IDENTIFICATION AND CLASSIFICATION OF ILLEGAL CONTENT ON TOR DARK WEB

Thesis submitted in fulfilment of the requirement for the degree of

Doctor of Philosophy

IN

INFORMATION TECHNOLOGY

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
LUCKNOW
प्रज्ञा शील करूणा
ESTABLISHED 1996

Submitted by

Mohd Faizan

Supervised by

Prof. Raees Ahmad Khan

Submitted to

DEPARTMENT OF INFORMATION TECHNOLOGY
BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
(A CENTRAL UNIVERSITY)
VIDYA VIHAR, RAEBARELI ROAD,
LUCKNOW – 226025, UTTAR PRADESH, INDIA

JULY-2021

DECLARATION

I, **Mohd Faizan**, solemnly declare that this thesis of research on **"Identification and Classification of Illegal Content on Tor Dark Web"** is my original work. The study has been conducted under the guidance of **Prof. Raees Ahmad Khan**, at Department of Information Technology, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow. It is further declared that to the best of my knowledge and belief it has not been submitted earlier for the award of any degree. I also undertake that the thesis is essentially free from all kinds of plagiarism.

Dated: 29-07-2021

(Mohd Faizan)
Researcher
Department of Information Technology
Babasaheb Bhimrao Ambedkar University
(A Central University)
Lucknow, Uttar Pradesh, India

CERTIFICATE

This is to certify that the thesis titled **"Identification and Classification of Illegal Content on Tor Dark Web"** submitted by **Mr. Mohd Faizan** is an original research work and has not been previously submitted in part or full for the award of any other degree or diploma to this or any other University.

This thesis submitted to Babasaheb Bhimrao Ambedkar University, Lucknow satisfies all the requirements as stipulated in the Doctor of Philosophy (Ph.D.) regulations-1999 as amended in 2008/2010/2013 and it is fit for submission and evaluation for the award of the degree of Doctor of Philosophy of the University.

Supervisor

Date: 29-07-2021

Head of the Department

HEAD
Department of Information Technology
School For Information Science & Technology
Babasaheb Bhimrao Ambedkar University
Lucknow

ACKNOWLEDGEMENT

The writing of this thesis has been an incredible journey and a monumental milestone in my academic life. I could not have embarked on this expedition and travelled this far without the passionate and continued support of supervisors, colleagues, friends, and family. Working as a Ph.D. scholar at Babasaheb Bhimrao Ambedkar University was a magnificent as well as challenging experience to me. In all these years, many people were instrumental directly or indirectly in shaping up my academic career. It was hardly possible for me to thrive in my research work without the precious support of these people. Here is a small gratitude to all those people.

I am indeed indebted to my distinguished guide and supervisor **Prof. Raees Ahmad Khan** for all his help during the course of the study. Certainly, I am short of words to convey my real feelings for his invaluable help and concern together with 'scholarly' insightful and 'critical' guidance. I do not hesitate to state that without his help it would not have been possible for me to complete the research work.

I must thank to **Prof. Raees Ahmad Khan** because despite of his tight time-schedule, he was always available to me all times, to answer my queries, discussion of matters, helped me to learn from my mistakes, without even forcing his opinion on me. I spent countless hours with him for discussing research, writing research papers, proof reading of research manuscripts, making this thesis and talking about my life in general. Moreover, being the Head of the Department, Prof. Khan gave me the opportunity to pursue research work at the Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow. I received lots of motivation, encouragement, and support from him during my entire duration.

I feel deep sense of gratitude for my parents who formed a base of my vision and taught me the good things that really matters in my life. In addition, I am also thankful to **Dr. Dhirendra Pandey** and **Dr. Alka**, for their continuous encouragement, guidance, moral support, and consultations during the course of the study. I am also thankful to all my friends and colleagues for providing encouragement and support especially Dr. Mohd Waris Khan, Manish Joshi and Dr. Virendra Singh. I express my sincere thanks to all the faculty members and office staff of the department for their time-to-time continuous encouragement and support. I express my sincere thanks to all the experts from India and abroad for providing me with their valuable observations during the peer review of my research papers. In summary, I would like to thank everyone for putting up with me for the entire period of my research work.

Mohd Faizan

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Background

The World Wide Web (WWW) has completed three decades since it was first proposed by Tim Berners-Lee. Now, the number of active websites on the Internet has crossed the billion mark and is growing rapidly. The Internet has become an inevitable part of our lives. It keeps people connected to their dear ones globally with the help of social media at the comfort of their homes.

We often resort to the Internet to get abreast with the live news and weather updates. Be it some knowledge gathering or entertainment, Wikipedia and YouTube are always there as a savior. The increasing dependence on the Internet for online shopping driven by the ease of access has created a boom in the electronic commerce sector so much so that its global worth is in trillion US dollars. The Internet currently has about 4.6 billion[1] users thereby reaching nearly to more than half of the planet's population.

1.1.1 Dark Web: Looking Beyond the World Wide Web

Internet and the World Wide Web (WWW) despite considered synonymous are two different things. The web is just a part of the Internet for sharing and accessing information. The websites that we search for the information on popular search engine like Google constitutes only a portion of the web called the "Surface Web". Beyond the easily available surface web lies another much larger layer of content that is not searchable through Google, Yahoo or any other traditional search engine. This layer of the Internet is called the "Deep Web" [1].

[1] https://www.statista.com/

The content on the deep web is not indexed by the search engines and thus it is not openly available for the general users. The content available on the intranet of government and corporate offices, academic institutions, databases of financial entities like the banks and insurance companies, dynamically generated content of websites through user filled forms and queries etc constitutes the deep web. The deep web content is not indexed for various reasons to ensure security, privacy and confidentiality of users data. Sometimes the content is not relevant to every user on the Internet hence it is not indexed and is accessible only through the deep web.

The deepest layer of the Internet is known as the "Dark Web" or the "Darknet" [2]. The websites on the dark web are hidden in such a way that it requires a special browser and mechanism to access them. Unlike the surface and deep web, the dark web utilizes the Onion Routing [3] technique to establish the connection between the user and the website. The dark websites can only be accessed via the Tor browser that ensures the anonymity of the users while surfing the dark web. Given the level of anonymity provided by the dark web platforms, the users may leverage this to perform several legal and illegal tasks. A comparison of the layers of the Internet is shown in Table 1.1.

Table 1.1: An Overview of the Web Content Layers.

Web Content	Accessibility	Indexing	Users
Surface Web	Regular Web Browser (like Chrome, Firefox, Safari)	Indexed by search engines	General user
Deep Web	Regular Web Browser (like Chrome, Firefox, Safari)	Not Indexed by search engines	Users who have permissions
Dark Web	Special Browser (like Tor browser)	Not Indexed by search engines	Users who want to remain anonymous

Accessing the Dark Web: The Onion Routing (Tor)

The dark web can be reached via multiple methods using decentralized and anonymous hops. Some of the methods include The Onion Router (Tor), the Invisible Internet Project (I2P)[2] and the Freenet[3]. Among all the tools available, the Tor is the most commonly used for accessing the dark web [4]. The Tor was initially developed by the United States Naval Research Laboratory as The Onion Routing Project in 2002 for the secret communication of electronic messages.

The idea behind the development of the onion routing technique was to conceal the communication between the two users. It aims to provide the utmost privacy and anonymity while using the Internet [5]. In onion routing, whenever communication is made between the client and the server, instead of a direct connection, a circuit of three random nodes is created between the two ends. These nodes are called the relays and each of the relays only knows about the relay it receives the data from and the relay to which it sends the data. All the relays in the circuit are unaware of the complete path taken by the data packet. The data is encrypted for each of the relays with separate encryption keys before reaching the destination.

The data packet from the sender with the three layers of encryption is passed on to the first node of the circuit called the guard node. The first layer of the data packet is decrypted by the entry node with its key. The entry node only knows about the sender. The packet is then transferred to the middle node which decrypts the second layer of encryption. The middle node does not have any information about the sender. Finally, the data packet is moved to the last node called the exit node which removes the last layer of encryption.

[2] https://geti2p.net/
[3] https://freenetproject.org/

The exit nodes can see the original data received from the sender but it does not know who the sender is. The data packet is then forwarded to its actual destination over the open Internet. The destination only has the information of the exit node and is unaware of the actual source of the packet. The destination can use the same connection to send the response to the actual source. Figure 1.1 shows the graphical representation of the onion routing and how it differs from the regular routing technique.

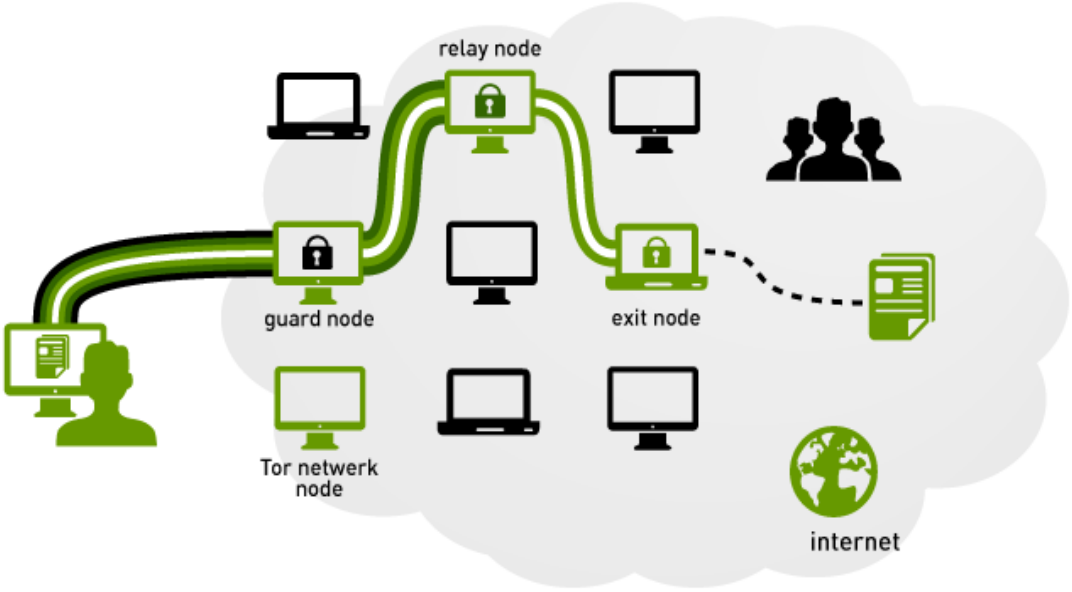

Figure 1.1: The Onion Routing Technique.

The Tor Browser

The Onion Router browser or the Tor browser is similar to the conventional Internet browsers like Google Chrome or the Internet Explorer except that it is modified to keep its user identity anonymous. The Tor browser disguises the original Internet Protocol (IP) address of its users in an attempt to provide anonymity while surfing the web. The Tor browser is available for download from the official website of the Tor Project[4]. It comes both as a standalone application bundle in the form of a modified Firefox browser or an embedded operating system package.

[4] https://torproject.org/

As per the Tor Project Inc., the Tor browser is being used to safeguard the privacy of the users while accessing the Internet. It may be beneficial to access the websites that are blocked in oppressive regimes, it also allows the residents of oppressive regimes to securely communicate over the Internet without the fear of censorship. Since the Tor browser employs the onion routing technique to conceal the connections, it can effectively be used to access the dark web. The websites on the dark web that are accessible through Tor are commonly known as the hidden services.

Hidden Services

The Tor allows setting up the servers that are hard to trace and can only be accessed through the Tor browser. It enables to host any type of content on the servers bypassing all forms of restrictions and monitoring. The actual location of these servers is unknown to the Internet Service Providers (ISPs), law enforcement agencies and even to the Tor project. These servers are commonly referred to as the Tor hidden services which constitutes the Tor dark web. Unlike the surface web, the web address of the hidden services consists of sixteen alphanumeric characters which are difficult to understand (for example, *uzxmoz3g7gsfvita.onion*) and ends with the top-level domain of *.onion*. The .onion domains do not map to an IP address and hence could not be traced anywhere.

There is no database (like the Domain Name System (DNS) on the surface web) of the hidden services that contain the onion addresses. Moreover, the hidden services are not indexed by traditional search engines like Google or Yahoo. The sites like *The Hidden Wiki*[5] provides some of the onion addresses on the dark web. Once the address of the hidden service is known, an encrypted connection is established both by the client and the server to an intermediary node called the rendezvous

[5] https://thehiddenwiki.org/

node that acts as a bridge between the two ends for the communication. In this way, the anonymity of both the client and the server is ensured.

1.1.2 Need for the Dark Web

The dark web may be used for many reasons by people to hide their online presence. It was primarily developed to ensure the privacy of communications in a variety of domains. The use of Tor has been extensively promoted to protect the freedom of speech, online privacy and security [6]. Some of the common purposes of the use of the anonymous platform are discussed below.

Secret Communication: The anonymous platform may be utilized by individuals or groups of people for discussing sensitive information on private chat services and online forums. The users may leverage this platform to discuss their private matters such as serious diseases or bullying or injustice they have faced. The corporate environment may also use this to discuss their business strategies and plans.

Whistleblowing: Whistleblowing is the act of disclosing government or organization confidential information in the public domain. Though the sharing of private government information to the public has been advocated as the right of the public [7], it is considered a treacherous act in many countries. Moreover, the disclosure of an organization or company's private information also falls under illegal acts in countries like the United States.

The dark web is a safe den for journalists, whistleblowers to connect with the general public to disseminate secret information regarding the government or an organization. They can secretly share classified documents and texts with other individuals. The prime example of whistleblowing on Tor was that of Edward Snowden. Snowden

reportedly used Tor to connect with the journalists to share some of the most secret information of the United States (US) security forces [8]. He exposed the mass surveillance program PRISM of the US that attempts to reveal the identity of people who use anonymizing tools like the Tor [9].

Censorship Evasion: Tor can be used by the citizens of the oppressive government to evade censorship on the blocked content [10]. They can securely access the websites that are banned from use by the government. People can express their political views, dissents with their governments without any action [11]. It is beneficial in the countries with strict monitoring and government-sponsored censorship against journalists, protestors and political activists. For example, Tor has enabled the journalist to evade government censorship and let the world know about the situation when the riots were happening in Egypt [12]. The political activists may utilize Tor in the repressive governments for secret communications with other individuals, advocating social reforms, etc.

