaluation of the automatic labeling algorithm

The proposed automatic labeling algorithm has been evaluated so as to ascertain its efficiency in terms of accuracy. The evaluation was done by labelling the dataset manually by an expert team just like in previous works, where the error rate in the manually labelling is zero percent.

The accuracy of the algorithm was measured using a simple code that was created; the code was used to compare the label column for manually labelled dataset with that which was labelled automatically. Also, the error rate and accuracy rate were calculated as shown in equations 3.1 and 3.2 below.

Error rate = number of error / total number of documents (3.1)

$$\text{Accuracy} = 1 - \text{Error rate} \qquad (3.2)$$

Algorithm 3.3 shows the steps used for the automatic labeling.

Algorithm 3.3: Automatic labeling dataset

Input: Set of dark web pages unlabeled as dataset, drug keywords list (DKL), weapon keywords list (WKL), fake-id keywords list (FKL), hacking keywords list (HKL), basic keywords list (BKL)

Output: Labeled Dataset

```
Begin
Step1: read the dark web page
Step2: read the keyword lists for classes
Step3: // compute score for each class based on keywords
        i = 1
        N = number of all tokens
        D = 0 // score of drug class
        W = 0 // score of weapons class
        F = 0 // score of fake-id class
        H = 0 // score of hacking class
        While i <= N do
        Begin
            If token [i] is found in DKL then
                D = D+1
            else
                If token [i] is found in WKL then
                    W = W+1
                else
                    If token [i] is found in FKL then
                        F = F+1
                    else
                        If token [i] is found in HKL then
                            H = H+1
                        End If
                    End If
                End If
            i = i+1
            End If
        End While
    Step4:// Choosing the highest scores for two classes
        i = 1
        N = number of dark web pages
        $S_0$, $S_1$ = highest scores for two class
        While i <= N do
        Begin
            If ($S_0$-$S_1$)>2 and $S_0$>4 then
                label = the name of ($S_0$) class
            else
                label = others
            i = i+1
            End If
        End While
    Step5: // comparison the basic keywords for classes [drug, fake, weapon,
        hack] with description dark web page
        i = 1
        N = number of all tokens in page description
        While i <= N do
        Begin
            If token[i] is found in BKL then
```

```
                label = token[i]
            i=i+1
            End If
        End while
End
```

3.7 Classification Methods

The dark web marketplaces were classified using three automatic classification systems (LSVM, RF classifier, and NB classifier); the classification was done based on what is sold and bought in a given market. Here, all the words on the page crawled by the crawler and the data collected from that page are considered as the features of that page.

The classification was performed subsequent to many other operations that were implemented on the dataset, including preprocessing, and labelling the dataset. After the classifiers have been trained, the system was ready for testing, which was done by entering the testing page into the system, and subjecting it to processing at many steps until the system predicts the correct class for page used for testing. It is important to note that, for the purposes of training and testing, the dataset was divided into sets (training and testing sets), with the training set constituting 70% and the testing set 30% of the entire dataset.

Subsequent to the application of the three classifiers on the dataset, it was observed that the linear SVM demonstrated superior performance compared to the RF classifier, and NB classifier. Thus, the use of the linear SVM was employed in this research to classify the dark web marketplaces.

3.7.1 SVMs Classifier

SVM algorithms could be described as the function of distinguishing the two positive and negative classes in a feature space. The data points are defined as positive or negative when the problem is to define a hyper-plane that serves as a separator between two kinds of data points.

Two types of SVMs can work with binary classification. The first one is Linear SVM, which is used if there is a hyper-plane can spate the data into two classes will linear. The second type is non-linear which is used when there is no linear line to separate the data and use hypotheses.

Algorithms usually use several mathematical functions known as the kernel function. Kernel functions for data sets, graphs, text, images, and vectors take input data and convert it to the required form; all SVM algorithms use different kernel functions, For example non-linear, linear, radial basis function (RBF), polynomial and sigmoid.

The kernel functions allocate the inner product to an appropriate space between two points. Therefore, by defining a definition of similarity, with minimal computational cost, in some very high-dimensional spaces.

Since we have five classes and the LSVM algorithm works to classify the data into two classes, we will take the first class and let it be for example drug and the rest of the four classes will all be as a second class to be separated linearly and based on the keywords extracted by the term frequency that will define the class label for each dark web page entered into the classifier, and so on for the rest of the categories.

With using linear equations (2.4 and 2.5) to plot the hyper-plane.

For any vector X can calculate y=w.x+b. where if the Y value is greater than zero then class y=+1 then the class is a drug, else if the Y value is less than zero then class y=-1belong to one of the other fourth classes.

The LSVM classifier will train on a portion of the data that is 70 percent of the dataset which is dark web pages with labelling and test on the remaining portion of the dataset that is 30 percent that is unlabeled dark web pages by applying the pre-processing steps and calculating the TF for all the words on the test page, which enters the SVM to calculate the similarity of each word on the test page with each word on the training page. The model will return the dark web page where the similarity or high similarity as a page for drug trafficking and so on for the rest of the classes.

Algorithm 3.4 illustrates the training phase and testing phase of the proposed system.

Algorithm 3.4: training and testing phase of the system

```
Input: Dataset as set of dark web pages.
Output: Dark web pages classification.
Begin
Step1: Building the dataset by dark crawler (algorithm 3.1)
Step2: preprocessing the dataset by using, data cleaning, tokenization, and
          remove stop words (algorithm 3.2).
Step3: labelling the dataset by our autolabeling algorithm (algorithm 3.3).
Step4: Training phase
          Read the labeled dataset.
Step5: Testing phase
          Read the unlabeled dataset.
Step6: Classify the dark web pages based on keywords.
End
```

Chapter Four

The Experimental Results and Discussion

Chapter Four
The Experimental Results and Discussion

4.1 Overview

The proposed methodology illustrated in chapter three was implemented so that the research objectives outlined in chapter one can be achieved. Therefore, some experiments were performed in order to achieve the objectives of the research, and the results of the experiments are presented in this chapter.

4.2 Software and Hardware Requirements

The proposed system was implemented using the following hardware and software requirements:

Hardware: Processor Intel i7, Freq. 2.1 GHz, RAM 4GB, Storage 500 GB.

Operating System: Windows10 (64) bit.

Programming language: Python language.

IDLE: the system was implemented by Python 3.9 shell, Visual Studio Code.

Browser: TOR browser.

4.3 Results of Crawling and Data Collection

The dark crawler proposed in this research was used to obtain about 2,363 dark web pages. After each process of crawling of any link, the link obtained is fed into the crawler and the data of that link is extracted. After this step, the data is subjected to pre-processing. The data is processed directly using selected data pre-processing techniques and afterwards saved in the database, which has been named MongoDB as noted in the previous chapter. Figure 4.1 shows a sample of dark web pages crawled by a crawler and their contents that were retrieved prior to data pre-processing.

```
'<!DOCTYPE html>\r\n<!--[if lt IE 7 ]><html class="ie ie6" lang="en">
<![endif]-->\r\n<!--[if IE 7 ]><html class="ie ie7" lang="en">
<![endif]-->\r\n<!--[if IE 8 ]><html class="ie ie8" lang="en">
<![endif]-->\r\n<!--[if (gte IE 9)|!(IE)]><!--><html lang="en"> <!--
<![endif]-->\r\n<head>\r\n\t<meta charset="utf-8">\r\n\t<title>Black
Market - Guns Arms Ammo Drugs for Bitcoin - Supplier since
2001</title>\r\n\t<meta name="description" content="Choose right now
what do you need, we are shipping worldwide and guarantee the successful
shipping of all our physical good.">\r\n\t<meta name="author"
content="Black Market">\r\n\t<meta name="viewport"
content="width=device-width,initial-scale=1,maximumscale=1">\r\n\t<link
rel="stylesheet"href="stylesheets/base.css">\r\n\t<linkrel="stylesheet"
href="stylesheets/skeleton.css">\r\n\t<linkrel="stylesheet"href="styles
heets/layout.css">\t\r\n</head>\r\n<body>\r\n\t\t\t\t\t\t\t\t\t\t\t\t\r
\n\r\n<div class="container">\r\n\t\t<div style="text-align:center;"><a
href="shop.php"><imgsrc="images/logo.png"alt=""width="267px"/></a></div
><br/><br/>\r\n\r\n\t\t <div class="one-third column">\r\n\t\t <div
id="imgdeag"></div>\r\n\t\t\t<divstyle="textalign:center;">\r\n\t\t\t<a
href="images/bm/dea1.jpg"target="_blank">1</a>-\r\n\t\t\t<a
href="images/bm/dea2.jpg"target="_blank">2</a>\r\n\t\t\t<ahref="images/
bm/dea3.jpg"target="_blank">3</a>\r\n\t\t\t</div>\r\n\t\t\t<br/>\r\n\t\
t\t<h3 style="text-align:center;">Desert Eagle 357 Mag GOLD TIGER
STRIPE</h3>\r\n\t\t\t<p><strong>Features:</strong> <br/>\r\n\t\t\t
Manufacturer: Magnum Research <br/>\r\n\t\t\t Model: Deseart Eagle 357
<br/>\r\n\t\t\t100%TitaniumGoldTigerstripe<br/>\r\n\t\t\tMagnumrated<br
/>\r\n\t\t\t<br/>\r\n\t\t\t<strong>Specifications:</strong><br/>\r\n\t\
t\tCaliber:357Mag<br/>\r\n\t\t\tFinish:\tTitaniumold<br/>\r\n\t\t\tBarr
elLength:\t6inch.<br/>\r\n\t\t\tCapacity:\t8 <br/>\r\n\t\t\tNumber of
Mags:1<br/>\r\n\t\t\tType:\tSemiAutomaticPistol<br/><br/>\r\n\t\t\tPric
eonmarket<strong>2200$</strong></p>\r\n\t\t\t<divstyle="margin:auto;pad
ding:5px11px9px11px;background:rgb(39,197,0);textalign:center;width:aut
o;position:relative;display:<br/><br/>\r\n\t\t</div>\r\n\t\t\r\n\t\t<di
vclass="one-thirdcolumn">\r\n\t\t\t<divid="imgrem"></div>\r\n\t\t\t<div
```

Figure (4.1): the dark web page crawled by a crawler in HTML format.

4.4 Results of Data Pre-processing

The results of the pre-processing stage are presented in this sub-section. The pre-processing steps involved the data cleaning of the pages by removing irrelevant information, tokenization of words (splitting of words), removing stop words, and extraction of URL, page title, and description. The outcomes of the pre-processing stage were a collection of tokens for each text.

- **Removing Tag and Extracting the Text**

Table 4.1 shows the results for tag removal and extraction of text from the content of the dark web page.

Table (4.1): Tag Removal.

Before Removing	'<!DOCTYPE html>\r\n<!--[if lt IE 7]><html class="ie ie6" lang="en"> <![endif]-->\r\n<!--[if IE 7]><html class="ie ie7" lang="en"> <![endif]-->\r\n<!--[if IE 8]><html class="ie ie8" lang="en"> <![endif]-->\r\n<!--[if (gte IE 9)|!(IE)]><!--><html lang="en"> <!--<![endif]-->\r\n<head>\r\n\t<meta charset="utf-8">\r\n\t<title>Black Market - Guns Arms Ammo Drugs for Bitcoin - Supplier since 2001</title>\r\n\t<meta name="description" content="Choose right now what do you need, we are shipping worldwide and guarantee the successful shipping of all our physical good.">\r\n\t<meta name="author" content="Black Market">\r\n\t<meta name="viewport" content="width=device-width, initial-scale=1, maximum-scale=1">\r\n\t<link rel="stylesheet" ...
After Removing	'Black Market - Guns Arms Ammo Drugs for Bitcoin - Supplier since 2001\r\n\t\r\n\t\r\n\t\r\n\t\r\n\t\r\n\t\t\r\n\r\n\r\n\t\t\t\t\t \t\t\t\t\t\t\t\r\n\r\n\r\n\t\t\r\n\r\n\t\t\r\n\t\t\r\n\t\t\t\r\n\t\t\t1-\r\n\t\t\t2-\r\n\t\t\t3\r\n\t\t\t\r\n\t\t\t\r\n\t\t\tDesert Eagle 357 Mag GOLD TIGER STRIPE\r\n\t\t\t Features:\r\n\t\t\t Manufacturer: Magnum Research \r\n\t\t\t Model: Deseart Eagle 357 \r\n\t\t\t100% Titanium Gold Tiger stripe \r\n\t\t\tMagnum rated \r\n\t\t\t\r\n\t\t\tSpecifications: \r\n\t\t\tCaliber: 357 Mag \r\n\t\t\t Finish:\tTitanium Gold\r\n\t\t\tBarrel Length:\t6 inch. \r\n\t\t\tCapacity:\t8\r\n\t\t\tNumber of Mags: 1\r\n\t\t\tType:\t Semi-Automatic Pistol \r\n\t\t\tPrice on market 2200$ \r\n\t\t\t\r\n\t\t\tour price \r\n\t\t\t$800=0.0162 BTC \r\n\t\t ..

- **Removal of Numbers**

In general, numbers are not an important expression in the scope of this research, so they have been removed from the page content to improve the content. Table 4.2 shows a sample of the original text of the page content after numbers were removed from the text.

Table (4.2): Removing numbers.

Before Removing	'Black Market - Guns Arms Ammo Drugs for Bitcoin - Supplier since 2001\r\n\t\r\n\t\r\n\t\r\n\t\r\n\t\r\n\t\t\r\n\r\n\r\n\t\t\t\t \t\t\t\t\t\t\r\n\r\n\r\n\t\t\r\n\r\n\t\t\r\n\t\t\r\n\t\t\t\r\n\t\t\t1-\r\n\t\t\t2-\r\n\t\t\t3\r\n\t\t\t\r\n\t\t\t\r\n\t\t\tDesert Eagle 357 Mag GOLD TIGER STRIPE\r\n\t\t\t Features:\r\n\t\t\t Manufacturer: Magnum Research \r\n\t\t\t Model: Deseart Eagle 357 \r\n\t\t\t100% Titanium Gold Tiger stripe \r\n\t\t\tMagnum rated \r\n\t\t\t\r\n\t\t\tSpecifications: \r\n\t\t\tCaliber: 357 Mag \r\n\t\t\t Finish:\tTitanium Gold\r\n\t\t\tBarrel Length:\t6 inch. \r\n\t\t\tCapacity:\t8\r\n\t\t\tNumber of Mags: 1\r\n\t\t\tType:\t Semi-Automatic Pistol \r\n\t\t\tPrice on market 2200$ \r\n\t\t\t\r\n\t\t\tour price \r\n\t\t\t $ 800=0.0162 BTC \r\n\t\t \r\n\t\t\r\n\t\t\r\n\t\t\r\n\t\t\t\r\n\t\t\t1 -\r\n\t\t\t2 - 3 - \r\n\t\t\t4 - 5 ...
After Removing	'Black Market - Guns Arms Ammo Drugs for Bitcoin - Supplier since \r\n\t\r\n\t\r\n\t\r\n\t\r\n\t\r\n\t\t\r\n\r\n\r\n\t\t\t\t\t\t\t\t\t\t\r\n\r\n\r\n\t \t\r\n\r\n\t\t \r\n\t\t \r\n\t\t\t\r\n\t\t\t -\r\n\t\t\t - \r\n\t\t\t \r\n\t\t\t\r\n\t\t\t\r\n\t\t\tDesert Eagle Mag GOLD TIGER STRIPE\r\n\t\t\tFeatures: \r\n\t\t\tManufacturer: Magnum Research \r\n\t\t\tModel: Deseart Eagle \r\n\t\t\t% Titanium Gold Tiger stripe\r\n\t\t\tMagnum rated\r\n\t\t\t\r\n\t\t\tSpecifications: \r\n\t\t\tCaliber: Mag \r\n\t\t\tFinish:\tTitanium Gold\r\n\t\t\tBarrel Length:\t inch. \r\n\t\t\tCapacity:\t \r\n\t\t\tNumber of Mags: \r\n\t\t\tType:\t Semi-Automatic Pistol\r\n\t\t\tPrice on market $\r\n\t\t\t\r\n\t\t\tour price\r\n\t\t\t$=. BTC\r\n\t\t \r\n\t\t\r\n\t\t\r\n\t\t\r\n\t\t\t\r\n\t\t\t -\r\n\t\t\t - - \r\n\t\t\t - - \r\n\t\t\t ...

- **Removal of Punctuations**

In this step, Punctuations were removed, text was converted into lowercase, and double space were removed from page texts. Table 4.3 contains a sample of the page contents before and after the punctuations were removed.

Table (4.3): Removal of Punctuations.

Before Removing	'Black Market - Guns Arms Ammo Drugs for Bitcoin - Supplier since \r\n\t\r\n\t\r\n\t\r\n\t\r\n\t\r\n\t\t\r\n\r\n\r\n\t\t\t\t\t\t\t\t\t\t\t\r\n\r\n\r\n\t\t\r\n\r\n\t\t \r\n\t\t \r\n\t\t\t\r\n\t\t\t -\r\n\t\t\t - \r\n\t\t\t \r\n\t\t\t\r\n\t\t\t\r\n\t\t\tDesert Eagle Mag GOLD TIGER STRIPE\r\n\t\t\tFeatures: \r\n\t\t\tManufacturer: Magnum Research \r\n\t\t\tModel: Deseart Eagle \r\n\t\t\t% Titanium Gold Tiger stripe\r\n\t\t\tMagnum rated\r\n\t\t\t\r\n\t\t\tSpecifications: \r\n\t\t\tCaliber: Mag \r\n\t\t\tFinish:\tTitanium Gold\r\n\t\t\tBarrel Length:\t inch. \r\n\t\t\tCapacity:\t \r\n\t\t\tNumber of Mags: \r\n\t\t\tType:\t Semi-Automatic Pistol\r\n\t\t\tPrice on market $\r\n\t\t\t\r\n\t\t\tour price\r\n\t\t\t$=. BTC\r\n\t\t \r\n\t\t\r\n\t\t\r\n\t\t\r\n\t\t\t\r\n\t\t\t -\r\n\t\t\t - - \r\n\t\t\t - - \r\n\t\t\t - - \r\n\t\t\t - - \r\n\t\t\t - \r\n\t\t\tRemington Defense ...
After Removing	'black market guns arms ammo drugs for bitcoin supplier since desert eagle % mag gold tiger stripe features manufacturer magnum research model deseart eagle titanium gold tiger stripe magnum rated specifications caliber mag finish titanium gold barrel length inch capacity number of mags type semiautomatic pistol price on market our $ price $= btc remington defense xm sass this rifle was one of the designs that remington defense submitted for the military sass trials ...

- **Removing Special Characters**

Table 4.4 demonstrates a sample of the page contents before and after remove Special Characters.

Table (4.4): Removal of Special Characters.

Before Removing	'black market guns arms ammo drugs for bitcoin supplier since desert eagle % mag gold tiger stripe features manufacturer magnum research model deseart eagle titanium gold tiger stripe magnum rated specifications caliber mag finish titanium gold barrel length inch capacity number of mags type semiautomatic pistol price on market our $ price $= btc remington defense xm sass this rifle was one of the designs that remington defense submitted for the military sass trials ...
After Removing	'black market guns arms ammo drugs for bitcoin supplier since desert eagle mag gold tiger stripe features manufacturer magnum research model deseart eagle titanium gold tiger stripe magnum rated specifications caliber mag finish titanium gold barrel length inch capacity number of mags type semiautomatic pistol price on market our price btc remington defense xm sass this rifle was one of the designs that remington defense submitted for the military sass trials ...

- **Word Tokenization**

Table 4.5 shows the results of splitting page texts into their constituent words.

Table (4.5): Word tokenization

Before Tokenization	'black market guns arms ammo drugs for bitcoin supplier since desert eagle mag gold tiger stripe features manufacturer magnum research model deseart eagle titanium gold tiger stripe magnum rated specifications caliber mag finish titanium gold barrel length inch capacity number of mags type semiautomatic pistol price on market our price btc remington defense xm sass this rifle was one of the designs that remington defense submitted for the military sass trials ...
After Tokenization	['black', 'market', 'guns', 'arms', 'ammo', 'drugs', 'for', 'bitcoin', 'supplier', 'since', 'desert', 'eagle', 'mag', 'gold', 'tiger', 'stripe', 'features', 'manufacturer', 'magnum', 'research', 'model', 'deseart', 'eagle', 'titanium', 'gold', 'tiger', 'stripe', 'magnum', 'rated', 'specifications', 'caliber', 'mag', 'finish', 'titanium', 'gold', 'barrel', 'length', 'inch', 'capacity', 'number', 'of', 'mags', 'type', 'semiautomatic', 'pistol', 'price', 'on', 'market', 'our', 'price', 'btc', 'remington', 'defense', 'xm', 'sass', 'this', 'rifle', 'was', 'one', 'of', 'the', 'designs', 'that', 'remington', 'defense', 'submitted', 'for', 'the', 'military', 'sass', 'trials', ...]

- **Stop Words Removal**

During the text analysis carried out in this research, unimportant words, which are called stop words, were removed from the original text of the page. Table 4.6 shows a sample of a page after the removal of stop words.

Table (4.6): Removing Stop Words.

Before Removing	['black', 'market', 'guns', 'arms', 'ammo', 'drugs', 'for', 'bitcoin', 'supplier', 'since', 'desert', 'eagle', 'mag', 'gold', 'tiger', 'stripe', 'features', 'manufacturer', 'magnum', 'research', 'model', 'deseart', 'eagle', 'titanium', 'gold', 'tiger', 'stripe', 'magnum', 'rated', 'specifications', 'caliber', 'mag', 'finish', 'titanium', 'gold', 'barrel', 'length', 'inch', 'capacity', 'number', 'of', 'mags', 'type', 'semiautomatic', 'pistol', 'price', 'on', 'market', 'our', 'price', 'btc', 'remington', 'defense', 'xm', 'sass', 'this', 'rifle', 'was', 'one', 'of', 'the', 'designs', 'that', 'remington', 'defense', 'submitted', 'for', 'the', 'military', 'sass', 'trials', ...]
After Removing	['black', 'market', 'guns', 'arms', 'ammo', 'drugs', 'bitcoin', 'supplier', 'since', 'desert', 'eagle', 'mag', 'gold', 'tiger', 'stripe', 'features', 'manufacturer', 'magnum', 'research', 'model', 'deseart', 'eagle', 'titanium', 'gold', 'tiger', 'stripe', 'magnum', 'rated', 'specifications', 'caliber', 'mag', 'finish', 'titanium', 'gold', 'barrel', 'length', 'inch', 'capacity', 'number', 'mags', 'type', 'semiautomatic', 'pistol', 'price',

	'market', 'price', 'btc', 'remington', 'defense', 'xm', 'sass', 'rifle', 'one', 'designs', 'remington', 'defense', 'submitted', 'military', 'sass', 'trials', ...]