1.2 Illegal Activities on the Dark Web

The use of the dark web is not limited only to the good purpose but is also used by bad actors for unethical and illegal activities. The dark web is infamous for being a safe ground for several illegal activities [13]. Some of the activities are illicit drug trafficking, firearm trafficking, fake documents and counterfeit currencies. Other disgusting things include child abuse content, violence and human experiments [14]. There have also been reports on the usage of dark web platforms by terrorist and extremist groups to spread their propaganda [15].

Malware has used the Tor network to sniff the victims off their private and confidential information [16]. Ransomware has also been found to be originating from the dark web where they take full control over the infected system by encrypting user's data and demand ransom in

the form of cryptocurrency for decryption [17]. The term dark web would refer to the Tor dark web unless specifically mentioned. A brief description of some of the illicit acts on the dark web is given in Table 1.2.

Table 1.2: Some of the illegal activities on the Dark Web and their description.

Activity	Description
Drugs	Prescription drugs like stimulants, depressants, narcotics, etc are offered illegally
Adult Content	Extreme pornographic content like bestiality, amputee porn and child abuse content
Firearms and Ammunitions	Illegal trafficking of firearms, ammunition, explosives and other defensive gears
Credit Card Dumps	Details of cloned and stolen credit cards, PayPal and bank accounts
Counterfeits	Fake currencies like the US Dollars, Euros and GB Pounds
Hitman	Contract killers to assassinate anyone on the planet (as claimed)
Human Experiment	Visual content depicting victims being mutilated
Gore	Graphic content of medical procedures, agony and bloodshed
Fake Documents	Forged documents like passport, academic degrees, citizenship documents etc
Wildlife Products	Products of endangered and vulnerable species (flora and fauna) like rhino horn, elephant ivory
Digital Material	Ebooks and guides on unethical hacking, 3D printing of firearms, DIY explosive preparation

The major hubs of illegal activities on the dark web are cryptomarkets or darknet marketplaces or simply marketplaces. Cryptomarkets are virtual markets that sell a range of products in a single place. Many of the products and services discussed in the Table 1.2 can easily be found on a single marketplace.

1.2.1 Cryptomarkets

Cryptomarkets are hidden services that deal in the trade of illegal and controversial goods and services. Some of the goods include illegal firearms, stolen products, counterfeits, credit card dumps, pirated software among others with drugs being the most popular product [18]. The cryptomarkets operate similarly to Amazon or eBay on the surface web except that the users are allowed to remain anonymous while completing transactions on the dark web cryptomarkets.

The most notable example of the cryptomarkets is the *Silk Road* marketplace that appeared in the year 2011. The rapidly growing marketplace had generated a revenue of around 15 million US dollars in the year 2012 bringing it under the lens of the law enforcement agencies [18]. In October 2013, Silk Road was shut down by the Federal Bureau of Investigation (FBI) followed by the arrest of its administrator. However, Silk Road version 2 was soon launched under a different domain. Moreover, the extensive popularity of the Silk Road leads to the emergence of several other cryptomarkets with a large number of vendors offering a multitude of both legal and illegal products [19].

Exploring the Cryptomarkets

An individual can explore the products available on the cryptomarkets by creating a user account [18]. Vendors advertise their products on the cryptomarkets through their respective vendor accounts. Each of the vendors has their profile information that shows their popularity, reputation, products they offer and their ratings. The products available online contain their vendors and price. It may also include the shipping location of the vendor and a review of the products purchased by the customers from that vendor. The majority of the products on the Silk Road goes out of the stock within three weeks and one-fourth of them were in stock for not more than three days [18].

The product may be available as an open listing or through a direct invitation link intended to be displayed only to specific customers. Such types of listings are only provided to users which are trusted by the vendors with positive transaction history. Many times, a user can directly contact the vendors through the messaging service to enquire about specific products of their interest. Once a product has been finalized by the user, it can proceed to buy either through the escrow service or by directly transferring the amount to the vendor.

Products Offered on Cryptomarkets

Cryptomarkets are known to offer a multitude of products including both physical and digital items. The users can comfortably go through the complete product catalog available under one roof. The marketplace usually has a similar layout displaying main product categories and the sub-categories. The individual product pages display the product specifications in detail along with the vendor description. A user can add its desired product into shopping cart of the marketplace and proceeds to checkout to confirm the order.

Among all the products available on the dark web, drugs were the most popular [18]. The type of drugs include some of the most harmful drugs like *heroin, crack cocaine, fentanyl* etc. Party drugs like *ecstasy* and *LSD* pills are also popular. The catalog of the Silk Road contains more than 200 distinct categories of items. Other products include digital items like e-books, hacking guides, software, hacking tools, credit card dumps, fake documents etc.

1.2.2 Cryptocurrencies

To complete the transaction on the cryptomarkets, users must have cryptocurrencies. *Bitcoin* is predominantly the preferred cryptocurrency followed by others like *Monero, Etherium, Litecoin* etc. Cryptocurrencies

are digital currency that employs cryptographic techniques for the execution of a transaction. Contrary to regular banking currency, cryptocurrencies are decentralized where each transaction is authorized by all the involved participants instead of a central nodal authority. This also keeps the Bitcoin transaction away from the control of any central monitoring agency [20].

For example, Bitcoin uses blockchain technology that maintains the historical record of all the Bitcoin transactions that ever happened and is publicly available. Whenever a new transaction has to be completed, it is added to the blockchain and has to be approved by all the participating entities. Bitcoins can be bought using platforms similar to that of online Internet banking. As per the writing of this thesis, there was near about 18.6 million Bitcoin currently in circulation in the market. The price of Bitcoin is highly unpredictable and varies irregularly. Moreover, the mining of Bitcoin requires high computing power.

The individual user can create online wallets to acquire Bitcoin and use it to perform any transaction. Once the user has a sufficient cryptocurrency in the account, he/she can begin to initiate the transaction on the cryptomarket.

Shipping of Goods

Cryptomarkets offers both electronic and physical items. The electronic or digital items can be delivered to the buyer directly through his/her email address. However, delivery of physical items is only possible at the buyer's location. The buyers can share their physical location with the vendors through secret communication services available on the dark web. The secrecy of the communication can be enhanced with the encryption techniques like Pretty Good Privacy (PGP). The cryptomarkets prefer the PGP to resist law enforcement actions [19].

PGP is an encryption technology that ensures the privacy and security of the messages involved in the communication between the two parties. It is a cryptographic technique where a pair of public-private keys is generated for encrypting the message. The sender of the message shares the public key with the receiver for encrypting the message. The receiver upon receiving the message from the sender decrypts it with the private key.

Upon receiving the address from the buyer, the vendors pack the bought item in different forms depending on the nature of the item to minimize the suspicion on the packaging. Drugs are packed in airtight packets and concealed in envelopes to minimize their smell so that the frequent posts from the same address do not look dubious. Vendors may also post shipments at different times and may also use different addresses to minimize the suspicion from the intelligence agencies. The buyers of the products may use the address of some other buildings in the locality or that of abandoned locations instead of their correct address. They then receive the package from that address to avoid possible law enforcement intervention [21].

Once the buyer receives the shipment successfully, he/she is expected to post the feedback of the purchase on the cryptomarkets via online feedback forms. Most of the buyer's feedback on the cryptomarkets indicates the positive review of the products [19]. The prospective customer can examine this feedback about vendors and their products to make any future purchase. The buyers can also use the cryptomarket forums to share their shopping experiences on the dark web. The cryptomarkets can directly host their discussion forums or they may provide a link to an individual hidden service specifically providing the forums. Cryptomarkets like Alphabay and Dream market have their forums to facilitate interaction between the customers and the vendors [19].

1.2.3 Why use Cryptomarkets?

The main reason behind the popularity of the cryptomarkets for purchasing drugs was the comfort of getting them delivered at the doorstep from the varied range of drug types with different qualities from the different vendors [22]. This is complemented by the transparent review of the products through discussion forums and online ratings [18], [22-24]. Many users have also acknowledged the level of safety when buying online instead of conventional street purchasing where there is a risk of violence [25].

Furthermore, the anonymity that comes with the cryptomarkets like Silk Road was also the critical factor in building user trust as it makes the buyer thinks that there was much less risk of being intercepted by the law enforcement agencies [24]. However, many users have reported several cases of monetary loss and unable to receive the products when shopping on the cryptomarkets [26].

The above findings of the user experiences on cryptomarkets indicate several advantages of buying goods especially drugs online instead of streets. Though they come with some risks buying on cryptomarkets that when happens could mar the joy of first-time shoppers.

1.3 Law Enforcement Actions on the Dark Web

The illicit nature of the products available on the cryptomarkets has put them under the radar of law enforcement and surveillance agencies. Several attempts have been made to take down the illicit cryptomarkets and were successful.

Silk Road was busted in the year 2013 by the FBI and reportedly recovered Bitcoin amounting to the value of around 3.6 million US dollars [21]. The agents of the FBI disguised as vendors have managed to reach

and communicate with Ross Ulbricht, the administrator of the Silk Road [27]. Consequently, after the closure of the Silk Road, Ross Ulbricht was taken into custody and was charged under the narcotics laws which sentenced him to life imprisonment [28].

The closure of the Silk Road did not have much effect on the dark web ecosystem and the sales of products grew rapidly after a short time [19]. The Dutch Police conducted Operation Commodore in February 2014 to bust another marketplace called *Utopia*. The Dutch Police utilized the same technique as that of the FBI in the case of Silk Road.

However, one of the most disruptive law enforcement actions was Operation Onymous that was carried out in November 2014. The Europol, Department of Homeland Security and the FBI collaborated to attack the Tor dark web. The result of Operation Onymous was that several Tor hidden services were taken down including eleven marketplaces namely: *Alpaca, Black Market, Bluesky, Cannabis Road 3, Cloud9, Flugswamp, Hydra, Pandora, Silk Road 2.0, Topix2* and *Torbazaar* [29]. Several individuals were arrested who were involved in the functioning of these cryptomarkets and were prosecuted [30].

In Operation Hyperion carried out in November 2016, a different approach was taken from the previous attempts. It was jointly conducted by the intelligence agencies of Australia, Canada, New Zealand, the United Kingdom and the US. It was not aimed to close the marketplace instead focused to contact the people who were allegedly involved in this trade. The specific persons were warned to stop their activity immediately and failing to do so would lead to legal action. Several thousand people were talked in this regard by the Swedish police only among others. The names of the individuals who were found to be involved were made publicly available by the Dutch police [31].

The most recent law enforcement action happened in July 2017 jointly conducted by the FBI, the Drug Enforcement Administration (DEA) and the Dutch Police. The operation resulted in the shutdown of *Hansa* and *Alphabay* cryptomarket. The Dutch police took control over the *Hansa* but did not shut it down, instead, it was kept online to keep watch on the user activities. They then collaborated with the DEA and FBI to watch the movement of users to *Hansa* after the closure of the *Alphabay*. As expected, the dislodged users of the *Alphabay* have moved to the *Hansa* giving an opportunity to the law enforcement agencies to better understand their activities [32].

1.3.1 Impact of the Law Enforcement Actions

The impact of law enforcement action on the cryptomarkets was analyzed in multiple studies. The closure of the Silk Road saw a drastic reduction in the sales volume but it began to increase rapidly after a short time [19]. After the shutdown, the usage of the cryptomarket remains popular with the rise of new marketplaces [33]. The same track of events happened after Operation Onymous indicating that such operations do not have a long-lasting impact on its users [34].

The dark web ecosystem is highly unstable and ever-evolving where new hidden services rapidly come up and go down [35]. Therefore, continuous monitoring of these websites is necessary to control their illicit activities. The law enforcement agencies should also focus on new techniques and methodologies to detect illegal activities being carried out on the cryptomarkets. This thesis aims to propose new methodology and techniques to counter illegal activities on the dark web which are important in terms of the law enforcement agencies. It specifically focuses on the two illicit activities: drugs and firearms trafficking on the Tor dark web. These activities were identified based on their prevalence on the dark web and their negative impact on society.

1.4 Motivations and Problem Formulation

In today's scenario, a popular search engine like Google is inevitable while gathering information from the Internet. However, they still cannot index the hidden services on the dark web despite employing the state of the art indexing and crawling algorithms. The profound amount of anonymity guaranteed by the Tor infrastructure has always tempted the criminals and fraudsters to expand their online trade thereby opening new challenges for the cyber-security and law enforcement experts [36].

Several individuals and collaborated attempts have been made by law enforcement agencies to disrupt the grey trade areas on the dark web with much success. However, the strong resilient nature of the hidden services has mandated continuous efforts by the law enforcement agencies to tackle illicit activities. Moreover, there also exist some key hidden services that are a crucial player in the sustenance of the dark web ecosystem [37]. Therefore identification and monitoring of such key websites could play an important role in studying the structure of dark web markets.

The dark web environment is dynamic in nature which requires continuous surveillance by law enforcement agencies. In such types of scenarios, automated tools and methodologies have a clear advantage over manual approaches. Hence, the researchers should focus on proposing automated approaches for tracking illicit activities over the dark web.

The literature review of the dark web in general and Tor hidden services, in particular, has uncovered that the trade of illicit drugs and other psychoactive substances is the most common illegal activity among others on the dark web. Moreover as discussed in section 1.3, the majority of the law enforcement actions on the cryptomarkets have the primary

goal of disrupting the drug trafficking network. Infact, the infamous Silk Road got busted by the FBI because of the volume of illicit drug business being carried out there. Therefore, preventing drugs sale on the dark web is on the priority list of the enforcement agencies. On the other hand, firearm trafficking on the dark web though small in number poses a great threat to the lives of people. A single pistol sold on a hidden service could endanger several lives if it gets into the hand of a criminal.

The existing studies have managed to expose only the qualitative characteristics of the dark web marketplaces. These studies have analyzed the data extracted from the marketplaces that have been shut down by law enforcement agencies. However, there is a need to develop dynamic methods that can be applied to any of the hidden services or marketplaces. Also, such methods should consider the content of the hidden services so that they can be applied to any type of hidden services. The present work in this thesis attempts to provide the solutions to these problems by proposing automated methods for identifying illicit activities on the dark web platform.

Machine learning and natural language processing methods have garnered good popularity in the domains of academic research. These methods have been proven to show good results accuracy and optimal performance in their relevant fields. These methods can also be applied in the current study to develop methods for identifying illicit activities on the dark web. The literature review of the existing work conducted in the next chapter reveals that albeit there has been several efforts to index the illicit activities on the dark web, still there are some gap and limitations in the existing methods.