At the last step of the pre-processing, the content of all the dark web pages that have been crawled are split into URL, title, title keyword, description, and word frequency. Table 4.7 presents a sample of the dataset after pre-processing.

Table (4.7): Sample of the dataset

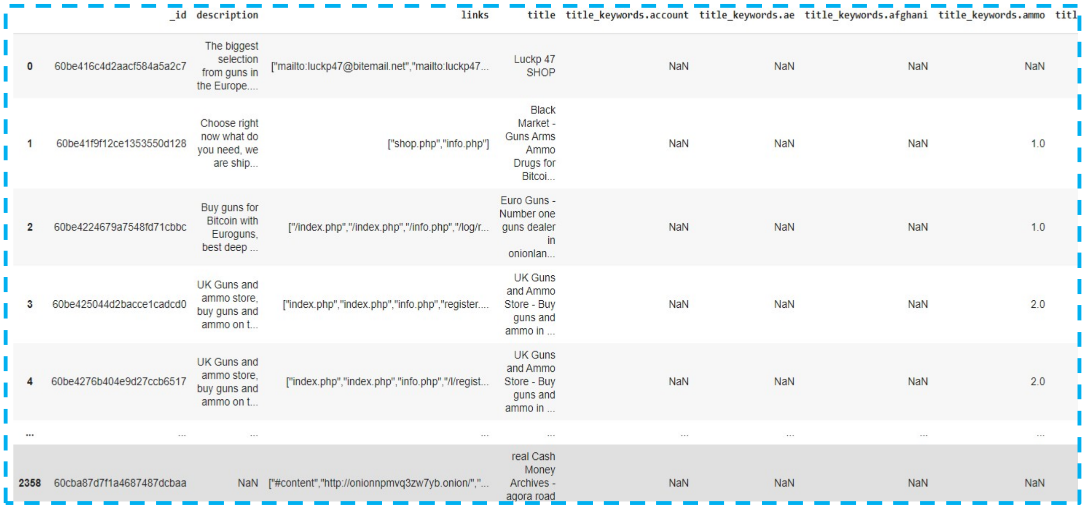
Page	**URL:** weapon5dj7fwz2zaqa22fqeaqrauclim5kfvecegphrvxeywxoa3wuid.onion/
	Title: Black Market - Guns Arms Ammo Drugs for Bitcoin - Supplier since 2001
	Title keyword: black:1, market:1, guns:1 , arms:1, ammo:1, drugs:1, bitcoin:1, supplier:1, since:1
	Description: Choose right now what do you need, we are shipping worldwide and guarantee the successful shipping of all our physical good
	Links: Array
	Word frequency: black:8, market:28, guns:1, arms:2, ammo:1, drugs:1, bitcoin:2, supplier:1, since:3, desert:1, eagle:2, mag:6, gold:6, tiger:2, stripe:2, features:3, manufacturer:8, magnum:5, research:1, model:12, deseart:1, titanium:4, gold:15, rated:2, specifications:24, caliber:15, finish:4, barrel:15, length:11, inch:4, capacity:9, number:3, mags:1, type:5, semiautomatic:4, pistol:1, price:48, btc:24, Remington:5, defense:3, xm:1, sass:2, rifle:7, one:4, designs:1, submitted:1, military:1, trials:1, ...

4.5 Results of Automatic labeling Algorithm

At the third stage of the proposed system, the dataset is labeled, using the automatic labelling algorithm proposed in this research, meaning the labelling was done automatically, rather than manually as done in several previous studies. This stage consists of several steps as follows:

- **Read the Dataset**

In this step, the dataset is read and displayed. These are illustrated in figure 4.2 below.



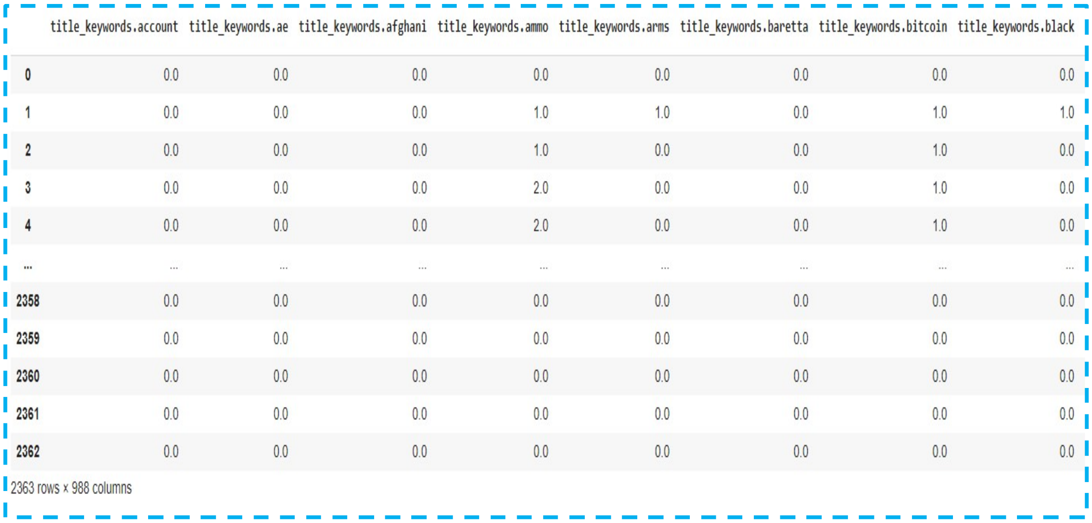
	_id	description	links	title	title_keywords.account	title_keywords.ae	title_keywords.afghani	title_keywords.ammo	titl
0	60be416c4d2aacf584a5a2c7	The biggest selection from guns in the Europe....	["mailto:luckp47@bitemail.net","mailto:luckp47...	Luckp 47 SHOP	NaN	NaN	NaN	NaN	
1	60be41f9f12ce1353550d128	Choose right now what do you need, we are ship...	["shop.php","info.php"]	Black Market - Guns Arms Ammo Drugs for Bitcoi...	NaN	NaN	NaN	1.0	
2	60be4224679a7548fd71cbbc	Buy guns for Bitcoin with Euroguns, best deep ...	["/index.php","/index.php","/info.php","/log/r...	Euro Guns - Number one guns dealer in onionlan...	NaN	NaN	NaN	1.0	
3	60be425044d2bacce1cadcd0	UK Guns and ammo store, buy guns and ammo on t...	["index.php","index.php","info.php","register....	UK Guns and Ammo Store - Buy guns and ammo in ...	NaN	NaN	NaN	2.0	
4	60be4276b404e9d27ccb6517	UK Guns and ammo store, buy guns and ammo on t...	["index.php","index.php","info.php","/l/regist...	UK Guns and Ammo Store - Buy guns and ammo in ...	NaN	NaN	NaN	2.0	
...	...	...	...	...	...	...	...	...	
2358	60cba87d7f1a4687487dcbaa	NaN	["#content","http://onionnpmvq3zw7yb.onion/","...	real Cash Money Archives - agora road	NaN	NaN	NaN	NaN	

Figure (4.2): Sample of the dataset before labeling.

It was noticed that the dataset contains the word (NaN) as indicated in figure 4.2 above. This is attributed to the fact that this is the stage in which data is pre-processed. All the texts contained in any dark web page that was crawled, were divided into tokens. These were regarded as the features of the document or webpage, and these tokens differ from one dark web page to another, so that the number of columns in the dataset is the number of total tokens for the dataset. Therefore, the token in a particular document cannot be present or repeated in a number of documents, so its value is left blank in the document in which it does not appear. This problem is addressed in the next step.

- **Results of Handling the Tokens Weights**

In this step, the weights of the tokens in the dataset are processed. Firstly, the dataset was divided into two parts; the first contains the title keywords columns and word frequency columns (which contain numeric values (weights)). Here, the tokens' weights were processed by placing a "zero" value in the empty cell. Figure 4.3 shows a sample of the first part that was processed.

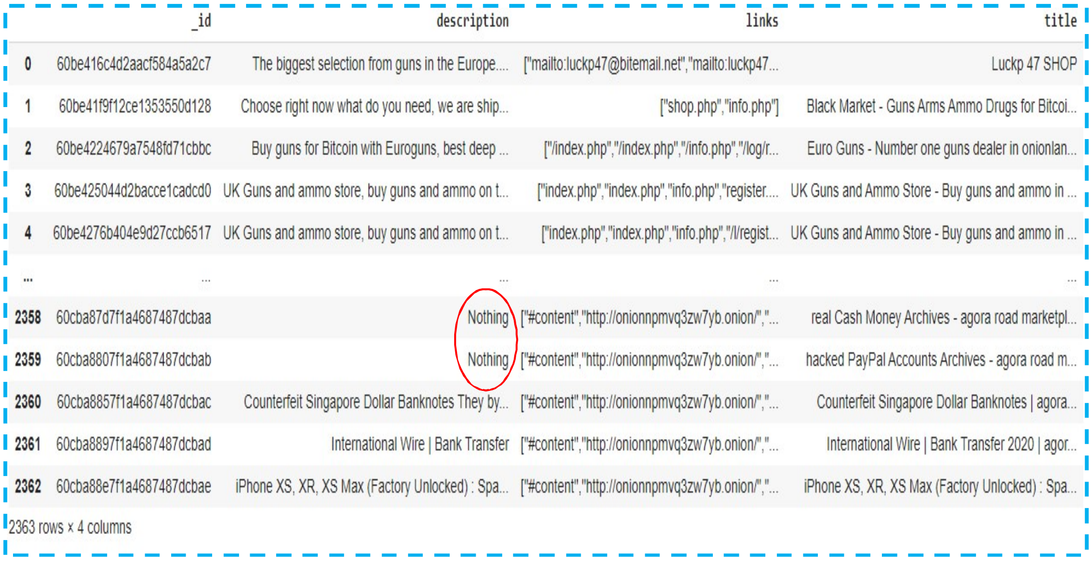


	title_keywords.account	title_keywords.ae	title_keywords.afghani	title_keywords.ammo	title_keywords.arms	title_keywords.baretta	title_keywords.bitcoin	title_keywords.black
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0
2	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
3	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0
4	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...
2358	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2359	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2360	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2361	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2362	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

2363 rows × 988 columns

Figure (4.3): Sample of first part after handling tokens value.

The second part of the dataset is the part that contains the title column of the crawled page and its description column. This part contains just texts, and as such, the missing texts are only found in the description column. This is because many of the pages that were retrieved do not contain descriptions, hence, the processing was done by putting the word "Nothing" in the empty cell. This is shown in Figure 4.4.

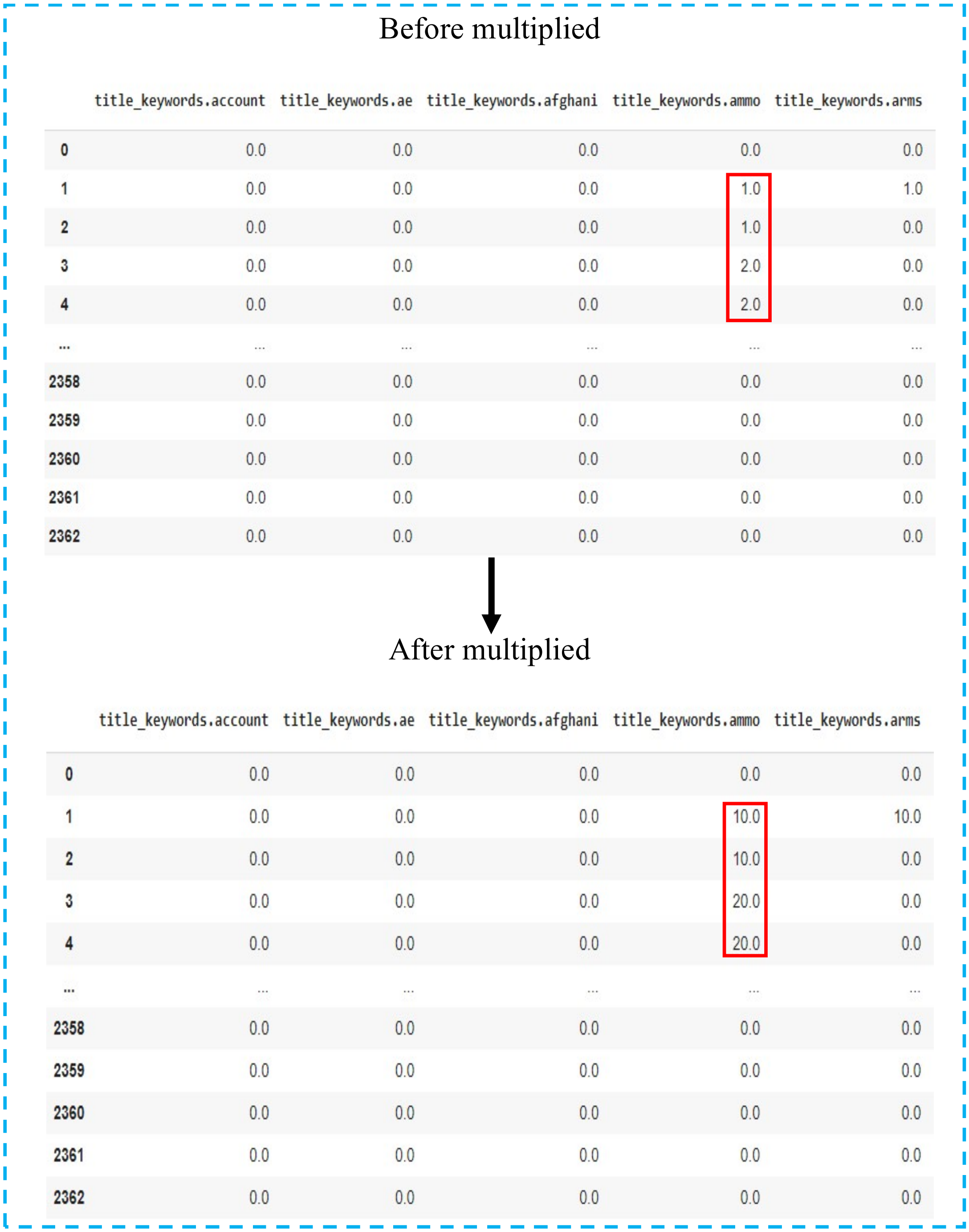


	_id	description	links	title
0	60be416c4d2aacf584a5a2c7	The biggest selection from guns in the Europe....	["mailto:luckp47@bitemail.net","mailto:luckp47...	Luckp 47 SHOP
1	60be41f9f12ce1353550d128	Choose right now what do you need, we are ship...	["shop.php","info.php"]	Black Market - Guns Arms Ammo Drugs for Bitcoi...
2	60be4224679a7548fd71cbbc	Buy guns for Bitcoin with Euroguns, best deep ...	["/index.php","/index.php","/info.php","/log/r...	Euro Guns - Number one guns dealer in onionlan...
3	60be425044d2bacce1cadcd0	UK Guns and ammo store, buy guns and ammo on t...	["index.php","index.php","info.php","register....	UK Guns and Ammo Store - Buy guns and ammo in ...
4	60be4276b404e9d27ccb6517	UK Guns and ammo store, buy guns and ammo on t...	["index.php","index.php","info.php","/l/regist...	UK Guns and Ammo Store - Buy guns and ammo in ...
...	...	...	...	...
2358	60cba87d7f1a4687487dcbaa	Nothing	["#content","http://onionnpmvq3zw7yb.onion/","...	real Cash Money Archives - agora road marketpl...
2359	60cba8807f1a4687487dcbab	Nothing	["#content","http://onionnpmvq3zw7yb.onion/","...	hacked PayPal Accounts Archives - agora road m...
2360	60cba8857f1a4687487dcbac	Counterfeit Singapore Dollar Banknotes They by...	["#content","http://onionnpmvq3zw7yb.onion/","...	Counterfeit Singapore Dollar Banknotes | agora...
2361	60cba8897f1a4687487dcbad	International Wire | Bank Transfer	["#content","http://onionnpmvq3zw7yb.onion/","...	International Wire | Bank Transfer 2020 | agor...
2362	60cba88e7f1a4687487dcbae	iPhone XS, XR, XS Max (Factory Unlocked) : Spa...	["#content","http://onionnpmvq3zw7yb.onion/","...	iPhone XS, XR, XS Max (Factory Unlocked) : Spa...

2363 rows × 4 columns

Figure (4.4): Sample of second part after handling missing texts.

After the step involving the processing of tokens weights and missing texts, the next step involved the automatic labelling of dataset. The tokens in the title keywords are important tokens that can enable the accurate labeling of the document, and for that reason, the weight has been multiplied by ten as seen in Figure 4.5.



Before multiplied

	title_keywords.account	title_keywords.ae	title_keywords.afghani	title_keywords.ammo	title_keywords.arms
0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	1.0	1.0
2	0.0	0.0	0.0	1.0	0.0
3	0.0	0.0	0.0	2.0	0.0
4	0.0	0.0	0.0	2.0	0.0
...	...	...	...	...	...
2358	0.0	0.0	0.0	0.0	0.0
2359	0.0	0.0	0.0	0.0	0.0
2360	0.0	0.0	0.0	0.0	0.0
2361	0.0	0.0	0.0	0.0	0.0
2362	0.0	0.0	0.0	0.0	0.0

After multiplied

	title_keywords.account	title_keywords.ae	title_keywords.afghani	title_keywords.ammo	title_keywords.arms
0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	10.0	10.0
2	0.0	0.0	0.0	10.0	0.0
3	0.0	0.0	0.0	20.0	0.0
4	0.0	0.0	0.0	20.0	0.0
...	...	...	...	...	...
2358	0.0	0.0	0.0	0.0	0.0
2359	0.0	0.0	0.0	0.0	0.0
2360	0.0	0.0	0.0	0.0	0.0
2361	0.0	0.0	0.0	0.0	0.0
2362	0.0	0.0	0.0	0.0	0.0

Figure (4.5): Sample of doubling the title keywords weights.

Results of Labeling the Dataset

After preprocessing stage and splitting text into tokens, the keywords of each class are extracted by calculating term frequency for all words in the dark web page that crawled and are placed in a list that carries the name of the same class as shown in Table 4.8. Afterward, the proposed algorithm begins to label the dataset automatically based on the keywords for each class. The automatic labeling is carried out in two steps.

Table 4.8: Sample the keywords for class drug as list.

Drugs Class	Important keywords
	array(['substance', 'drug', 'heroin', 'cocaine', 'LSD', 'cannabis', 'hush', 'chemical', 'mdma', ' meth', 'kush', 'crystal', 'coca', 'weed', 'afghani', 'dragon', 'yellow', ' red', 'orange'. 'fealar', 'bullets', 'steriod', ...])

The first step:

The algorithm calculates the scores for each class based on the keywords by measuring the similarity between class and document keywords, and counting the number of keywords that are found in that document. Therefore, the document is labeled based on the class that achieves the highest score. Usually, the activities of many dark web markets are not limited to the trade of a specific type of merchandise, such as drugs, weapons, or others. Rather, the market includes a number of illegal activities, such as the Dream Market and Dark Fox. Therefore, the scores for a number of categories are very close, and to solve this problem, a fifth category was added to carter for that category of activities. Table 4.9 shows the result of this step.

Table (4.9): Sample of the outcomes of labeling dataset step one.

No.	Drugs Score	Fake ID Score	Weapons Score	Hacking Score	Class Label
1	30	1	93	4	Weapons
2	28	0	37	0	Weapons
3	36	11	90	11	Weapons
4	5	0	66	0	Weapons
5	4	0	61	0	Weapons
...	...	...	...	...	...
2359	18	13	15	12	Drugs
2360	18	16	13	14	Others
2361	19	4	20	6	Others
2362	5	3	20	5	Weapons
2363	21	7	17	8	Drugs

- **The second step:**

Subsequent to the automatic labeling of the dataset which was performed in the first step, it was observed some errors occurred during the labeling of the documents, where many documents contained basic keywords (such as Drug, Fake, Weapon, and Hack) in the description indicating its content. Through these words, the document can be labelled directly, placing it under the class it belongs to. These errors were corrected at this stage by comparing the description of each document with these keywords and labeled based on them. Table 4.10 presents the outcomes of this step.

Table (4.10): Sample of the outcomes of labeling dataset step two.

No.	Drugs Score	Fake ID Score	Weapons Score	Hacking Score	Class Label St.1	Class Label Final
1	30	1	93	4	Weapons	Weapons
2	28	0	37	0	Weapons	Weapons
3	36	11	90	11	Weapons	Weapons
4	5	0	66	0	Weapons	Weapons
5	4	0	61	0	Weapons	Weapons
...	...	...	...	...	...	...
2359	18	13	15	12	Drugs	Drugs
2360	18	16	13	14	Others	Hacking
2361	19	4	20	6	Others	Others
2362	5	3	20	5	Weapons	Weapons
2363	21	7	17	8	Drugs	Drugs

4.5.1 Result of evaluate the automatic labeling algorithm

To ensure the accuracy of the automatic labeling algorithm, the dataset used in this research was manually labeled by the judge and the result of the class label was compared that which was labelled automatically. In other words, a comparison was done between the manually labelled dataset and the automatically labelled dataset so as to determine the performance of the proposed automatic labelling algorithm in terms of accuracy. The number of errors in labeling documents for automatic labeling was calculated as shown in table 4.11. It is worth mentioning that, the error rate in manual labeling is almost zero percent, so the error rate was calculated to determine the accuracy of the proposed algorithm.

Table (4.11): Sample of the dataset auto-labeling to calculate error

No.	Drugs Score	Fake ID Score	Gun Score	Hacking Score	Autolabel	Judge Label
1	13	0	12	0	Others	Drugs
2	27	0	12	0	Drugs	Drugs
3	16	0	10	0	Drugs	Drugs
4	30	0	19	0	Drugs	Drugs
5	10	1	8	2	Hacking	Hacking
6	52	0	36	0	Drugs	Drugs
7	12	0	12	0	Others	Drugs
8	10	0	21	0	Weapons	Weapons
9	26	0	12	0	Drugs	Drugs
10	11	0	8	2	Drugs	Drugs

The results showed that an accuracy rate of 85% was achieved by the proposed automatic labeling algorithm. The accuracy rate is calculated as presented below:

- Error rate = number of error / total number of document

  = 343/2363 = **0.145**

- Accuracy = 1 – Error rate

  = 1 – 0.145 = **0.854**

The resulting dataset (dark web-db) consists of 2363 samples distributed over five classes i.e., the four classes plus the other one created to carter for the other activities that do not fall under the four classes used in this research. Table 4.12 and figure 4.6 show the number of documents for each class. Upon completion of the automatic labelling process, it was found that the largest percentage of the dark web pages that were crawled was related to drug trade, as indicated in figure 4.7.