To help support the law enforcement and intelligence agencies for monitoring and further investigation of illicit activities on the dark web,

we propose techniques based on machine learning and natural language processing. These techniques could prove beneficial in identifying key hidden services on the dark web involved in the illicit trade of drugs and firearms trafficking. Therefore, a study focusing on providing solutions for tackling the illegal dark web trade is proposed under the following title:

"*Identification and Classification of Illegal Content on Tor Dark Web*"

The research solutions to the problems presented in this work can be broadly divided into the following parts:

1.4.1 Hidden Services Classification

The classification unit can be applied to classify a hidden service into one of the pre-defined categories based on the textual content of the service. The objective here is to classify the illegal hidden services into various classes. There are different types of content present on the dark web including legitimate content but the law enforcement agencies are interested in only those services which offer illegal content. Moreover, manual monitoring of a large number of hidden services continuously may prove to be an impractical approach in terms of resources and cost-effectiveness. This motivated us to propose an automatic and efficient tool for the monitoring of the Tor hidden services by the law enforcement agencies. The main advantage of this tool is the automatic tagging of the hidden services that host illegal content.

1.4.2 Content Based Identification

This component could be used to identify key hidden services dealing in drugs and specific firearm listings using content based textual features. The proliferation of the hidden services on the Tor network

brings us the need of identifying the key domains and specific product listings among others. The content based identification could provide solutions to the problems like *What are the most harmful drug providing services in the Tor network?* or *Where are the firearm listings available on the Tor network?*

The answer to these questions may enhance the ability of law enforcement agencies in locating the key hidden services by channeling their efforts in monitoring them instead of the less significant ones. The content based identification would easily recognize a hidden service hosted under a new domain with the same content if it is previously been taken down by law enforcement actions. Moreover, it easily identifies a newly established service before it gains popularity among the users if it hosts similar content to a service that has previously been tagged by our methodology. In this way, the law enforcement agencies could take down newcomers before they could impose any threat to the users.

1.4.3 Link Based Identification

The last component could be applied to identify influential hidden services based on their hyperlink connectivity to the other services in the network. The link based identification of influential domains works by representing the Tor network into a directed graph where vertex represents a hidden service and an edge denotes the hyperlinks between the services. In this approach, the edges of the graph determine the influence of the node in the whole network. Contrary to the content based approach, the link based does not require a previously trained model for its working.

The link based can also identify the influential domains irrespective of the content hosted on it, thus it can be uniformly applied for any type of hidden services. However, one major drawback of the link based

approach is that it cannot identify the isolated nodes in the graph even if it is an influential one. The link based identification could act as a complementary approach along with the content based approach enabling the law enforcement agencies to direct their resources towards monitoring the specific and key services.

1.5 Research Objectives

The dark web is a hotspot of criminal activities. Drugs and firearm trafficking, child abuse, and even human trafficking and kidney racket are some of the most disgusting things that could be found online. The dark web can be regarded as an advanced and more sophisticated mechanism used by miscreants as compared to cybercrime. Moreover, it also serves as an example of how criminals leverage anonymous platforms for meeting out their ill intentions. It is complemented by the global reach of these activities that surpass the local geographical boundaries to mark their ubiquitous presence. The government and law enforcement agencies strive to ensure that these activities should not happen on anonymous platforms.

This thesis aims to put forth methods to identify and monitor the nefarious activities of the hidden services on the Tor dark web. To achieve this task, we proposed various methodologies based on machine learning, natural language processing, feature engineering and graph analysis that could be used for monitoring and further investigation by the law enforcement agencies.

The following objectives are adopted to accomplish this task:

- To design a crawler for scraping the textual content of the Tor hidden services.
- To study the network topology of the Tor dark web using the graph metrics.

- To use machine learning techniques for the classification of illicit activities on the dark web.
- To propose dimensionality reduction techniques for optimal classification performance.
- To propose a metric for calculating the harmful impact of the hidden services that is selling illicit drugs.
- To detect and identify the popular firearm listings using ensemble machine learning methods on hidden services offering the illegal trafficking of firearms.
- To identify important hidden services in the Tor network using hyperlink analysis.

1.6 Main Contributions

The major contributions of the thesis are summarized below:

- A dataset of the Tor hidden services consisting of 4102 Tor hidden services categorized into 31 different categories is presented. Each of the hidden services is manually labeled into the corresponding category.
- In first of its kind, a language-wise categorization of the hidden services in the dataset is performed. Besides English, hidden services were separately categorized for each of the 31 non-English languages.
- A two-step dimensionality reduction scheme is proposed for obtaining an optimal feature set to be used in the textual classification of the hidden services.
- A text classifier using the two-step dimensionality reduction scheme is proposed to classify the hidden services into five categories with better performance than the baseline methodology.
- A metric is proposed to rank the hidden services based on the harm level of the illicit drugs they offer to the users.

- An ensemble classification model with an engineered feature set is proposed to predict the specific firearm listings (handguns and rifles) on the cryptomarkets with good classification performance.
- We study the effect of the part-of-speech tagged feature set on the performance of the proposed ensemble model and demonstrate their superiority over the original feature set through experiments.
- We study the statistical mechanics of the Tor network by analyzing the topological and graph-theoretic properties. We uncover the small world and scale-free nature of the Tor dark web network.
- We propose a link based ranking methodology to identify the influential domains in the Tor network.

1.7 Limitations

Every study is limited by some facts that need to be discussed. The advantage of proposing new methods is that it provides whole new dimensions to the area under consideration while limitations reveal the possible points that the proposed study is lagging. Despite several reasons that move the current work in the direction of applying it in real practical scenarios, there are some limitations also. Some of them are listed below:

- The current approach only considers the textual content though there are multimedia content that could enhance the current work.
- The present work only consider the two illegal activities among a range of others on the dark web.
- The current work has only studied the Tor dark web skipping the I2P and the Freenet environment.

1.8 Thesis Outline

The thesis consists of seven chapters. The first chapter gives an introduction to the problem, motivation behind the study, research objectives and main contributions. A brief description of the remaining chapters is given as follows.

Chapter 2: Literature Review

This chapter reviews the existing literature for the methodologies and techniques associated with the objectives of research as discussed. It discusses the published work on the exploration of the dark web networks, automatic methods for identification and classification of illegal content on the dark web. Next, it analyses the work on focusing on the vulnerabilities in the Tor software and tools to further enhance the privacy of the users.

Moreover, it explores the studies on the topological properties of the dark web as well as the surface web. Next, it focuses on the variety of illegal products and services available on the Tor marketplaces, their mode of operations and other characteristics. Finally, it explores several link analysis algorithms for identifying important nodes in the web graph.

Chapter 3: Hidden Services Classification

This chapter describes the automatic classification of hidden services using the machine learning models. It starts by explaining the process of collection of the dark web text dataset followed by its analysis and manual categorization. Additionally, it addresses the problem of feature selection and proposed a two-step dimensionality reduction for generating an optimal feature set for the classifier. The effectiveness of the two-step dimensionality reduction scheme is tested on the two datasets using standard classification metrics. The results are also compared with the baseline methodology for the classification of Tor hidden services.

Chapter 4: Content Based Identification-Ranking Drugs Hidden Services

Chapter four describes the content based techniques proposed for identifying the illegal activities on the Tor dark web. A ranking methodology is proposed for ordering the hidden services involved in

illicit drug trafficking based on the severity and harmful effects of the controlled drugs. Next, it describes the generation of the ground truth data to be used as a benchmark for evaluating the performance of the proposed ranking technique.

Chapter 5: Content Based Identification-Predicting Firearm Listings

In this chapter, the methodology for predicting the specific firearm listings on the Tor dark web has been proposed. Furthermore, it describes the construction of the ensemble classification model and the engineered feature set for the prediction of firearm listings. Finally, it concludes with the results of the experiments followed by a discussion.

Chapter 6: Link Based Identification

Chapter 6 describes the algorithm for ranking the hidden services to identify the influential domains among them. It describes several factors that govern the influential nature of the hidden service followed by the formulation of the algorithm to rank them by their influence. It further describes the graph robustness metric that will be used for evaluating the performance of the ranking algorithm. Finally, it concludes with the experimental setup, results and their interpretation.

Chapter 7: Conclusions

The final chapter of the thesis outlines the research work carried out along with its conclusions followed by the possible future work that can be undertaken.

CHAPTER 2
LITERATURE REVIEW

The Internet has been the main driver behind the tremendous change in the lifestyle of the global population. It has surpassed the geographical boundaries to mark its appearance to even in the remotest place on the earth. However, it also opens up new directions for criminals to carry out unlawful activities. The dark web is one such platform that is infamous for its controversial activities carried out under the umbrella of sophisticated routing techniques that guarantee anonymity. In this chapter, detailed literature of the existing work on the various aspects of the dark web platform and its usage is discussed.

2.1 On Content Analysis of Tor Hidden Services

The research on the dark web has focused on exploring and analyzing the nature of the content available on the Tor hidden services and their classification based on the legality of the content. One of the first studies in this direction was conducted by Guitton in the year 2013 [38]. The author has managed to collect around 1171 *.onion* addresses from three available databases. The collected data of hidden services were analyzed and classified into two main categories namely: ethical and unethical. They were further classified into 23 sub-categories. Around 45 percent of the content was reported to be into the unethical category with child abuse content having a major share. The results have so much shaken the author that he advocates stopping the further development of the Tor platform.

Biryukov *et al.* [39] perform the content analysis of the Tor hidden services on a much larger dataset than Guitton's [38] study. They analyzed the content of nearly forty thousand hidden services and classified them

into eighteen categories using automatic tools. They have utilized the MALLET and the *uClassifiy* tool for classification purposes. Around 44 percent of content was related to counterfeits products, drugs and weapon trafficking, personal and identifying information from the hacked accounts. They also examine the linguistic diversity of the hidden services and found content available in 16 different languages other than English. They concluded that both the legal and illegal content is present on the dark web albeit their volume is not clear.

Biryukov *et al.* [39] also performed the popularity analysis of the hidden services by recording the number of requests for a hidden service descriptor from the client-side and found that the botnet-related services were in demand. Following the same line, Spitters *et al.* [40] performed the classification of around a thousand hidden services using topic modeling and machine learning approaches. They confirm the findings of previous work [38] about the major presence of child abuse content along with other illegal and controversial content. However, they identify thirty languages in which content was present, much diverse than the previous study about the variety of languages present.

A study spanning six months has tried to explore the content and usage of the Tor hidden services [41]. The authors have set up the Tor servers and extracted the data from the Tor distributed hash table which resulted in the collection of a list of around 80,000 hidden services. According to the study, only 45,000 services were accessible of which drugs-related hidden services occur for around 15 percent of the total with fraudulent websites and marketplaces behind it with nine percent each. One of the important results that came out of the study was that a significant amount of the generated traffic (around eighty percent) contains the requests for the child abuse content. Additionally, they also discovered other illegal and unethical content in their work.

A study by Intelliag [42], has identified nearly 30,000 *.onion* domains, however, due to the transitory nature of the hidden services, only half of those could be reachable. It was found that the sites on file sharing, leaked data and financial fraud were prevalent among others, however, the sites promoting illegal weapon trafficking were much less (less than one percent). Also, the study did not try to explicitly categorize the content into the legal and illegal parts, though they reported that the majority of the content was of questionable nature that requires law enforcement attention.

The dubious nature of the dark web has perpetually attracted the researcher community. The studies focusing on indexing the hidden services have been continuously published. In this regard, Moore and Rid [2] has performed the classification of hidden services into 12 categories. They aim to reveal the true picture that depicts the nature of content hidden behind the curtain of the Tor network. They designed the customized web crawler that scrapes the text content and skips other types of content (like multimedia, graphics) to avoid immoral and unethical content which is a common problem in this domain. Similar to the results obtained in the previous studies, they also confirmed the presence of the child abuse content. Another notable finding in their results was the negligible presence of the extremist websites on the Tor.

The picture of the Tor dark web gets more cleared in yet another empirical analysis where the number of hidden services categories were refined into 26 different classes [43]. The size of the dataset under study was relatively large containing around eight thousand records. Out of 26 categories, eight were related to illegal activities. The authors of the study have also proposed the pipeline of the text classification for automatic labeling of the dark web textual content into one of the predefined categories.

2.2 On Automatic Classification of Tor Hidden Services

The literature review related to the content based classification of the web pages has been conducted in this section. The objective of the content based classification of web pages is to label them with one of the predefined classes. The text based classification is increasingly being applied in the detection of spam, sentiment analysis, hate content identification and identifying controversial content on the dark web network [39], [44-46]. The success of such methods depends on the availability of an adequate amount of training data [47] as well as the quality of the representative features [48].

Initially, when the dark web was not much popular among Internet users, the scientific community has focused its efforts on the surface web [49], [50]. However, the shutdown of the Silk Road had brought the dark web networks into the limelight after which several researchers have moved on to propose automatic methods for exploring the dark web. In a study specifically on the *Agora* marketplace, the authors presented the classifier for classifying the available products on the marketplace into twelve classes [51]. They have employed the feature selection in the combination of Principal Component Analysis for dimensionality reduction. The Support Vector Machine (SVM) classifier has achieved 79% classification accuracy.

In another study, the researchers proposed a hybrid term weighting technique to specifically target the terrorist content present on the dark web. They have combined the Entropy, Term Frequency-Inverse Document Frequency (TF-IDF) and Glasgow techniques and assessed their efficacy on the five different classifiers including SVM [52]. Al-Nabki *et al.* [43] have employed the automated approach for the classification task. The TF-IDF along with the logistic regression classifier have produced the highest classification accuracy. They have also set the baseline

methodology to assess the performance of the text classifier for categorizing the dark web content.

The automated approach was also used to classify products from the different discussion forums. More than four lakhs of forum posts were manually annotated to generate the training set for the classifier. The Neyman-Pearson prediction was found to be the most accurate, though not as good as the accuracy achieved through manual annotation [53].

An automatic classification model called the Tor-use Motivation Model (TMM) was proposed that was specifically designed to help the law enforcement agencies in identifying the illegal activities [54]. Their model has classified the hidden services into the illegal and legal and found that the majority of the content was of nefarious nature.