Table (4.12): The number of documents in each class in the dataset.

No.	The name of class	The number of documents in each class
1	Drugs	968
2	Others	614
3	Hacking	259
4	Fake ID	253
5	Weapons	251

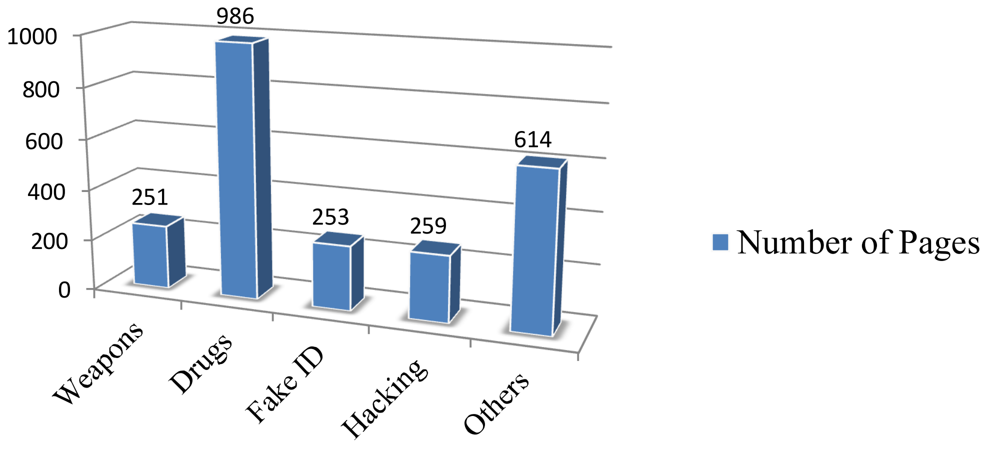


Figure (4.6): The number of pages of each class in the dataset.

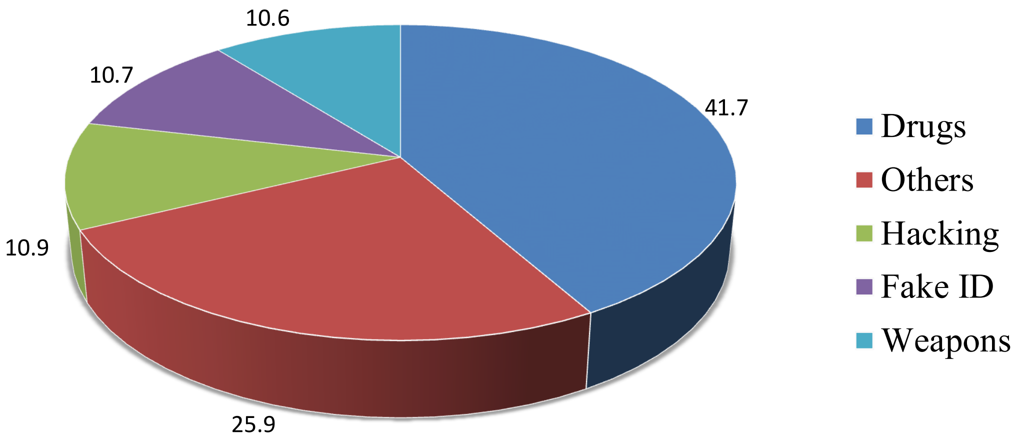


Figure (4.7): The percentage of each class in the dataset.

Figure 4.8 shows the number of documents of each class in the dataset for auto-labeling and manually labeling.

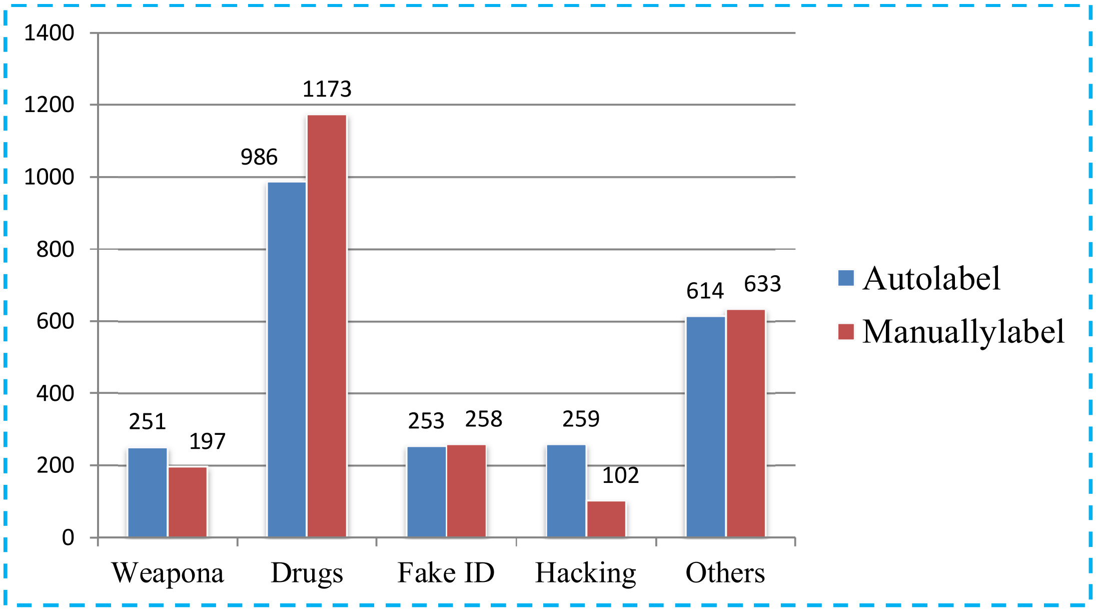


Figure (4.8): The deferent between autolabeling and manually labeling.

4.6 Results of Classification Methods

The total number of classified dark web pages reached 2,363. The entire dataset was divided into two (training and testing set). The training set had 70% of the dataset, and the remaining 30% was made up the testing set. The pages obtained were used for classification. The classification of the dark web pages was done using three techniques, including linear SVM, Random Forest (RF), and Gaussian Naïve Bayes Classifier (NB). Linear SVM outperformed other classifier with an accuracy of 91%, while the accuracy of RF was 89%, and NB is 81%. The reason for this is that linear SVM works well with text categorization of high-dimensional input space. Figure 4.9 depicts the accuracy chart of the classifiers that were implemented.

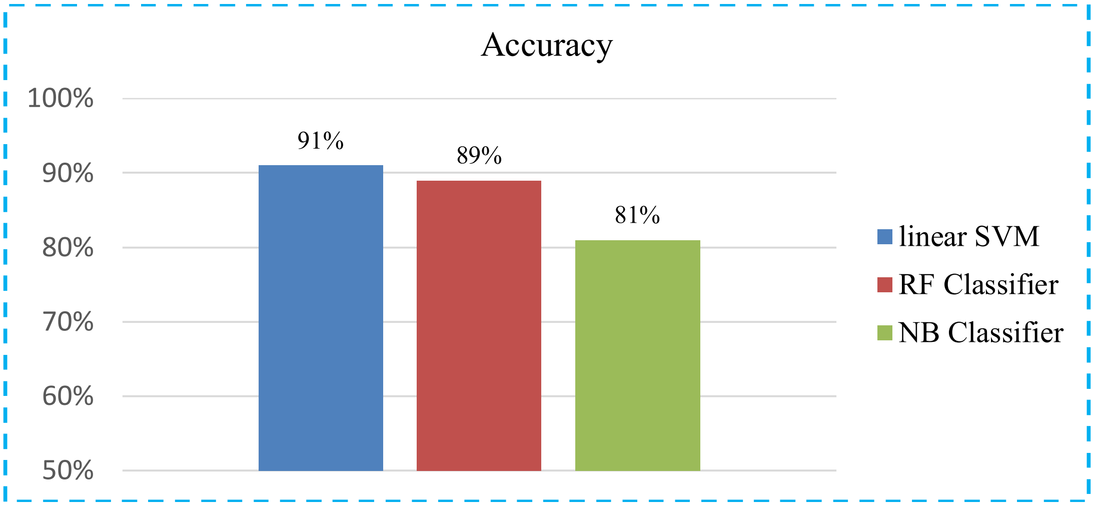


Figure (4.9): The accuracy chart of the classifiers.

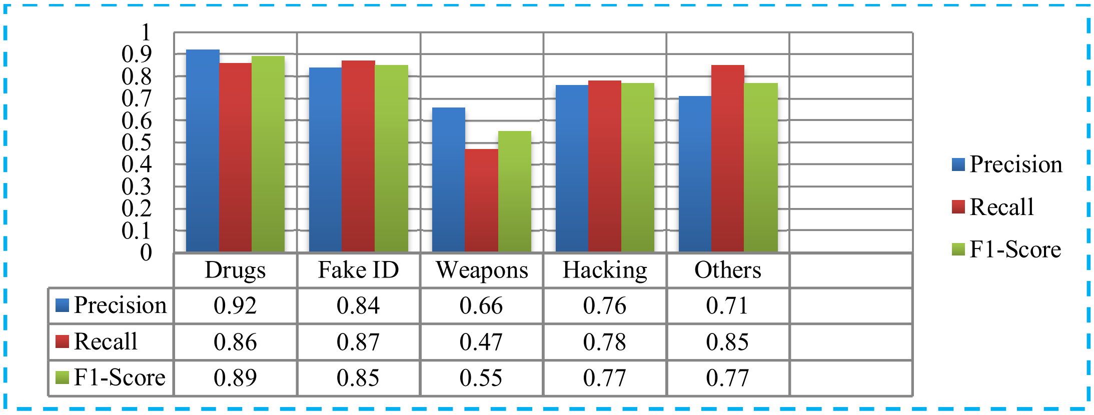
In addition to accuracy, other parameters like precision, recall, and f1-score were used to evaluate the performance of the classifiers.

4.6.1 Naive Bayes Classifier (NB)

Table 4.13 and Figure 4.10 show the performance metrics of the algorithm which was implemented using 5 classes containing 2363 documents. The Naïve Bayes algorithm was used for the classification. At the stage where the classifiers were designed, the names of the classes were converted into numbers (Drugs:1, Fake ID:2, Weapons:3, Hacking:4, and Others:5). There are two axes, the X-axis represents training, and the Y-axis represents the target (the label).

Table (4.13): Evaluation Metrics for NB Classifier.