In another study, a parallel Tor browser framework was proposed with the aim of enhancing the scraping mechanism of the Tor hidden services [55]. A supervised classifier employing TF-IDF weighting and Naïve Bayes classifier was proposed for the categorization of the illicit activities on the Tor dark web [56].

Table 2.1 summarizes the various indexing efforts of the Tor hidden services. The categories defined and the size in the percentage of some key categories (if reported by the author) for each of the study is shown. It can be seen that the weapon-related hidden services remain almost constant across the studies. The drug-related services were changed from five to fifteen percent in different studies. However, the number of Bitcoin-related services depicts an upward trend with the percentage getting almost doubled in the recent study. The increasing popularity of Bitcoin also comes along with the related scams and frauds that could affect naïve users.

Table 2.1: Categories of the Tor hidden services content identified in different studies.

Research Work	Categories of Content Identified			
Guitton, 2013 [38]	- Child Abuse - Hacking - Black Market - Pornography - Drugs (4) - Hit man - Weapons (1) - Surveillance	- Bitcoin (2) - Everything - Search engine -Racial Discrimination - Personal - File Sharing - Informatics	-Subversion of the state power -General forum with unethical topics -Ethical and specific topic	- Politics - Anarchism - Energy - Politics - Communism - Unknown
Biryukov, *et al.*, 2014 [39]	- Adult - Drugs (15) - Politics - Counterfeit - Weapons (4) - FAQs, Tutorials	- Security - Anonymity - Hacking - Software, Hardware - Art	- Services - Games - Science - Digital libs - Sports - Technology	- Other
Owen and Savage, 2015 [41]	- Drugs (15) - Market - Fraud - Bitcoin (6) - Mail - Wiki	- Whistleblower - Counterfeit - Forum - Anonymity - Search - Hacking	- Hosting - Porn - Blog - Directory - Books - Abuse	- News - Guns (~1) - Gambling
Intelliag, 2016 [42]	- File Sharing - Leaked Data - Financial Fraud - News Media - Promotion	- Discussion Forum - Drugs (4) - Internet / Computing	- Porno/Fetish - Hacking - Weapons (0.3) - Other	
Moore and Rid, 2016 [2]	- Arms - Drugs - Extremism - Finance - Hacking - Nexus	- Illegitimate Pornography - Other Illicit - Social - Violence - Other	- Unknown	
Al Nabki *et al.*, 2017 [43]	- Violence - Services - Drugs - Marketplaces - Pornography - Social Network - Cryptocurrency - Art/Music	- Gambling - Down - Empty - Forum - Hacking - Hosting and Software - Fraud	- Wiki - Leaked Data - Locked - Personal - Politics - Religion - Library/Books	- Counterfeit Money - Counterfeit Credit Cards - Counterfeit - Personal Identification - Human Trafficking
Dalins, 2018 [54]	- Drugs/Narcotics - Extremism - Finance - Child Exploitation - Hacking	- Identification /Credentials - Intellectual Property/ Copyright Materials	- Pornography-Adult - Search Engine/ Index - Unclear	- Violence - Weapons - Other-Not of Interest

2.2.1 On Language Analysis of Tor Content

A limited number of studies have also tried to identify the variety of languages other than the English in which the content is offered on the Tor dark web. Table 2.2 shows the non-English languages identified in the Tor content across the three studies. Biryukov *et al.* have applied the *Langdetect* tool to identify the type languages in their dataset [39]. They have reported that 84 percent of the total content was in English followed by the other languages. They have identified 16 different non-English languages of which the majority of them were European languages.

Table 2.2: Non-English languages identified in the Tor content in different studies.

Research Work	Number of languages identified	Languages			
Biryukov, *et al.*, 2014 [39]	16	Arabic	Czech	Hungarian	Portuguese
		Bantu	Dutch	Italian	Russian
		Basque	French	Japanese	Spanish
		Chinese	German	Polish	Swedish
Spitters *et al.*, 2014 [40]	29*	Italian	Bulgarian	Finnish	Greek
		German	Spanish	Portuguese	Croatian
		Russian	Polish	Swedish	Chinese
		French	Danish	Dutch	Japanese
		Albanian	Hungarian	Romanian	
		Slovene	Czech	Turkish	
Intelliag, 2016 [42]	30	German	Italian	Kinyarwanda	Amharic
		Chinese	Tai-kadai	Maltese Irish	Georgian
		French	Polish	Greek	Malay
		Russian	Norwegian	Bulgarian	Kannada
		Spanish	Danish	Luxembourgish	Romanian
		Dutch	Swedish	Kurdish	Hebrew
		Quechuan	Finnish	Chinese	
		Portugese	Armenian	Korean	

*Some languages with less percentage were not explicitly mentioned

Spitters *et al.* [40] also use the automated classifier to identify the non-English languages. They have identified 30 different languages with English being the majority with 83.27 percent. The Intelliag report [42] published in the year 2016, has identified 31 different languages in their collected dataset. However, the report explicitly did not tell about the method of language detection. With a small dip in the percentage as compared with the previous findings, English again topped the chart with 76 percent followed by German and Chinese languages.

All three studies have only identified the percentage of content present in non-English languages. The bifurcation of content among different categories for the non-English content has not yet been explored in any of the studies to the best of our knowledge.

2.3 On Tor Vulnerabilities

A number of studies have focused on identifying the loopholes and vulnerabilities in the Tor software, or customized attacks on the network and providing adequate measures to tackle the Tor anonymity. In this regard, TorFlow, a collection of tools was proposed for quantifying the effectiveness of Tor routers in terms of their reliability [57]. Through the Tor directory, such measurement could be made available to the Tor client in choosing efficient relays. In a study, it was found that a malicious router could corrupt the information funneling through it by modifying its settings [58]. To counter such types of attacks, the authors introduced the opportunistic algorithm for measuring the bandwidth that could help the users in achieving better anonymity.

In another study, StegoTorus [59], tool was proposed to conceal the Tor usage from the protocol analysis technique. It achieves this by disguising the packets in such a way that it resembles it coming from the other protocols like the HTTP. In the same way, at the transport layer, the

Tor circuits are spread over a number of small connections for each packet to hide the actual protocol. In Biryukov's, *et al.* work [39], they took the advantage of the entry nodes or guard nodes of the onion routing circuit to reveal the identity of the users of a hidden service. They also performed the popularity analysis of the hidden services by analyzing the traffic directed towards them.

TorBen [60], a deanonymization technique was put forth that leverages the two vulnerabilities in a web page. The first of them was loading malicious content from the un-trusted source and ineffective protection of the request-response pair size. TorBen was designed to manipulate a web page in a way that could expose the identity of the user visiting that specific web page.

2.4 On Tor Network Structure

Limited work has discussed the graph-theoretic properties of the Tor dark web graph. A web graph is a graph consisting of a set of nodes (or vertices) and edges connecting the nodes. A node in a web graph may either refer to the complete website or some web pages of the website. The edge in a web graph is the hyperlinks between any two websites.

Griffith *et al.* [61] were one of the first to analyze the Tor hidden services from a graph-theoretic perspective. They have used the *scrapinghub.com* service to crawl the Tor hidden services via the *tor2web* proxy. The nodes in the web graph were the domains and the edge denotes a hyperlink between any of the web pages of the two domains. The most notable result they found was that most of the hidden services (87 percent) do not have outgoing links to the other services. Also, the entire network is susceptible to disintegration upon removal of the central nodes with high in-degree values. They have also confirmed the presence of the bow-tie structure in the graph, albeit relatively smaller in size.

In the same year, Bernaschi *et al*. [62] analyzed the Tor hidden services graph at three aggregation levels on a relatively large dataset. They have used the BUbiNG crawler that made connections to nearly a million of the *.onion* addresses. Several graph metrics were reported at the page level, host level and service level for both directed and undirected versions of the graph. They have also reported that the majority of the hidden services have zero out-degree. In fact, around 90 percent of the hidden services have at most one out-degree.

Consequently, the graph was highly disconnected with some of the directory/Wiki services located at the center of the graph. The graph resembles much like the forest of directed extended stars where a small number of nodes connects to nearly all the services in the graph having very small in-degree and almost zero outgoing links. A semantic analysis of the results was performed to identify any relationship with the topology of the network structure.

In a similar, study Bernaschi *et al*. [63] again analyzed the Tor web graph constructed from the data captured at the three different time periods. They have tried to relate the Tor graph web with the three well-known random graph models namely: Erdos-Renyi (ER), Watts-Strogatz (WS) and Barabasi-Albert (BA) models. Though the Tor graph shows many distinctive features, a deeper analysis is required to affirm that it matches with the characteristics of any of the random network models. However, it was quite clear that the Tor graph bears the properties of the small-world networks.

The Tor hyperlinks graph was also used to perform social network analysis to know about how the criminals and fraudsters take advantage of such structures to perform their illicit activities [64]. They have collected the data from 1120 dark web hidden services in their study. The social

network analysis found that the mutual connection among the Tor nodes facilitates the homophily effect in the network. A group of around 60 websites was found that form the core-periphery of the network that acts as a gateway to the incoming users. The top sites in this group were providing the directory/services to other hidden services and removal of these sites could be proved beneficial in limiting its further use. The network was not resilient to the removal of the hubs leaving the network vulnerable to the law enforcement or hacktivist that always look to disrupt such criminal networks. Moreover, the removal of central hubs may also restrict a large number of users from entering the network.

In one of its first kind, the authors study the network structure properties of the dark web forums on the Tor network and compared their properties to their counterparts on the open web [65]. They aimed to reveal the effect of the anonymous forum users on the structure of the network. To achieve their aim, they have created the interaction graph from the forum data where the participant makes up the nodes and their mutual interactions were represented by the edges of the graph. The degree distribution of the graph shows that most of the users have either one or two degree values while few highly active users have much larger degree values.

Moreover, a public forum user easily posts a message while a dark web forum user prefers not to leave any comments on the post. This also led to a slower decay in the life-time of public forum threads as compared with the dark web forum. The hierarchical structure of the network is governed by the behavior of the root commentator, the user who initiates the conversation in a thread. The levels of hierarchy in the network decrease if the root commentator responds back to a large number of other users in the thread. However, the structure starts to take a more hierarchical form if the root commentator does respond back to others.

Several works are also carried out for the surface web graph with some of the early studies that focused on studying the topological properties of the network [66], proposing family of the graph models that suited the surface web [67], identification of the bow-tie structure in the graph [68] and performing the analysis of the surface web graph at the different aggregation levels [69].

Some of the early studies include Kumar *et al*. [66] that uncover the polynomial distribution of the in-degree and out-degree of the nodes. Contrary to that, Barabási *et al*. [70] showed that the degree distribution follows the power law. Broder *et al*. [68] analyses the graph constructed from the sufficiently large dataset and found that the web graph resembles a structure similar to the bow-tie arrangement consisting of six different components. They also confirm the presence of the power law in the degree distribution as claimed in the earlier study. In yet another study, the authors again confirm the power law in the degree distribution [70]. They also compare the size of the bow-tie components with that of the Broder *et al*. [68].

Lehmberg *et al*. [71] perform the analysis of the aggregated version of the web graph. Their dataset consists of 3.5 billion web pages and 128 billion hyperlinks among them. As found in the earlier studies, power-law was followed by the degree distribution along with the presence of the bow-tie structure. They have proposed a hypothetical structure corresponding to the web graph that is composed of two layers. The Low Degree Layer contains sparsely connected websites while the High Degree Layer contains the densely connected websites.

Meusel *et al*. [69] separately analyze the web graph at the three aggregation levels: page, host and pay-level domain. The data for the web graph was compiled by the Common Crawl Foundation. They observed a

considerable increase in the connectivity and average in-degree value of the page graph. They also computed and compared the various graph metrics at different aggregation levels. A study has also analyzed the effect of the crawling mechanism for data on the various graph metrics [72]. They suggest the need for a framework to analyze the sampling errors that arise due to different scraping mechanisms. This would help in defining the correct structure of the web unaffected by the change in the scraping technique.

These studies could act as a base for further research in getting insight into the network structure and their probable role in facilitating criminal activities on the dark web. The application of the social network analysis can help identify the nodes that form communities that allow for better traversal of the Tor network and could better help in identifying the criminal activities occurring there.

2.5 On Tor Cryptomarkets

Many research studies have been performed on the various aspects of the cryptomarkets like the nature of products, customer and vendor characteristics, business environment, quality of products, the geographical reach of shipments etc. Some studies have tried to classify drugs into different categories while others have focused on analyzing the extent of illegal wildlife and firearm trafficking. The relevant literature pertaining to the above-discussed aspects shall be explored in the subsequent sections.

2.5.1 On Customers and Vendors

A study has observed and collected data from 16 cryptomarkets spanning over the period of four years to get insight into the vendor characteristics [19]. They have continuously monitored the active vendors and found that over 10 percent of the vendors remain active throughout

their monitoring time period. They also noticed the surge in the number of vendors after the shutdown of the Silk Road. The majority of vendors have only managed to sell under USD 1000 of the products with only a few have crossed the 100 thousand USD mark. This evaluation was based on the price of the product and the corresponding number of user reviews it has got. Two-thirds of the vendor chooses to use the PGP keys for ensuring their security and it further increased after several law enforcement actions.

Later on, a group of eight cryptomarkets i.e. *Agora, Blue Sky, Evolution, Silk Road 2.0, Cloud 9, Pandora, Hydra* and *Andromeda* was studied [73]. Around two hundred vendor profiles and nearly four thousand product listings were scraped. Most of the vendors were found to be operating only on the single cryptomarket with few listings, while, the vendors operating across multiple cryptomarkets have a greater number of listings. Also, the vendors were found to be selling different products on different marketplaces.

Afterward, the same author has tried to look into the drift in the shipping of the offered products on the *Evolution* marketplace [74]. For that purpose, more than four thousand vendor's data was analyzed. The analysis found that the English-speaking countries along with the western countries are predominantly present with each country dealing in specific drug types. The study also found variation among the shipping locations of the different countries where some countries ship only to the domestic locations while other countries cater to the international locations.

Several attempts by law enforcement actions have also been carried out to disrupt the illicit trade on the cryptomarkets. The effect of these actions on the business of the vendors has also been analyzed in several studies and most of them found that they are resilient to such actions.

After the shutdown of the Silk Road, the vendors chose to shift to other marketplaces instead of leaving the business. The study has shown that the number of vendors got almost doubled on the *Black Market Reloaded* and the *Sheep* [75].

In order to quantify the overall sales volume on the cryptomarket, the number of feedback received for each product was used as estimated sales of that product [18], [19], [76]. However, the number of feedback could not adequately represent the sales value as the single user could buy multiple products but gives a review for any single product or a user may not provide feedback of the product at all [19]. Using this technique may result in underestimating the sales volume. Moreover, this method could only be applied to the marketplaces that provide support for the feedback and rating forms.

Customer satisfaction was also measured by analyzing the feedback data where more than 97 percent of the customers were responded with positive feedback. Only a few customers have given negative feedback indicating that only a few of them were not satisfied with the vendor products [18]. In this regard, the reputation of the vendor was also evaluated on the Silk Road, where around 10000 sales data were analyzed. It was found that the decrease in the vendor reputation affects the price of the products. The fall in the specific product rating has a greater impact than the overall vendor rating [77]. In a similar study on the *Silkkitie* marketplace, it was found that both the reputation of the vendor and the availability of the products have a direct impact on the sales of the vendor [78]. They also found that the vendors have a short life span and many of their products could not be able to get even a single buyer.

To summarize, the above studies have highlighted the importance of the trustworthiness between the customer and vendors while dealing

online. The reputation of vendors on the surface web markets plays a significant role in constructing a healthy customer base where the customer always has alternatives to express dissatisfaction with the products elsewhere. In case of cryptomarkets, trustworthiness becomes more important which is driven by anonymity, quality of products and the near absence of law enforcement agencies.

2.5.2 On Drugs Trafficking

The initial studies have used the information provided by the cryptomarkets to categorize the available drugs [18], [19], [79]. A study on the Silk Road was conducted to check whether it deals especially in drugs or not. Around 13000 pages were scraped containing more than 10000 listings. It was concluded that most of the items were for resale [76]. The analysis of the data scraped from the Agora and Evolution cryptomarket found more than forty thousand listings from over two thousand vendor profiles [80]. The majority of the product listings were related to drugs and other controlled substances. They statistically identified the significant difference in the price of the drugs and non-drugs items on sale.

The impact of the cryptomarkets on the global drugs network was analyzed by retrieving the vendor locations of advertisements that offer cannabis, cocaine and opioids on four cryptomarkets. A non-negligible difference was found between the drug producer countries and the vendor locations which indicate that the dark web drug trades focused mainly on the drug consumer countries [81].

In another study, the locations of the vendors and their drug supply on the *Agora* were studied by analyzing data collected on the randomly selected seven days [79]. The number of listings and drug types differed from country to country. They also noticed the relation between the

country and the type of drug offered by the vendors from that country. For example, US-based vendors offer a large number of cannabis than vendors from other countries.

The user profile of the cryptomarket was studied in a survey-based study [26]. Most of the users were quite young males of which the majority of them were employed followed by the students. A relatively large number of users were highly educated with around 38 percent of them were university pass outs. Ecstasy, cannabis and LSD were reportedly the most popular among users who they buy for their personal consumption or other people. The survey also revealed that the buyer prefers the online platform for buying drugs because it is more convenient and safe from violence as compared to traditional street buying where possibility of violence do exist.

2.5.3 On Firearms Trafficking

RAND Corporation conducted exclusive research on illegal firearm trafficking over the Tor dark web [82]. Around 800 product listings related to firearms across twelve marketplaces were analyzed in the study. They found that the small firearms were most common followed by some rifles and sub-machine guns. Besides physical items, they also reported a decent volume of digital material like manuals and handbooks on the topics like manufacturing explosive items using household materials. The firearms were comparatively expensive than that available in offline stores. Since the study has covered the firearm trafficking on the cryptomarkets, the stand-alone shops managed by a single vendor were not analyzed.

The vendor shops on the Tor network were analyzed in another study similar to that of the RAND Corporation in which the six online vendor shops were taken into consideration [83]. In this study also, they

found that small firearms like pistols were the most sought after products among others. However, the majority of the pistol listings were offered by only one shop out of the six. Additionally, each of the identified vendors across all the shops has at least a single listing of a pistol to offer to its customers. The vendors frequently provide a range of premium products to their buyers.

In another study, the listings of nine popular marketplaces were studied to get the latest status of firearm trafficking on the Tor dark web [84]. A total of 386 weapon listings were identified across the nine locations. Personal defense weapons including tasers, knuckle dusters, batons, pepper sprays etc were 28 percent of the total listings. Lethal weapons like firearms were only 25 percent. Across the nine marketplaces, 96 unique vendor profiles were found many of them were present in two or more marketplaces. Overall, the study concludes that the firearm trafficking on the Tor network is relatively small when compared to other illegal items like drugs and addictive substances, CC dumps and financial frauds.

Recently, a study has proposed an automated approach for the detection of firearm-related activities on the dark web discussion forums [85]. Basically, the machine learning techniques were utilized to identify discussion threads on the forums that are suspicious for procuring illegal weapons. The dataset for the model was created from the four discussion forums promoting terror propaganda on the dark web. The threads in the dataset were manually labeled by multiple experts as to either it is for procuring an illegal weapon or not. The annotated dataset was used to train the machine learning classification models to predict the class of a new thread. The study has also performed a comparative analysis of the existing work on the dark web discussion forum promoting a militant mindset.

2.5.4 On other Illegal and Controversial Content

The dark web cryptomarkets hold a bad reputation for drug trafficking but besides this, the marketplace also facilitates other illicit services including the sale of credit card (CC) dumps, personal information, illegal wildlife products, suicidal content, ransomware, etc.

A study has scraped the data across the eight marketplaces to get insight into the cybercrime business on the dark web [86]. The author analyzed the sale of different such products (like hacked e-mail accounts, botnets, ransomware, forged documents, digital piracy etc) and estimated the sale of cybercrime services and products to be approximately 15M USD in a period of about five years. However, according to the authors, the sales figures could be underestimated due to the limitations of the data collected for the study. The authors went on to identify the key product for each of the categories of products and services using the topic modeling techniques. They reported that the top three products come out to be the fake CC details, CC dumps and manuals for recruiting money mules.

The dark web has also been reported of providing content that could provoke suicidal thoughts on the users [87], [88] or doxing that could lead the victim to commit suicide [89]. In this regard, a study has systematically explored the dark web with the intent of finding suicidal content [90]. The authors of the work have identified nine search engines exclusively used on the Tor network and collected the first 30 results provided by selected search engines when searched for the words “suicide” and “suicide methods”. Their search result only found 4 percent of the total hits related to the suicidal content. Moreover, more than half of the search results were of outdated and unreachable content or content that is not related to suicide. Overall, they could not find any hidden service that exclusively offers suicidal content.

A research study by the INTERPOL has found clear evidence of illegal wildlife trade on the dark web [91]. Although few in numbers, the evidence clearly indicate that the products of some of the critically endangered species like rhinoceros horn were found in the study. The five existing rhino species native to Africa and Asia are listed in Appendix I of the Convention on International Trade in Endangered Species (CITES) thereby prohibiting the trade of rhino or their products between countries. However, the rise in the demand for the rhino horn has put this species in crisis [92].

2.6 On Identifying Key Hidden Services

In the past several years, a surge in the number of research work has been witnessed related to the illicit activities on the Tor network on different categories like the illegal drugs trafficking [73], [93], [94], terrorism [95], firearm trafficking, violence [84] and cybercrime [96]. However, only some of the work has focused on exploring the dark web data to identify the key hidden services. The detection of the important hidden services may either be performed by analyzing the hyperlinks in the network which is referred to as link based analysis [97] or by examining the content of the hidden services, called the content based analysis [98]. However, the two approaches can be combined into a single hybrid approach for better results [99].

The link based analysis utilizes the graph theory and the related metrics to compute the significance of a particular website or a node. A combination of the SVM classifier and the link based ranking method was proposed to identify the eligible applicants who had applied for bank loans [100]. Similarly, graph based metrics like degree distribution and centrality measures were employed to detect money laundering crimes on the social networks [101]. The robustness of the terrorist networks against the removal of components of the networks having high in-degree value or

betweenness centrality was examined in another study [102]. In yet another study, the betweenness centrality along with the Katz centrality was used to determine the influential nodes in the network [103].

The social network analysis technique was used in identifying the key users in a Facebook group using the values of the betweenness and eigenvector centrality metrics [104]. Additionally, a study has used graph-theoretic properties of the GitHub network to identify the most influential repositories among them by creating a star relation graph [105]. On the same line, a modified PageRank algorithm was proposed called the UserRank algorithm for the GitHub network [106]. A sentiment analysis algorithm was proposed to identify the social media influencers on leading social networks and microblogging platforms [107]. Additionally, the influence of social media users was also estimated by proposing a new score metric. The metric considers the interaction of a user with its follower users in the network on the tweets posted by the user to calculate the overall influence of the concerned user [108].

In a recent study on ranking the Tor hidden services, a link based algorithm called the ToRank algorithm is proposed to identify the influential Tor domains [37]. The proposed algorithm works by building a graph structure where each node represents the Tor hidden service and the edges represent the hyperlinks. Each node is assigned a weight based on its in-degree and out-degree values and is subsequently ranked based on its assigned weights. The nodes with top ranks are the influential or key hidden services in the entire network.

The ToRank algorithm is compared with other state-of-the-art algorithms like PageRank, HITZ and Katz. The empirical comparison shows the superiority of the proposed algorithm over the others. The ToRank algorithm was also tested on other large networks. ToRank

algorithm could be employed by the law enforcement agencies to detect the most suspicious domains on the Tor network. However, it suffers from one major drawback of not being able to detect the isolated nodes in the network given it is link based algorithm that does not consider the content of the node.

The link based algorithms may not apply to those nodes that are not connected to other nodes via hyperlinks or have few links to other nodes in the networks despite hosting influential content. In such scenarios, the content based approach becomes indispensable that has the ability to detect even the isolated nodes if they offer any significant content in terms of law enforcement perspective. However, at the time of the writing of this thesis, only a few studies have applied the content based approach in identifying the prominent entity on the Tor dark web networks. Some of the work on the content based methodology is discussed as follows.

The terrorist organizations are attracted by the anonymous nature of the dark web platforms for spreading their propaganda, fundraising, etc [95], [109-111]. A study was undertaken to identify the influential people in the group of radicals on the dark web discussion forums [99]. They collected the user data from their profiles and leverage the textual content to calculate the metric that reflects the radicalness of the user. They integrated the metric value with the modified PageRank algorithm to generate a list of radical users in a particular group in the discussion forum.

A similar study was conducted to detect the influential users on the Twitter platform [112]. They captured the relevant details from the user's Twitter profile like the number of tweets, nature of tweets and used them to represent them in a graph-like structure. The graph metrics were then used to identify key users among all others on Twitter.

CHAPTER 3

HIDDEN SERVICES CLASSIFICATION

The users of the dark web and Tor network enjoy an advanced level of privacy and anonymity. This creates a fortified ground for criminals to conduct illegal and unethical activities online. Most of the studies have claimed that services and content offered on the dark web were illegal drugs, arms and ammunition, child abuse content, forged documents and counterfeits [38-40]. The monitoring of these activities by the law enforcement agencies is limited by the unavailability of a search engine that could index the Tor hidden services. Motivated by the need to identify and detect the illegal content, we aimed to design and develop machine learning models to classify the illegal content in the Tor network.

3.1 Background

Text classification is a supervised machine learning technique that accepts text documents in an n-dimensional vector format and outputs its class label. The member elements of the vectors are words or a group of words in the text documents commonly referred to as *terms* or *features*. The collection of all the features in the dataset is called feature set or feature space. A dataset containing large number of samples may produce a relatively high dimensional feature set consisting of many irrelevant and redundant features. These unwanted features may increase the computational complexity of the classification model and degrade its performance. This problem is famously known as the *"curse of dimensionality"*. To overcome this problem, Dimensionality Reduction (DR) techniques are used to eliminate the unwanted features from the feature space. DR techniques are of two types: *Feature Selection* and *Feature Extraction* [113].

3.1.1 Feature Selection

Feature Selection is the method of selecting only the most suitable and relevant terms from the original feature set based on some criteria. There are three different approaches to achieve this task: *filter*, *wrapper* and *embedded* approach. The filter approach performs statistical analysis of the feature space to calculate scores of each feature indicating its relevance. High-scoring features are selected for further analysis. The filter techniques work independently of the underlying classification algorithms and thus are computationally cheap. Some of the commonly used filter approaches are: Mutual Information (MI) [114], Gini Index (GINI) [115], reliefF [116], chi-square (CHI) [117], Document Frequency (DF) [118], Term Variance (TV) [118], Information Gain (IG) [119], Laplacian Score (LS) [120], Absolute Cosine (AC) [121] method etc.

The wrapper approach uses a learning algorithm to evaluate the performance of the selected features. They are computationally expensive due to the involvement of the learning algorithms as compared to the filter methods. The resultant feature set may produce generalized results for the algorithm it was used for the selection process. The embedded approach generally uses machine learning algorithms for classification. The searching criteria for the optimal feature subset are built into the classifier construction.

3.1.2 Feature Extraction

Feature Extraction uses mathematical transformations to project the original feature space into a new space with reduced dimensions. It replaces the redundant features with other new features that are much less in number but appropriately represent the information of the original feature space. Some of the methods of feature extraction are Principal Component Analysis (PCA) [122], Latent Semantic Indexing (LSI) [123] and Linear Discriminant Analysis (LDA) [124].

Our classification model presented here for the dark web text content uses a two-step DR technique to obtain a reduced feature set. In the first step, the filter approach is used to implement feature selection using Mutual Information. The second step applies the feature extraction on the feature set obtained from the preceding step. Here Linear Discriminant Analysis is applied to the feature set. The two-step DR technique should be able to achieve better classification performance than the individual DR schemes.

The classification of illegal content on the dark web is a multi-class classification problem where input text content is assigned a class label by the classifier from the set of predefined classes of illegal content. To correctly predict the class labels, the classification model needs to be trained on a pre-labeled corpus or dataset. Hence we create a labeled dataset of the Tor hidden services for classification problems.

3.2 Dark Web Text Dataset

In this section, we describe the procedure of collecting the dataset of Tor hidden services followed by the manual labeling of each of the sample in the dataset. The language wise categorization of the content is also performed.