Method	Classes	Precision	Recall	f1-score
NB Classifier	1	0.92	0.86	0.89
	2	0.84	0.87	0.85
	3	0.66	0.47	0.55
	4	0.76	0.78	0.77
	5	0.71	0.85	0.77
weighted avg.		0.81	0.81	0.81



	Drugs	Fake ID	Weapons	Hacking	Others
🟦 Precision	0.92	0.84	0.66	0.76	0.71
🟥 Recall	0.86	0.87	0.47	0.78	0.85
🟩 F1-Score	0.89	0.85	0.55	0.77	0.77

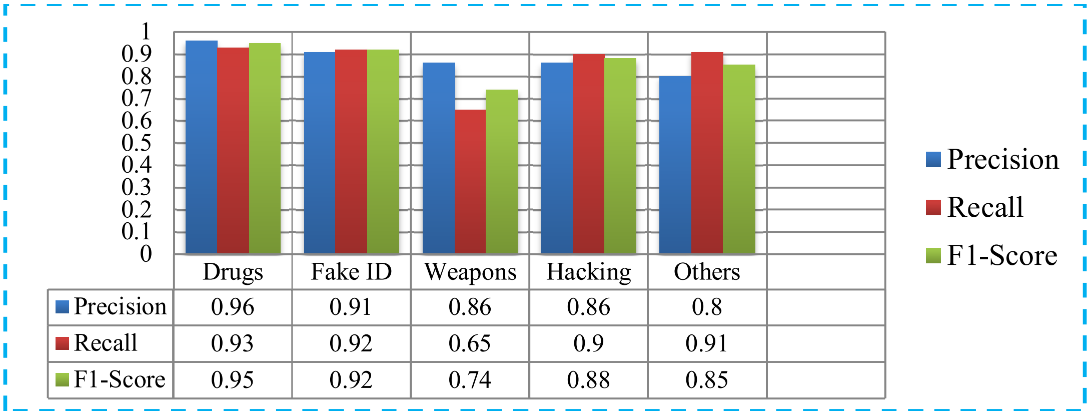
Figure (4.10): The performance metrics of the NB algorithm.

4.6.2 Random Forest Classifier (RF)

Table 4.14 and Figure 4.11 show the performance metrics of the implemented algorithm that there were 5 classes with 2363 documents classified using the RF algorithm.

Table (4.14): Evaluation Metrics for RF Classifier

Method	Classes	Precision	Recall	f1-score
RF Classifier	1	0.96	0.93	0.95
	2	0.91	0.92	0.92
	3	0.86	0.65	0.74
	4	0.86	0.90	0.88
	5	0.80	0.91	0.85
weighted avg.		0.89	0.89	0.89



	Drugs	Fake ID	Weapons	Hacking	Others
🟦 Precision	0.96	0.91	0.86	0.86	0.8
🟥 Recall	0.93	0.92	0.65	0.9	0.91
🟩 F1-Score	0.95	0.92	0.74	0.88	0.85

Figure (4.11): The performance metrics of the RF algorithm.

4.6.3 Linear Support Vector Machine (LSVM)

Table 4.15 and Figure 4.12 show the performance metrics used in evaluating the implemented algorithm, which was implemented on 5 classes containing 2363 documents. The classification was done using the linear SVC algorithm.

Table (4.15): Evaluation Metrics for LSVM.

Method	Classes	Precision	Recall	f1-score
LSVM	1	0.95	0.95	0.95
	2	0.95	0.90	0.92
	3	0.82	0.75	0.78
	4	0.87	0.90	0.88
	5	0.86	0.90	0.88
weighted avg.		0.91	0.91	0.91

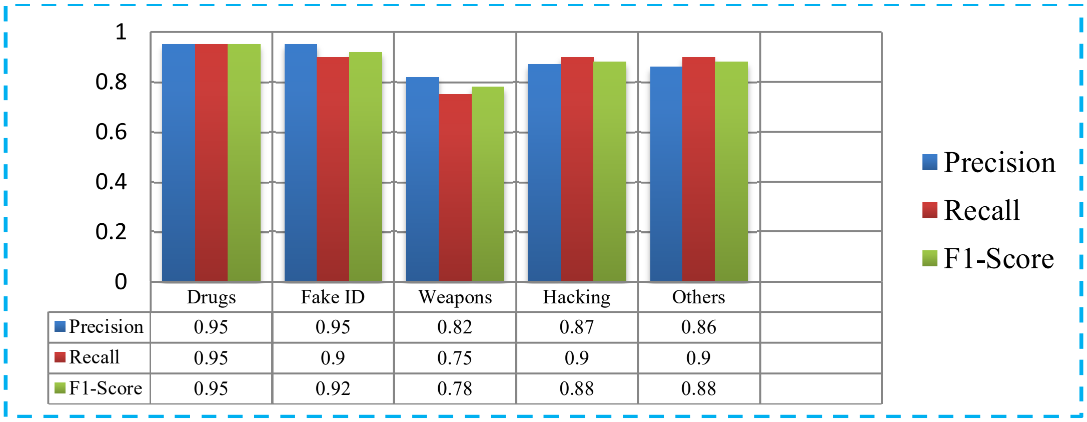


Figure (4.12): The performance metrics of the SVM algorithm.

The Performance Metric of the SVM, RF, and NB shows in figure 4.13.

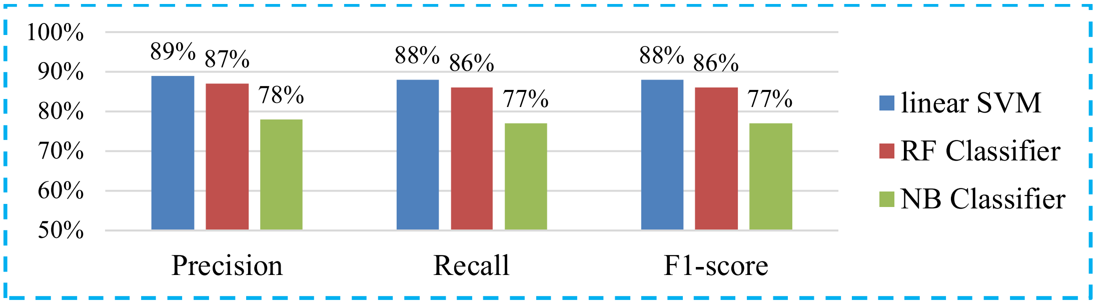


Figure (4.13): The performance metric of the SVM, RF, and NB algorithm.

4.7. Comparison between proposed System with Related Works

Compared to related works, the problem of manually labeling the dataset has been solved in this research, as an automatic labelling algorithm was proposed. To the best knowledge of the researcher, this is the first of its kind. 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 in the dark web, and the proposed model achieved high accuracy. Table (4.16) illustrates the comparison between the proposed system and the related works. Based on the crawling data process, labeling the dataset, and machine learning algorithm that used in classification the dark web and comparison between the system performance and the related work in term of accuracy.

Table (4.16): Comparison between the proposed system and the related works.

No.	**Method**	**Data Collection System**	**Labelling the Dataset**	**Classification**	
				Technique	Accuracy
1	Al Nabki et al. 2017 [5]	Crawler	Manually	LR	96%
2	Siyu He et al. 2019 [6]	Scraper	Manually	NB	93%
3	Khare et al. 2020 [23]	Smart Crawler	Manually	Neural Network	95%
4	The proposed system	Smart Crawler	automatically	SVM	91%

Chapter Five

Conclusions and Future works

Chapter five
Conclusions and Future Works

5.1 Research Summary

The entire activities carried out in this research to ensure its completion are summarized as follows:

- The collection of data lasted for almost a month, begining from early May to early June, 2021. The time frame is considered to be good enough due to the fact that the crawler developed in this research was used in identifying the pages' hyperlinks, thereby reducing the time needed for the search of hyperlinks of the pages.
- In this research, data was subjected to pre-processing using different techniques due to its importance. The pre-processing of data is crucial as it helps in changing the format of the dataset to a more desirable one that can be easily processed by machine learning algorithms. In addition, it purifies the dataset, eliminating any form of noise in it, which in turn improves the accuracy of features extraction.
- Dataset labelling is a critical task that must be performed prior to classification, especially, at the training stage of the system. Thus, the labelling of dataset in this research was carried out using the automatic labeling system created in this research. The labeling algorithm demonstrated high level of accuracy and effectiveness.
- The Linear Support Vector Machine (LSVM) demonstrated superior performance, therefore, being the best algorithm for the classification of dark web pages.
- Based on the evaluation of the performance of the proposed system, which was done based on four parameters (accuracy, recall, precision, and F1-Score), it can be concluded that the system yielded excellent

results. The accuracy is 91 %, precision is 89%, recall is 88% and F1-Score is 88%.

5.2 Future Works

There are many areas in which the presented research can be further explored and improved on, and they are given as follows:

- The use of other methods of classification like K-Nearest Neighbor (KNN), Decision Tree (DT), and Logistic Regression (LR), can be used, given that they are regarded as intelligent methods.
- In this research, the best accuracy rate was recorded for Support Vector Machine. Nevertheless, adaptations can be made to the SVM, also use feature extraction and feature selection techniques to reduce the number of features and choose the important features 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 on a different dataset. More so, improvements can be made to the dataset so that its accuracy can be better.
- The dataset can be further boosted in terms of size by harvesting more HS sources like Freenet and I2P from the dark web, and other ports apart from the HTTP port should be explored. Moreover, we plan to get the benefit of the HTML tags and the hyperlinks by weighting some tags or parsing the hyperlink's text.
- To further ascertain the reliability and detection performance of the proposed system, more experiments will be carried out on a number of new kinds of illicit activities.

References

[1] 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.

[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.

[3] N. Agrawal and S. Johari, “A Survey on Content Based Crawling for Deep and Surface Web,” *Proc. IEEE Int. Conf. Image Inf. Process.*, vol. 2019-Novem, pp. 491–496, 2019.

[4] U. Noor, Z. Rashid, and A. Rauf, “A Survey of Automatic Deep Web Classification Techniques,” *Int. J. Comput. Appl.*, vol. 19, no. 6, pp. 43–50, 2011.

[5] 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.

[6] 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.

[7] S. Ghosh, A. Das, P. Porras, V. Yegneswaran, and A. Gehani, “Automated categorization of onion sites for analyzing the darkweb ecosystem,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. Part F1296, pp. 1793–1802, 2017.

[8] 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.

[9] K. Michael, “White Paper : The Deep Web : Surfacing Hidden ,” Journal of Electronic Publishing Volume 7, Issue 1: August, 2014.

[10] F. Zhao, J. Zhou, C. Nie, H. Huang, and H. Jin, “SmartCrawler: A two-stage crawler for efficiently harvesting deep-web interfaces,” *IEEE Trans. Serv. Comput.*, vol. 9, no. 4, pp. 608–620, 2016.

[11] W. Mhd, Supervised Machine Learning For Classification, Mining, And Ranking Of Illegal Web Contents, doctoral diss., University Of León Department Of Electrical, Systems And Automatic Engineering, 2019.

[12] Y. Wu *et al.*, *Python Scrapers for Scraping Cryptomarkets on Tor*, vol. 11611 LNCS, no. October 2013. Springer International Publishing, 2019.

[13] M. Chertoff, “A public policy perspective of the Dark Web,” *J. Cyber Policy*, vol. 2, no. 1, pp. 26–38, 2017.

[14] M. B. Box, “the Deep Web and the Darknet : a Look Inside the Internet ’ S,” no. October, 2015.

[15] D. Moore and T. Rid, “Cryptopolitik and the darknet,” *Survival (Lond).*, vol. 58, no. 1, pp. 7–38, 2016.

[16] B. AlKhatib and R. Basheer, “Crawling the Dark Web: A Conceptual Perspective, Challenges and Implementation,” *J. Digit. Inf. Manag.*, vol. 17, no. 2, p. 51, 2019.

[17] P. Kaur, “Web Content Classification: A Survey,” *Int. J. Comput. Trends Technol.*, vol. 10, no. 2, pp. 97–101, 2014.

[18] M. Graczyk and K. Kinningham, “Automatic Product Categorization for Anonymous Marketplaces,” *Comput. Sci.*, pp. 1–6, 2015.

[19] 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.

[20] 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.

[21] E. Marin, M. Almukaynizi, E. Nunes, and P. Shakarian, “Community finding of malware and exploit vendors on darkweb marketplaces,” *Proc. - 2018 1st Int. Conf. Data Intell. Secur. ICDIS 2018*, pp. 81–84, 2018.

[22] M. Pannu, I. Kay, and D. Harris, *Using Dark Web Crawler to Uncover Suspicious and Malicious Websites*, vol. 782. Springer International Publishing, 2019.

[23] 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.

[24] V. V. Mahale, M. T. Dhande, and A. V. Pandit, “Advanced web crawler for deep web interface using binary vector page rank,” *Proc. Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), I-SMAC 2018*, pp. 500–503, 2019.

[25] B. Hawkins, “Information Security Reading Room Under The Ocean of the Internet - The Deep Web,SANS institute information security” 2020.

[26] S. Kaur and S. Randhawa, “Dark Web : A Web of Crimes,” *Wirel. Pers. Commun.*, no. 0123456789, 2020.

[27] 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.

[28] M. F. Bin Rafiuddin, H. Minhas, and P. S. Dhubb, “A dark web story in-depth research and study conducted on the dark web based on forensic computing and security in Malaysia,” *IEEE Int. Conf. Power, Control. Signals Instrum. Eng. ICPCSI 2017*, pp. 3049–3055, 2018.

[29] A. Montieri, D. Ciuonzo, G. Bovenzi, V. Persico, and A. Pescape, “A Dive into the Dark Web: Hierarchical Traffic Classification of Anonymity Tools,” *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, pp. 1043–1054, 2020.

[30] I. Karunanayake, N. Ahmed, R. Malaney, R. Islam, and S. Jha, “Anonymity with Tor: A Survey on Tor Attacks,” 2020.

[31] E. Nunes *et al.*, “Darknet and deepnet mining for proactive cybersecurity threat intelligence,” *IEEE Int. Conf. Intell. Secur. Informatics Cybersecurity Big Data, ISI 2016*, pp. 7–12, 2016.

[32] A. Elgzil, C. E. Chow, A. Aljaedi, and N. Alamri, “Cyber anonymity based on software-defined networking and Onion Routing (SOR),” *2017 IEEE Conf. Dependable Secur. Comput.*, pp. 358–365, 2017.

[33] A. T. Zulkarnine, R. Frank, B. Monk, J. Mitchell, and G. Davies, “Surfacing collaborated networks in dark web to find illicit and criminal content,” *IEEE Int. Conf. Intell. Secur. Informatics Cybersecurity Big Data, ISI 2016*, pp. 109–114, 2016.

[34] H. YU, Y. YANG, L. YANG, and G. ZHU, “Dark Web Threat Intelligence and Market Analysis,” *DEStech Trans. Environ. Energy Earth Sci.*, no. iccis, pp. 470–477, 2019.

[35] M. Manke, K. K. Singh, V. Tak, and A. Kharade, “Crawdy : Integrated crawling system for deep web crawling,” vol. 4, no. 9, pp. 389–393, 2015.

[36] D. Eldhose T John, B. Skaria, and P. X. Shajan, “An Overview of Web Content Mining Tools,” *Bonfring Int. J. Data Min.*, vol. 6, no. 1, pp. 01–03, 2016.

[37] H. Chen, X. Li, M. Chau, Y. J. Ho, and C. Tseng, “Using open web APIs in teaching web mining,” *IEEE Trans. Educ.*, vol. 52, no. 4, pp. 482–490, 2009.

[38] P. Kolari and A. Joshi, “Web mining: Research and practice,” *Comput. Sci. Eng.*, vol. 6, no. 4, pp. 49–53, 2004.

[39] K. Jayamalini and M. Ponnavaikko, “Research on web data mining concepts, techniques and applications,” *2017 Int. Conf. Algorithms, Methodol. Model. Appl. Emerg. Technol. ICAMMAET 2017*, vol. 2017-Janua, pp. 1–5, 2017.

[40] S. A. Alasadi and W. S. Bhaya, “Review of data preprocessing techniques in data mining,” *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017.

[41] E. Haddi, X. Liu, and Y. Shi, “The role of text pre-processing in sentiment analysis,” *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013.

[42] A. Jović, K. Brkić, and N. Bogunović, “A review of feature selection methods with applications,” *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.*, pp. 1200–1205, 2015.

[43] P. C. Gaigole, L. H. Patil, and P. M. Chaudhari, “Preprocessing Techniques in Text Categorization,” *Natl. Conf. Innov. Paradig. Eng. Technol.*, pp. 1–3, 2013.

[44] A. Krouska, C. Troussas, and M. Virvou, “The effect of preprocessing techniques on Twitter sentiment analysis,” *IISA 2016 - 7th Int. Conf. Information, Intell. Syst. Appl.*, 2016.

[45] S. Sun, C. Luo, and J. Chen, A review of natural language processing techniques for opinion mining systems, *Information Fusion*, vol. 36. 2017.

[46] V. Srividhya and R. Anitha, “Evaluating Preprocessing Techniques in Text Categorization,” *Int. J. Comput. Sci. Appl.*, pp. 49–51, 2010.

[47] S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, “Preprocessing Techniques for Text Mining -An Overview,” vol. 5, no. 1, pp. 7–16.

[48] P. Tan and M. Steinbach, “Introduction to Data Mining Instructor ’ s Solution Manual,” *Names*, vol. 28, no. 1, pp. 9–35, v, 2006.

[49] P. Ristoski and H. Paulheim, “Semantic Web in data mining and knowledge discovery: A comprehensive survey,” *J. Web Semant.*, vol. 36, pp. 1–22, 2016.

[50] S. Beniwal and J. Arora, “Classification and Feature Selection Techniques in Data Mining,” *Int. J. Eng. Res. Technol.*, vol. 1, no. 6, pp. 1–6, 2012.

[51] A. Sharma, R. Sharma, V. Sharma, and V. Shrivatava, “Application of Data Mining– A Survey Paper,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 2023–2025, 2014.

[52] A. Jafar Hamid and T. M. Ahmed, “Developing Prediction Model of Loan Risk in Banks Using Data Mining,” *Mach. Learn. Appl. An Int. J.*, vol. 3, no. 1, pp. 1–9, 2016.

[53] V. Gupta and G. S. Lehal, “A survey of text mining techniques and applications,” *J. Emerg. Technol. Web Intell.*, vol. 1, no. 1, pp. 60–76, 2009.

[54] B. Baharudin, L. H. Lee, and K. Khan, “A Review of Machine Learning Algorithms for Text-Documents Classification,” *J. Adv. Inf. Technol.*, vol. 1, no. 1, 2010.

[55] E. K. M. Al-yasiri, “Arabic Sentiment Analysis for Identifying Terrorism Supporters on Twitter Using Data Mining Techniques,” master diss., University of Babylon, College of Information Technology, Software Department, 2019.

[56] Y. Lin and J. Wang, “Research on text classification based on SVM-KNN,” *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, pp. 842–844, 2014.

[57] F. Colas and P. Brazdil, “On the behavior of SVM and some older algorithms in binary text classification tasks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4188 LNCS, pp. 45–52, 2006.

[58] R. Inglehart, “Chapter 10. From Elite-Directed To Elite-Directing Politics: The Role Of Cognitive Mobilization, Changing Gender Roles, And Changing Values,” *Cult. Shift Adv. Ind. Soc.*, pp. 335–370, 2019.

[59] E. K. Al-Yasiri and A. Al-Azawei, “Improving Arabic sentiment analysis on social media: A comparative study on applying different pre-processing techniques,” Compusoft, vol. 8, no. 6, pp. 3150–3157, 2019.

[60] B. Luo, Q. Zhang, and S. D. Mohanty, “Data-Driven Exploration of Factors Affecting Federal Student Loan Repayment,” 2018.

[61] A. L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, “Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 6, pp. 493–507, 2012.

[62] J. Kazmierska and J. Malicki, “Application of the Naïve Bayesian Classifier to optimize treatment decisions,” *Radiother. Oncol.*, vol. 86, no. 2, pp. 211–216, 2008.

[63] W. Ding, S. Yu, Q. Wang, J. Yu, and Q. Guo, “A novel Naive Bayesian text classifier,” *Proc. - Int. Symp. Inf. Process. ISIP 2008 Int. Pacific Work. Web Min. Web-Based Appl. WMWA 2008*, pp. 78–82, 2008.

[64] S. Karthika and N. Sairam, “A Naïve Bayesian classifier for educational qualification,” *Indian J. Sci. Technol.*, vol. 8, no. 16, 2015.

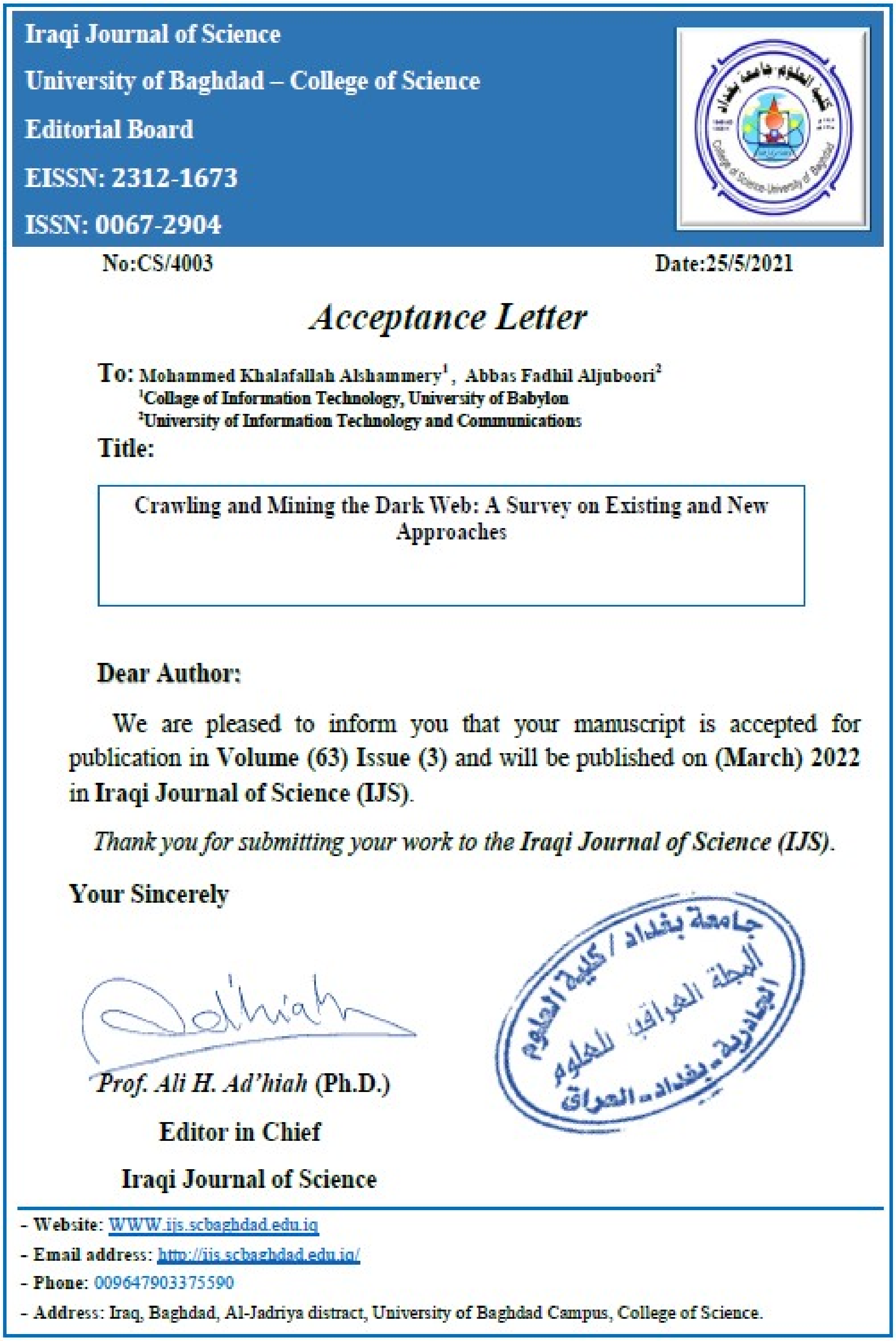
[65] H. Adel and M. Bayati, “Building Bi-lingual Anti-Spam SMS Filter,” *Int. J. New Technol. Res.*, vol. 4, no. 1, p. 263147, 2018.