3.2.1 Collection of Data

The dataset was collected in two phases by web crawlers. In the first phase, a customized crawler called *Link-Grabber* was built in Python for collecting as many *.onion* URLs as possible in a month. Link-Grabber was supplied with initial seeds from the Hidden Wiki on the surface web to make an entry into the Tor dark web. It connects to the Tor network through port 443 via the SOCKS proxy [125]. Each of the initial links was crawled by the Link-Grabber in the hunt for new links. The fresh links discovered by the crawler were stored in a separate file for further

processing. After eliminating the duplicate links from the collection, we were left with 25742 unique *.onion* URLs pointing to the hidden services.

With a list of onion domains in hand, another customized Python crawler *Content-Grabber* was employed to extract the textual content of the hidden services. The HTML of the home page of each of the onion domains was downloaded and stored in an individual file. The crawler skipped other content like the audio, image, video, hyperlinks and downloadable content. The Content-Grabber could only extract the content of 6,227 hidden services out of the total of 25,742. The remaining services were inaccessible at the time of crawl given the short-lived nature of the hidden services [35]. Thus crawling all the available hidden services on the Tor at any moment is infeasible. However, few hidden services restricted the automated crawlers from accessing; therefore we did not bypass the restrictions and skipped those services.

Before the actual labeling or classification of the hidden services under various categories, the collected data was put to cleansing by removing the hidden services that consist of either of the following:

- The textual content of not more than three words;
- Image with no accompanying text;
- Blank or empty web page;
- Redirection links only; and,
- Error messages like database error, server configuration error, client side-script error, etc.

A total of 2125 hidden services were removed during cleansing operation. The dataset was reduced to 4,102 hidden services after the cleansing process. Now, the labeling of the cleansed hidden services is carried out.

3.2.2 Labeling of the Hidden Services

Several attempts have been made to categorize the Tor content into various classes. However, the dynamic environment of the Tor dark web requires continuous efforts for indexing the hidden services to maintain their record. To get a clearer picture of the content, we have identified 31 different classes for categorizing the dataset. Moreover, in a first-ever attempt, we have tried to categorize the non-English content separately from the English content. The hidden services were manually classified to obtain precise categorization of the content.

Of the 4,102 hidden services in the dataset, 3,480 use the English language while the remaining 622 use the other languages. The resulting categorization of the English content is listed in Table 3.1. The legality of the content is determined by the USA laws.

Table 3.1: Categories and count of the Tor hidden services.

Category	Count	Category	Count	Category	Count
Adult Content	165 (4.7%)	Electronics	60 (1.7%)	Other Cryptocurrencies	62 (1.8%)
Bitcoin Doubling	188 (5.4%)	Ethical Hacking	112 (3.2%)	Personal Websites	50 (1.4%)
Bitcoin Mixer	117 (3.4%)	Forged Documents	40 (1.1%)	Political	9 (0.3%)
Bitcoin Trading	125(3.6%)	Forums & Others	337 (9.7%)	Religious	6 (0.2%)
Bitcoin Wallets	68 (2%)	Gambling-Betting	31 (0.9%)	Services	358 (10.3%)
Books	36 (1%)	Hosting	131 (3.8%)	Software	131 (3.8%)
CC Dumps & Others	271 (7.8%)	Login	127 (3.6%)	Tor	66 (1.9%)
Counterfeits	37 (1.1%)	Marketplace	355 (10.2%)	Uncensored Journalism	70 (2%)
Directory	128 (3.7%)	Music-Entertainment	44 (1.3%)	Violence	98 (2.8%)
Drugs	179 (5.1%)	News	26 (0.7%)	Whistleblowers	48 (1.4%)
Educational	5 (0.1%)	**Total Illegal**		**1315(38%)**	
		Total Legal		**2165(62%)**	
Grand Total				**3480**	

The services related to *Bitcoin* holds the major proportion of the dataset, so it was split into four categories: *Bitcoin doubling*, *Bitcoin mixer, Bitcoin trading* and *Bitcoin wallets*. The doubling services claimed to double the Bitcoin though their authenticity could not be verified. A large number of these services reflect the popularity of Bitcoin among other crypto-currencies for trading and shopping on the dark web.

The second-largest category is *Services* with a proportion of around ten percent of the dataset. They offer a variety of services both legal and illegal including e-mail service, privacy-enhancing services, encryption-decryption service, escrow etc. Some of the illegal and controversial ones include: arranging a Distributed Denial of Service (DDoS) attack, helping in illegal border crossing in wake of recent immigrant crisis [126] from the persecuted lands etc.

The category *CC dumps & Others* consist of the sites dealing in stolen credit card data for cloning purpose, social security numbers, account information of compromised bank accounts, PayPal accounts and other financial accounts, and personally identifiable information. Furthermore, the category *Counterfeits and Electronics* refers to the sale of fake currencies (mostly US dollars and Euros), counterfeit electronic items and stolen smart-phones. The *Drugs* include illegal drug trafficking including some of the most harmful drugs like *crack cocaine*, *heroin* and *cocaine*.

The class *Adult Content* is associated with pornographic content including extreme content. The pornographic content was not further analyzed due to moral and ethical grounds. Likewise, the category *Violence* as the name suggests contains violent and gory content capable of causing mental as well as physical trauma including fatality. It also includes the infamous *Hitmans,* the professional contract killers that can

be hired to kill a person. Though sounds exaggerated, some of the hitmen even claim to assassinate world leaders at high fees.

The category *Marketplace* refers to all the hidden services that are functioning as black markets commonly known as the cryptomarkets. The marketplace provides all the required products at a single site just like the traditional offline markets. However, they are more interested in offering illegal products to their customers. The *Directory* acts as the address book of the Tor network providing the onion links to the other services and their current status.

The category *Forums & Others* refer to all the blogs, discussion forums, social networks, chat groups, anonymous post groups on Tor. The dark web forums are key services concerning further research in studying the dark web where users discuss the services and review the products available on the marketplaces.

A notable feature of the dark web that contrasts it from the surface web is the near absence of educational content. Only a fraction of services were offering the educational content. Moreover, the buzz that was created globally over the probable use of the dark web platform for the propagation of the so-called Islamic terror ideology seems to be fizzling out in our results.

3.2.3 Non-English Content

The hidden services with non-English content were 622 in the collected dataset. Google Translate[6] was used to identify the language and 31 different languages were detected of which European languages like Russian, German, French etc tops the chart. The complete list of the languages and their frequency is shown in Table 3.2.

[6] https://translate.google.co.in/

The number of categories identified in the non-English content was 18 out of 31 categories of English content. The hidden services that displayed under maintenance message (68 out of 622) were removed before categorization. The nature of the content was pretty much legal based on the US laws as applied for English content. As much as 65% percent of the total content was legal while most of the illegal content was accounted by the adult content, drugs and forged documents similar to that in the English content. Discussion forums and their associated services had the largest share of content with 25% of hidden services. The distribution of categories among different non-English languages is shown in Table 3.3.

Table 3.2: Hidden Services in non-English content and their count.

S.No.	Language	Count	S.No.	Language	Count
1.	Russian	183	17.	Turkish	7
2.	German	87	18.	Polish	5
3.	Spanish	56	19.	Catalan	4
4.	French	53	20.	Hebrew	4
5.	Portuguese	44	21.	Swedish	4
6.	Chinese	30	22.	Bulgarian	1
7.	Indonesian	28	23.	Danish	3
8.	Arabic	19	24.	Bosnian	2
9.	Italian	18	25.	Afrikaans	1
10.	Finnish	15	26.	Bengali	1
11.	Latin	14	27.	Esperanto	1
12.	Japanese	9	28.	Greek	1
13.	Czech	8	29.	Luxembourgish	1
14.	Dutch	7	30.	Ukrainian	1
15.	Korean	7	31.	Vietnamese	1
16.	Thai	7	Total		622

Table 3.3: Distribution of categories among non-English content.

Language	Categories and Count		Total
Russian	Adult Content	9	Illegal: 77 (51%) Legal: 75 (49%) Total: 152
	Drugs	31	
	Forged Documents	7	
	Forums & Others	38	
	Hosting	24	
	Login	4	
	Marketplace	13	
	Political	3	
	Religious	6	
	Violence	17	
German	Adult Content	2	Illegal: 20 (26%) Legal: 58 (74%) Total: 78
	Drugs	4	
	Forums & Others	22	
	Ethical Hacking	2	
	Login	2	
	News	1	
	Services	36	
	Uncensored	9	
	Journalism		
Spanish	Adult Content	13	Illegal: 22 (39%) Legal: 34 (61%) Total: 56
	Drugs	1	
	Forged Documents	8	
	Forums & Others	19	
	Ethical Hacking	2	
	News	6	
	Personal Websites	3	
	Political	4	
French	Adult Content	5	Illegal: 22 (42%) Legal: 31 (58%) Total: 53
	Drugs	4	
	Forged Documents	9	
	Forums & Others	11	
	Ethical Hacking	2	
	Hosting	4	
	Login	1	
	Marketplace	4	

	News	2	
	Personal Websites	2	
	Religious	1	
	Uncensored Journalism	8	
Portuguese	Adult Content	8	Illegal: 9 (20%) Legal: 35 (80%) Total: 44
	Bitcoin Trading	1	
	Drugs	1	
	Forums & Others	13	
	Ethical Hacking	5	
	Music-Entertainment	4	
	News	8	
	Political	1	
	Religious	3	
Chinese	Adult Content	2	Illegal: 10 (38%) Legal: 16 (62%) Total: 26
	Forums & Others	5	
	News	5	
	Political	1	
	Services	8	
	Software	5	
Indonesian	Adult Content	3	Illegal: 7 (28%) Legal: 18 (72%) Total: 25
	Forums & Others	3	
	Ethical Hacking	1	
	Music-Entertainment	3	
	News	5	
	Personal Websites	2	
	Services	4	
	Software	1	
	Uncensored Journalism	3	
Arabic	Forums & Others	4	Legal: 15 (100%) Total: 15
	News	6	
	Personal Websites	4	
	Software	1	
Italian	Adult Content	4	Illegal: 5 (29%) Legal: 12 (71%) Total: 17
	Drugs	1	
	Forums & Others	7	

	News Political Religious	2 2 1	
Finnish	Drugs Forged Documents Forums & Others Hosting Login Software	2 4 2 2 2 3	Illegal: 6 (40%) Legal: 9 (60%) Total: 15
Latin	Drugs	1	Illegal: 1 (100%)
Japanese	Adult Content Forums & Others Software	6 1 2	Illegal: 6 (67%) Legal: 3 (33%) Total: 9
Czech	Bitcoin Trading Music-Entertainment Forums & Others	2 3 2	Legal: 7 (100%) Total: 7
Dutch	Adult Content Forums & Others News Personal Websites Software	1 1 1 1 2	Illegal: 1 (17%) Legal: 5 (83%) Total: 6
Korean	Forums & Others Login Personal Websites	3 1 3	Legal: 7 (100%) Total: 7
Thai	Forged Documents Forums & Others Ethical Hacking	2 1 4	Illegal: 2 (29%) Legal: 5 (71%) Total: 7
Turkish	Forums & Others Software Uncensored Journalism Violence	3 1 1 2	Illegal: 2 (29%) Legal: 5 (71%) Total: 7
Polish	Login Software	3 1	Legal: 4 (100%) Total: 4
Catalan	Political	4	Legal: 4 (100%)
Hebrew	Forums & Others Uncensored	2 2	Legal: 4 (100%) Total: 4

	Journalism		
Swedish	Drugs Forums & Others News	2 1 1	Illegal: 2 (50%) Legal: 2 (50%) Total: 4
Bulgarian	News	1	Legal: 1 (100%)
Danish	Ethical Hacking Hosting Violence	1 1 1	Illegal: 1 (33%) Legal: 2 (67%) Total: 3
Bosnian	Hosting Marketplace	1 1	Illegal: 1 (50%) Legal: 1 (50%) Total: 2
Afrikaans	Religious	1	Legal: 1 (100%)
Bengali	Violence	1	Illegal: 1 (100%)
Esperanto	Forums & Others	1	Legal: 1 (100%)
Greek	Political	1	Legal: 1 (100%)
Luxembourgish	Personal Websites	1	Legal: 1 (100%)
Ukrainian	Political	1	Legal: 1 (100%)
Vietnamese	Political	1	Legal: 1 (100%)
Total	Adult Content Bitcoin Trading Drugs Forged Documents Forums & Others Ethical Hacking Hosting Login Marketplace Music-Entertainment News Personal Websites Political Religious Services Software Uncensored Journalism Violence	53(10%) 3(1%) 47(8%) 30(5%) 139(25%) 17(3%) 32(6%) 13(2%) 18(3%) 10(2%) 38(7%) 16(3%) 18(3%) 12(2%) 48(9%) 16(3%) 23(4%) 21(4%)	**Illegal: 195 (35%)** **Legal: 359 (65%)** **Total: 554**

3.3 Methodology

The classification of text documents consists of the following tasks: Data Preprocessing, Term Weighting, Dimensionality Reduction, and Classification. Here we propose a two-step dimensionality reduction scheme for the classification of the dark web content. The application of the proposed scheme is evaluated using the three classifiers and the results are compared with other DR techniques.

3.3.1 Data Preprocessing

The documents in the dataset were preprocessed in three stages. Stop words are removed in the first stage. Stop words are the most commonly used words in a language (like *is*, *am*, *are*, *this*, *it*, etc) whose removal does not affect the meaning of the text. Python Natural Language Toolkit[7] (NLTK) package consisting of 58 such words is used for removing stop words. After removing stop words, symbols, numeric characters, punctuations and other special characters were removed. The complete text is also converted into lowercase. Finally, tokenization is performed on each of the documents to split the sentences into individual words.

3.3.2 Term Weighting

The documents in the dataset are represented in the *Bag of Words (BoW)* model. This model first creates a vocabulary of unique words that appears in the corpus [113]. The ordering of the words in the document is not taken into account while constructing the vocabulary. Moreover, the words that occur in less than two documents are not considered and removed from the vocabulary.

The text documents are turned into vectors of fixed length equal to the size of the vocabulary. Each element of the vector represents the term

[7] https://www.nltk.org/

of the vocabulary. These terms are assigned weights depending on their presence or absence in the document. We have used the *Term Frequency-Inverse Document Frequency (TF-IDF)* [113] to assign weights to the terms in the document vector. The idea of this measure is to assign a high weight to the term if it appears frequently in a document. However, if it also appears frequently in other documents of the corpus then it would not have been an important term and it is assigned a low weight. In short, an infrequent term in the corpus is an important term for the document and reverse. Given a corpus of size n, $tf - idf$ is defined by Equation (3.1):

$$tf - idf(t,d) = tf(t,d).idf(t) \tag{3.1}$$

where $tf(t,d)$ is the term frequency of term t in the document d i.e. the frequency of term t for each document d of the corpus.

$idf(t)$ is the inverse document frequency of term t is given by Equation (3.2):

$$idf(t) = \log(\frac{n}{df(t)}) + 1 \tag{3.2}$$

3.3.3 Dimensionality Reduction

The proposed DR scheme consists of two steps. In the first step, Mutual Information (MI) of each of the features in the vocabulary is computed followed by the application of LDA on the selected features from the first step to obtain a transformed feature space with reduced dimension. The two-step DR scheme is shown in Figure 3.1.

Mutual Information

Mutual Information is adopted from the field of information theory to estimate the mutual dependence between the feature and the corresponding class. MI is also defined as the measure of the amount of

information held by one random variable about another variable [127]. For selecting features, MI can be used to measure the importance of a feature with respect to its class. The MI between two variables X and Y is given by Equation (3.3):

$$MI(X:Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} p\big(X(i), Y(j)\big).\log\left(\frac{p(X(i),Y(j))}{p\big(X(i)\big).p(Y(j))}\right) \quad (3.3)$$

In case if $p\big(X(i), Y(j)\big) = p\big(X(i)\big).p(Y(j))$ i.e. if X and $Y$are statistically independent, then MI between X and Ywill be zero.

The features are ranked from high to low according to their MI score with the highest scoring feature at the top. From the ordered list of features, the top k percent of features are selected to reduce the dimension of feature space. Given a feature space having n documents and m features, the top k percent of ranked features will be selected for further analysis where. The reduced feature space will contain n documents with k percent of features.

The second step will further reduce the feature space obtained from step one. LDA technique is used for projecting the feature space into a new space having much lower dimensions.

Linear Discriminant Analysis

Linear Discriminant Analysis [124] is a statistical technique that finds new axes from the original dataset using class labels in a way that maximizes the distance among different classes. Both PCA and LDA follow a similar approach however unlike PCA, LDA is a supervised technique. It takes the original high dimensional feature space divided into classes as the input and transforms it into a new low dimensional feature space. The class structure remains unchanged by the

transformation. The differentiation between classes is obtained by minimizing the intra-class distance and maximizing the inter-class distance.

Given a dataset having d classes with feature space of length h, LDA can find new dimensions of size $\min(d-1, h)$. The eigenvectors of the training data are calculated and stored in inter-class and intra-class scatter matrices. The eigenvalues corresponding to the eigenvectors are arranged in the decreasing order of their length. The eigenvectors with the highest eigenvalues are kept as they are more informative than the smaller length eigenvectors. The selected eigenvectors are employed for transforming the dataset into a new space. After the end of the second step, the feature space is reduced to size $d-1$ or h whichever is smaller.

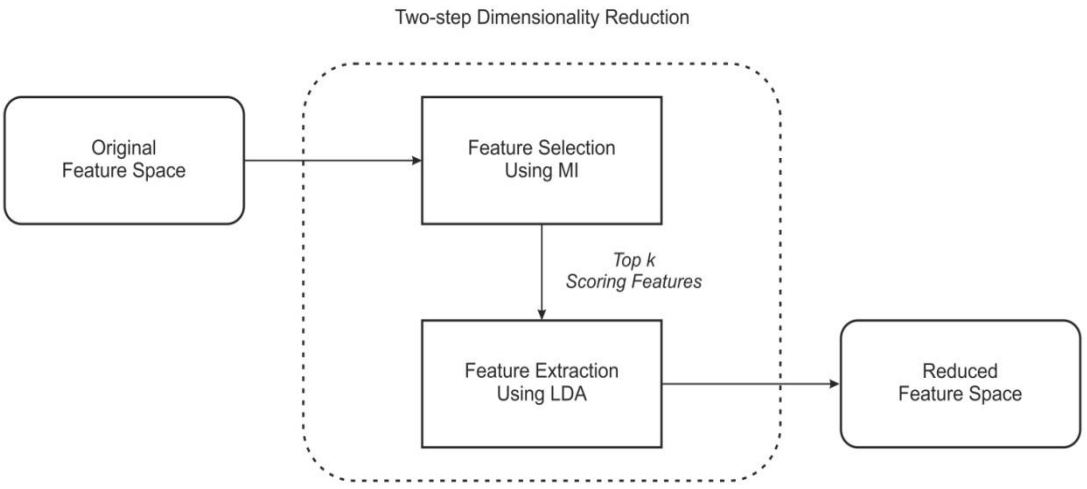

Figure 3.1: The two-step DR scheme.

3.3.4 Classifiers

The reduced feature space obtained after the application of the proposed DR scheme is supplied to the suitable classifiers for evaluation. The following three classification algorithms are employed: Logistic Regression, Support Vector Machines and Naïve Bayes.

Logistic Regression

Logistic Regression (LR) is a linear classification algorithm borrowed from the field of statistics. It passes the input data through the sigmoid function that returns a probability value which can be used to classify a particular data point to its corresponding class. A decision boundary or a threshold is computed for mapping the data points to their respective classes depending on the value of the sigmoid function. It is based upon the assumption that the probability of mapping a data point to a particular class can be obtained from the linear combination of the features of that point [128].

Support Vector Machines

Support Vector Machine (SVM) performs classification by computing decision hyperplane that demarcates instances of different classes. In an n-dimensional Euclidean space, a hyperplane is an n-1 dimensional subspace that separates the original space into two halves. The hyperplane is positioned in the hyperspace such that the distance between the different classes is maximized [128]. As SVM does not require the training examples to be transformed into different spaces, thus it can handle a very large feature subset. SVM has been used in several previous studies [130], [131] due its high classification accuracy [129].

Naïve Bayes

Naïve Bayes (NB) is a fast and simple probabilistic classifier. It is based on the Bayes theorem from probability theory that relies on prior knowledge about an event to predict the probability of occurrence of that event. It assumes that for a class, all the features are independent however practically it is not true [128]. Despite its naïve assumption about feature independence, the NB classifier gives a good competition in terms of performances to other classifiers. The easy implementation and low-cost training make NB a popular choice for classifiers [132], [133].

3.4 Experimental Setup

The main theme of our work is focused on identifying the criminal and illegal activities on the dark web. To match the theme, we selected five categories out of 31 of the dark web text dataset presented in the Section 3.2. The selected categories have content that is either illegal or of controversial nature that demands investigation and surveillance. Additionally, we include a sixth category called *Others* that holds a mixed content of other dark web categories.

The *Others* category may help better train the classification model by simulating the real-time practical situations where input data contains both the illegal and legal content. The details of the selected subset of categories are shown in Table 3.4. A total of 1006 samples distributed among six classes are taken for the experiment. It is to be noted that the distribution of samples in the categories is imbalanced as the difference between the largest and the smallest category is of more than 300 samples.

Table 3.4: Distribution of categories in the Dark Web Text Dataset.

Category	**Document Count**
CC Dumps	271
Counterfeits	37
Drugs	179
Forged Documents	40
Violence	97
Others	382
Total	**1006**

The proposed DR scheme is also evaluated on the Reuters-21,578 dataset, a benchmark dataset for the text classification problems. The dataset is a collection of documents on various topics from the Reuters news agency. The original version of the corpus has 135 categories and 21,578 samples [134]. As used in previous studies [135], [136], we are

taking only the top ten categories consisting of 9980 documents for our experiment. The distribution of the documents in the top ten categories is unbalanced as shown in Table 3.5.

Table 3.5: Distribution of categories in the Reuters-21,578 dataset.

Category	**Document Count**
Acquisition	2369
Corn	237
Crude	578
Earn	3964
Grain	582
Interest	478
Money-fx	717
Ship	286
Trade	486
Wheat	283
Total	**9980**

3.4.1 Evaluation Metrics

The performance of text classification is evaluated by Precision, Recall and Accuracy. Though accuracy can be used for an almost balanced dataset, it may not provide a good measure in the case of an unbalanced dataset (as in our case). On the other hand, a high value of precision is obtained at the cost of a low value of recall [137]. To get a balanced measurement, an *f-score* is used which is the harmonic mean of precision and recall [138].

For the problem of multi-class classification, a micro-averaged and macro averaged *f-score* is used [113]. For this, one binary *f-score* for each class is calculated and that class is represented as the positive class while all the other classes are treated as the negative classes. The micro averaged is calculated by using true positives, false positives and false negatives over all the classes. In this work, a micro averaged *f-score* is computed for evaluation.

3.4.2 Validation and Parameter Settings

The *K*-fold cross-validation method is used to validate the experimental results. In this method, samples are split into *K* mutually exclusive parts known as folds and the classification algorithm is executed for *K* rounds. At each round, the *K*-1 folds are used for training the classifier and the remaining part for testing. The final classification results are obtained by merging the individual results of each round. In this method, all the samples from the dataset get the chance to be in the training set and each sample is employed once to test the model. Therefore, we apply the 10- fold cross-validation with the three classifiers.

The proposed DR scheme and the classification algorithms are implemented in Python language and its associated library for machine learning Scikit-Learn[8]. During the construction of the feature vector with TF-IDF, the *N-grams* was set to unigrams only and all the terms with a frequency less than three were ignored. To obtain the optimal parameters for the classifiers, the grid search has been applied that performs the exhaustive search over the grid of parameter value and their possible combinations. The values of the parameters for the three classifiers were set as given in Table 3.6. The experiments were conducted on an Intel Core i5 1.6GHz machine running on Windows 8.1 operating system with 4GB of RAM.

Table 3.6: Parameter values for the three classifiers.

Classifier	Parameter
NB	Default
SVM	Kernel: linear Default values for other parameters
LR	C=10 Default values for other parameters

[8] https://scikit-learn.org/

3.5 Results and Discussion

The size of the feature space after the data preprocessing stage was 1886 and 15,665 for the dark web text dataset and the Reuters-21,578 dataset respectively. The first step of the DR was applied on both the datasets by calculating the MI score of each of the features. The top k percent of features are selected from the initial feature set. LDA is then applied in the next step to transform the subset of features into a new space. After the application of the two-step DR scheme, the resulting size of the feature space for the dark web text dataset and the Reuters-21,578 dataset is five and nine respectively (which is one less than the number of classes). Table 3.7 and Table 3.8 shows the classification performance on the dark web text dataset and the Reuters-21,578 when the proposed two-step DR scheme is applied at different values of k in the first step of feature selection.

Table 3.7: Classification performance on the Dark Web Text Dataset after the application of the two-step of DR.

Percentage of features selected in the first step (k)	Number of features	Classifier Performance (*f-score*)		
		NB	SVM	LR
10	189	93.63	93.63	93.33
11	207	93.53	93.83	93.93
12	226	94.73	94.73	94.63
13	245	95.72	95.42	95.61
14	264	95.32	95.12	95.92
15	283	95.52	95.92	96.22
16	302	95.93	95.62	96.42

In Table 3.7 we can see, with 16 percent of features selected in the first step, the NB classifier gives the best classification results. Also, SVM and LR attain the highest score when 15% and 16% of features are selected respectively. Therefore, we can say that the top 15-16 percent of features are the most discriminative as all three classifiers produced the best results in this range.

Table 3.8: Classification performance on the Reuters-21,578 after the application of the two-step of DR.

Percentage of features selected in the first step (k)	Number of features	Classifier Performance (*f-score*)		
		NB	SVM	LR
10	1566	88.84	89.01	89.62
11	1723	88.98	89.08	89.73
12	1880	89.06	89.02	89.67
13	2036	89.26	89.08	89.83
14	2193	89.42	88.82	89.96
15	2350	89.50	88.86	89.95
16	2506	89.65	88.91	90.04

For the Reuters-21,578 dataset, NB and LR achieve their highest *f-score* when the top 16% of features were selected in the first step. Though SVM achieves its best performance with only 11-13 percent of the feature is selected, it could not surpass the *f-score* of both the NB and LR classifiers, LR being the best among all three. Thus the LR classifier gives the best classification performance with the two-step DR technique

.

Table 3.9 compares the performance of the classifiers when no feature selection is applied, features selected from step one alone and with the two-step DR. The proposed approach achieves a significant improvement in the classification performance over the other.

Table 3.9: Comparison of the classification performance with different feature set sizes.

Dataset	Size of Feature Set	Classifier		
		NB	SVM	LR
Dark Web	Full Feature Set	83.17	91.32	91.63
	With MI (k=16)	73.14	82.81	87.56
	With MI-LDA	95.93	95.62	96.42
Reuters-21,578	Full Feature Set	79.50	84.14	81.25
	With MI (k=16)	70.65	85.68	82.11
	With MI-LDA	89.65	88.91	90.04

Now, we compare the proposed DR approach with the other feature extraction methods PCA and LSI when used in combination with MI on the two datasets. Table 3.10 shows the classification performance of the three classifiers with different feature extraction methods at 16% of the feature being selected using MI in the first step. The number of components in PCA and LSI were one less than the total classes as used in LDA. As evident from the results, the MI-LDA combination outperforms the other methods with a significant increase in the *f-score*. This increment in performance is complemented with the considerable reduction in the dimensionality of the feature space leading to lower computational cost as compared to other methods.

Table 3.10: Comparison of the two-step DR scheme with other feature extraction methods.

Dataset	Size of Feature Set	Classifier		
		NB	SVM	LR
Dark Web	MI-PCA	55.96	65.51	69.00
	MI-LSI	61.22	65.31	69.50
	MI-LDA	95.93	95.62	96.42
Reuters-21,578	MI-PCA	69.21	73.76	73.67
	MI-LSI	68.06	74.07	74.58
	MI-LDA	89.65	88.91	90.04

Figure 3.2 shows the learning curves for the logistic regression classifier when combined with the two-step DR technique. In Figure 3.2, at the beginning there is a narrow gap between the two curves depicting a bit of variance but the cross-validation score gradually improves as the training samples are increased. However, in the end, the training curve and validation curve are almost close to each other with a high score value which represents a reasonably good fitting of the model. Figure 3.3 shows the time taken (in second) to train the model with different sizes of training samples. Now, we compare the results obtained by the proposed methodology with baseline approach set forth in previous work [43].