[66] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015.

[67] M. A. Al_Masoudy, “Predicting Learners’ Performance on Virtual Learning Environment (VLE) based on their Behavior Using Data Mining Techniques,” master diss., University of Babylon, College of Information Technology, Software Department, 2020.

Appendix A

The Published Paper 1



Iraqi Journal of Science

University of Baghdad – College of Science

Editorial Board

EISSN: 2312-1673

ISSN: 0067-2904

No:CS/4003 Date:25/5/2021

Acceptance Letter

To: Mohammed Khalafallah Alshammery[1] , Abbas Fadhil Aljuboori[2]

[1]Collage of Information Technology, University of Babylon

[2]University of Information Technology and Communications

Title:

Crawling and Mining the Dark Web: A Survey on Existing and New Approaches

Dear Author:

We are pleased to inform you that your manuscript is accepted for publication in **Volume (63) Issue (3)** and will be published on **(March) 2022** in **Iraqi Journal of Science (IJS)**.

*Thank you for submitting your work to the **Iraqi Journal of Science (IJS)**.*

Your Sincerely

جامعة بغداد / كلية العلوم
المجلة العراقية للعلوم
الجادرية - بغداد - العراق

Prof. Ali H. Ad'hiah (Ph.D.)

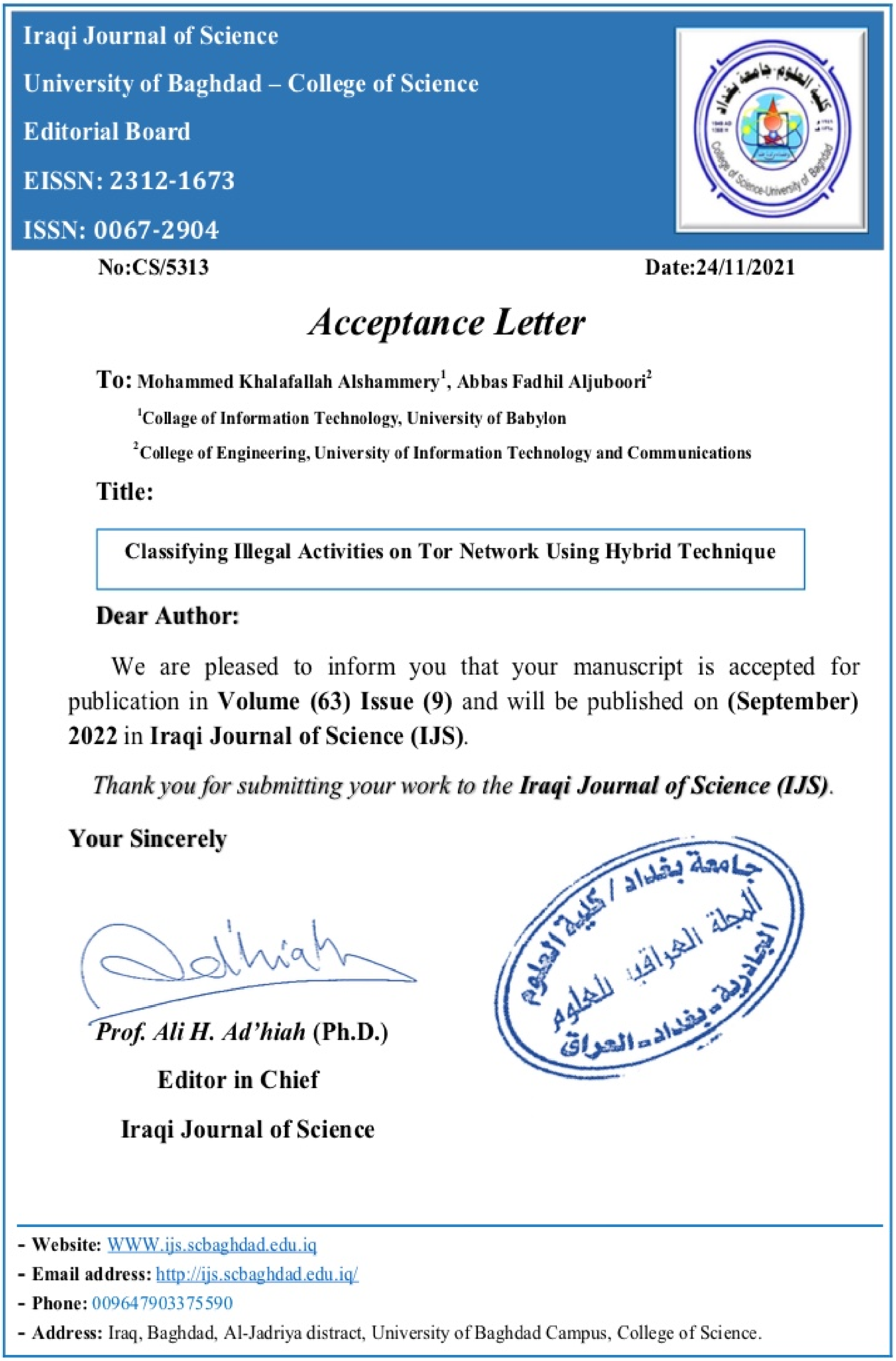
Editor in Chief

Iraqi Journal of Science

- Website: WWW.ijs.scbaghdad.edu.iq
- Email address: http://ijs.scbaghdad.edu.iq/
- Phone: 009647903375590
- Address: Iraq, Baghdad, Al-Jadriya distract, University of Baghdad Campus, College of Science.

Appendix B

The Published Paper 2



Iraqi Journal of Science
University of Baghdad – College of Science
Editorial Board
EISSN: 2312-1673
ISSN: 0067-2904

No:CS/5313 Date:24/11/2021

Acceptance Letter

To: Mohammed Khalafallah Alshammery[1], Abbas Fadhil Aljuboori[2]

[1]Collage of Information Technology, University of Babylon

[2]College of Engineering, University of Information Technology and Communications

Title:

Classifying Illegal Activities on Tor Network Using Hybrid Technique

Dear Author:

We are pleased to inform you that your manuscript is accepted for publication in **Volume (63) Issue (9)** and will be published on **(September) 2022** in **Iraqi Journal of Science (IJS)**.

Thank you for submitting your work to the ***Iraqi Journal of Science (IJS)***.

Your Sincerely

Prof. Ali H. Ad'hiah (Ph.D.)

Editor in Chief

Iraqi Journal of Science

جامعة بغداد / كلية العلوم
المجلة العراقية للعلوم
الجادرية - بغداد - العراق

- **Website:** WWW.ijs.scbaghdad.edu.iq
- **Email address:** http://ijs.scbaghdad.edu.iq/
- **Phone:** 009647903375590
- **Address:** Iraq, Baghdad, Al-Jadriya distract, University of Baghdad Campus, College of Science.

الخلاصة

توفر آليات إخفاء الهوية القوية والآليات التي يصعب تتبعها لشبكة الويب المظلمة المأوى للأنشطة غير القانونية. المحتوى غير القانوني على الويب المظلم متنوع ويتم تحديثه بشكل متكرر. يستخدم تصنيف الويب المظلم التقليدي صفحات ويب واسعة النطاق للتدريب الخاضع للإشراف. ومع ذلك ، أصبحت صعوبة جمع ما يكفي من محتوى الويب المظلم غير القانوني واستهلاك الوقت في وضع العلامات يدويًا على صفحات الويب من التحديات التي تواجه البحث الحالي.

يتكون النظام المقترح في هذا البحث من اربعة مراحل رئيسية. في المرحلة الأولى ، يتم جمع البيانات. في هذه الدراسة ، تم تقديم مجموعة بيانات جديدة تُعرف باسم "قاعدة بيانات الويب المظلم" للنطاقات النشطة على شبكة الويب المظلمة. تم إنشاؤها باستخدام نظام قوي لديه القدرة على اختراق الشبكة السوداء وجمع البيانات من صفحاتها ، والاستكشاف والتنقل عبر الارتباطات التشعبية في وقت واحد. لتحقيق ذلك ، تم أخذ عينات من شبكة (تور) كعناوين اولية ومن خلال هذه العنوان يستخرج النظام كل الروابط التشعبية والسير عليها واستخراج بياناتها. تتضمن المرحلة الثانية المعالجة المسبقة للبيانات باستخدام مجموعة متنوعة من تقنيات المعالجة المسبقة. تأتي هذه المرحلة مباشرة بعد كل عملية اسخراج بيانات ناجحة ينفذها نظامنا في الشبكة السوداء، وبعد ذلك ، يتم تخزين النتائج في قاعدة البيانات. تضمنت المرحلة الثالثة تم انشاء خوارزمية تعرف باسم خوارزمية وضع العلامات التلقائية ، والتي تم تطبيقها على مجموعة البيانات، وتم عنونة كل رابط ويب مظلم حسب المحتوى تلقائيًا الى احدى الفئات الخمس المستخدمة في بحثنا. أظهرت النتائج أن الخوارزمية حققت معدل دقة بنسبة ٨٥٪. في المرحلة الرابعة، وهي المرحلة الأخيرة ، تم استخدام مصنف ناقل الدعم الخطي (SVM) ، ومصنف Random Forest (RF)، ومصنف Naive Bayes (NB) لتطبيق طريقة التصنيف.

تم تنفيذ هذه التقنيات لإلقاء الضوء على أكثر الأساليب المستخدمة دقةً في تصنيف بيانات الويب المظلم. علاوة على ذلك ، تم تقييم نتائج خوارزميات التصنيف بناءً على مقياس الأداء باستخدام قياسات الدقة و precisionو recall و F1-score. أظهرت النتائج أن أعلى دقة تم تحقيقها من خلال تطبيق LSVM. حيث بلغت الدقة ٩١٪ ، وF1-score هي ٨٨٪. 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.

جمهوريـــة العـــراق
وزارة التعليم العالي والبحث العلمي
جامعـــة بابـــل
كلية كنولوجيا المعلومات - قسم البرمجيات

تصميم نظام متكامل للزحف وتصنيف الشبكة المظلمة

رسالة مقدمة إلى
مجلس كلية تكنولوجيا المعلومات - جامعة بابل كجزء من متطلبات
نيل درجة الماجستير في تكنولوجيا المعلومات / البرمجيات

من قِبَل

محمد خلف الله محمد نزال

بإشراف

الاستاذ الدكتور عباس فاضل الجبوري

١٤٤٣هـ **٢٠٢٢م**