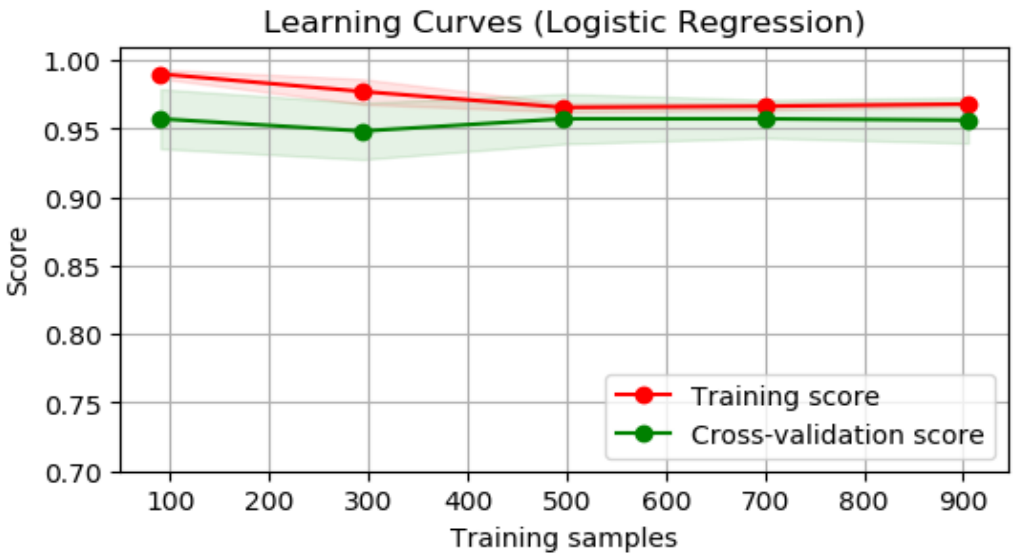

Figure 3.2: Learning Curves of the LR classifier with the proposed DR scheme.

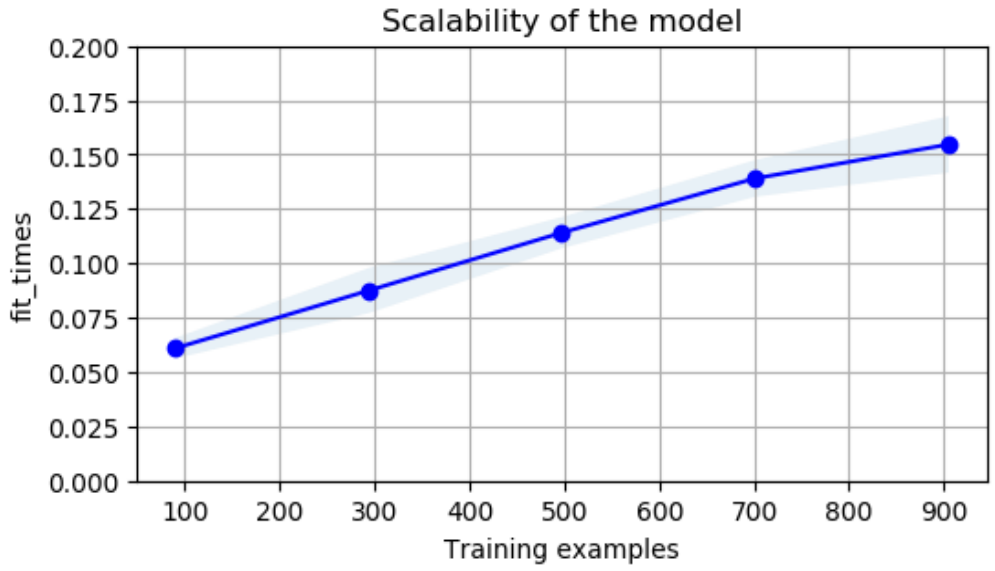

Figure 3.3: The time required for fitting the LR model.

The baseline method creates a pipeline for classification by identifying the key performance influencer in the method. They have used BoW and TF-IDF for document representation and three classifiers NB, LR and SVM for evaluation. The *f-score* of the three classifiers on the dark web text dataset is shown in Table 3.11 along with the result of our

method (at k =16%). The parameters of the classifiers and other settings of the experiment were the same as used in the baseline approach.

Table 3.11: Comparison of the proposed approach with the baseline approach.

Classifier	Feature Representation	Classifier Performance (*f-score*)	
		Baseline	Two-Step DR
NB	BoW	84.18	92.05
	TF-IDF	84.38	95.93
SVM	BoW	83.60	91.95
	TF-IDF	87.27	95.62
LR	BoW	86.67	91.76
	TF-IDF	90.35	96.42

As evident from the Table 3.11, the proposed DR approach has much better performance than the baseline approach with LR being the best among the three classifiers. Moreover, the better performance of our classification model is achieved at a much smaller feature set. The proposed classification model could prove beneficial to law enforcement agencies in predicting the illegal content on the Tor dark web network.

CHAPTER 4

CONTENT BASED IDENTIFICATION RANKING DRUGS HIDDEN SERVICES

4.1 Background

The emergence of cryptomarkets on the dark web has eased access to illicit drugs without compromising the privacy of the user. The ubiquitous presence of the narcotic substance on these marketplaces has lead to gradual ill effects on public health. The significant health issues with the use of narcotics have driven the government and law enforcement agencies into closing down marketplaces. However, a multitude of marketplaces and stand-alone shops dealing in illicit drugs requires much larger resources and time in their closure. Also, some of these avenues may be selling highly dangerous substances while some deal only in less harmful or even safe drugs (like *tobacco, khat*). Therefore, Law Enforcement Agencies (LEA) should prioritize their efforts in detecting the markets that sell the most harmful drugs. Note that this does not mean the other marketplaces are left untouched instead the LEA should focus their attention on marketplaces that need to be immediately monitored.

Ranking the marketplaces based on the amount of harm could help the LEA in identifying the influential marketplaces which subsequently helps in identifying the key vendors, customer profiles and the countries involved in the drug trafficking. Therefore, we aim to help the LEA by proposing the content based ranking scheme in detecting the most harmful hidden services including both cryptomarkets and stand-alone shops involved in the illicit drugs trade. For each of the hidden service (HS), we calculate a harm score that indicates the degree of threat posed by the illicit drugs on that HS. The harm score is then used to rank the HS. The ranking methodology can either be link based or content based or a combination of the former two.

The link based methods use the hyperlinks between the HS to rank them. However, the link based methods may not be efficient in ranking the dark web content as the majority of Tor HS web pages have low out-degree [62], even some of the Tor HS may have zero out-degree making them hard to find. Therefore, content based ranking methodology has an edge in the case of the dark web as they utilize the content of the HS rather than the hyperlinks. Also, the content base approach considers even the *isolated* (zero in-degree and out-degree) HS which are not taken into account by the link based methods.

4.2 Overview of the Drugs available on the Tor Network

The controlled drugs are classified into several categories based on their activity on the nervous system of the person. These drugs can be abused in several ways like swallowing, sniffing, smoking or injecting. The intensity of the effect depends on the type of the drug and the dose consumed. Here we briefly describe different categories of drugs and their effect on the consumer as mentioned by the US Drugs Enforcement Administration[9].

Stimulants: As the name suggests, these drugs stimulate the nervous system of the body to generate an instant sensation called *rush* or *flash*. It is generally abused to achieve the better physical and mental activity, boost self-esteem and produce a sense of exhilaration and prolonged wakefulness. When taken in large doses at a time may result in headaches, dizziness, chest pain, excessive vomiting and sweating. In case of overdoses, fatality may occur if not treated for symptoms.

Depressants: Depressants produce a calming effect on the mind and body that relieve anxiety and muscle spasms and induces sleep. Higher doses of depressants are abused by a person to experience euphoria. Long-term

[9] https://www.dea.gov/

intake of these drugs can cause physical and psychological dependence and tolerance.

Hallucinogen: Hallucinogens are capable of changing the mood and perception of a person. The psychological effects include distorted thoughts related to time and space, time appears to stand still and the person experiences hallucinations and flashbacks. The physical symptoms include increased heartbeat and blood pressure, dilated pupils sometimes followed by nausea and vomiting. The acute overdose of these drugs rarely causes death, however, severe doses may result in respiratory failure followed by death.

Narcotics (Opioids): Narcotics are primarily used to alleviate pain and dim human senses. A person feels calm, relaxed and drowsy by reducing anxiety, aggression and tension. The prolonged use of drugs results in psychological dependence and its withdrawal process can cause frustration, restlessness, depression, drug craving and other physical symptoms. Overdose of narcotics cause extreme drowsiness, confusions, slowed breathing that is often life-threatening

Inhalant: Inhalants are the volatile chemicals that are present in common house products like paints, correction ink, cleaning sprays etc. They produce psychoactive effects on the mind. It interferes with the common brain functions like thinking, hearing, vision and moving. Overdose of inhalants causes unconsciousness due to the accumulation of the toxic substance in the lungs leading to suffocation and death.

Steroid: Steroids are the synthetic replacement of the male hormone testosterone-responsible for the development of manly features like muscle growth, physical fitness and endurance, physical appearance and performance. Steroids are abused to achieve fast growth of muscles and

improved physical endurance. Their illicit intake has a number of side effects depending on the age and sex of the person though their overdoses are uncommon.

Cannabis: Cannabis or marijuana is a psychoactive drug extracted from different parts of the *Cannabis Sativa* plant including stems, leaves and flowers. Marijuana smoke can be inhaled directly to experience instant effects, however, it can also be eaten or drink. It affects different systems of the body and may cause impaired judgment, anxiety and accelerated heartbeat. The compounds found in marijuana can weaken the immune system of the body making it more vulnerable to diseases.

The drugs that are commonly available on the Tor HS that will be considered in our ranking are given in Table 4.1.

Table 4.1: Commonly available drugs on the Tor hidden services and their types.

Drug Class	Drug Type
Stimulants	Amphetamines Cocaine Crack Cocaine Khat Methamphetamine Mephedrone
Depressants	Benzodiazepines Gamma-Hydroxybutyric Acid (GHB)
Hallucinogen	Ecstasy Ketamines Lysergic Acid Diethylamide (LSD) Mushrooms
Narcotics (Opioids)	Buprenorphine Heroin Methadone
Inhalant	Butane
Steroid	Anabolic Steroids
Cannabis	Cannabis

4.3 Methodology

The proposed methodology is developed to target the domain of drug trafficking on the Tor dark web. The time complexity of the methodology can be reduced by providing it a pre-identified dataset of drug-related hidden services. For this, we can apply the classification technique presented in Chapter three for classifying dark web textual content. The proposed ranking methodology shall then be applied to extract the potentially harmful HS among other services. The proposed ranking methodology is graphically shown in Figure 4.1. It consists of the four steps: i) Data Preprocessing, ii) Illicit Drugs Name Extraction, iii) Calculation of Harm Score and iv) Final Rankings.

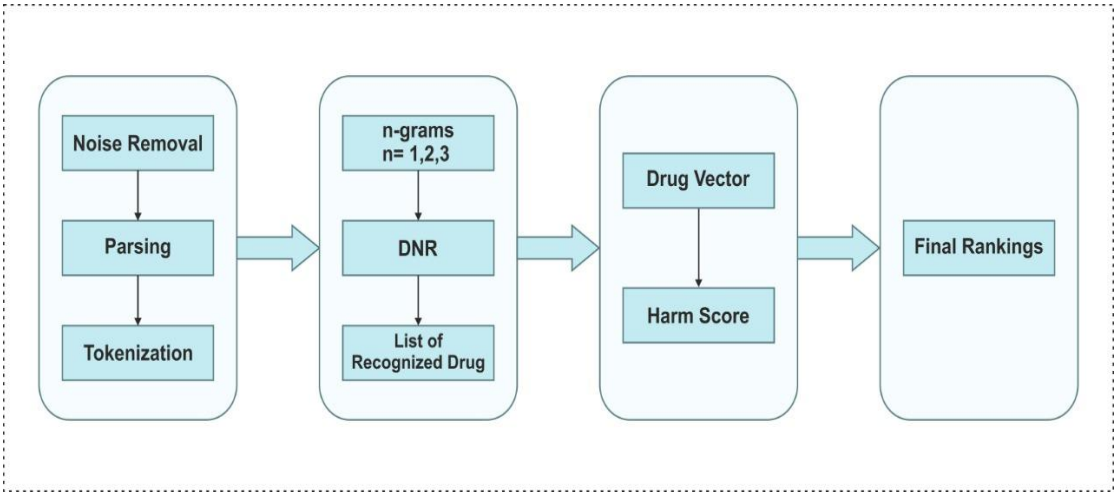

Figure 4.1: The proposed ranking methodology.

4.3.1 Data Preprocessing

The content of the HS is represented in the HTML file format. The product listings available on each of the HS are extracted by the customized Python script from the HTML files. These listings contain different products along with drugs from various vendors as the single HS (like cryptomarkets) may offer a large range of products. The textual content of all the product listings is extracted by removing the HTML tags and is saved in a separate plain text format for each individual HS. All the

unwanted content like the hyperlinks, URLs, script, meta-data, and white spaces are removed from the plain text file.

Stop words and duplicates are eliminated and the whole text is converted into the lower case. All the numbers consisting of either single-digit or more than three digits are identified and removed with the help of a regular expression parser. The relevance and need of removing such numbers shall be described in the following step. Finally, the tokenization is applied to the processed text to break the long sentences and phrases into smaller entities called tokens consisting of single words and numbers (with two or three digits only). Upon completion of the tokenization step, the obtained tokens are stored in a list format for subsequent steps.

4.3.2 Illicit Drugs Name Extraction

The list of tokens obtained from the data preprocessing steps contains the names of illicit drugs and also the name of other products that are irrelevant and unwanted in our task and needs to be eliminated from the token lists. Drug Name Recognition (DNR) could be used to identify and extract the name of drug-related items. DNR is a special type of Named Entity Recognition (NER) to identify and extract the drug names from the unstructured text [139].

DNR is a crucial and challenging step in our ranking methodology for extracting the names of illicit drugs. The customers and vendors involved in drug trafficking often use street names and slang for the prohibited drugs to confuse the law enforcement agencies. These slangs and street names consist of common words and English phrases used in daily life that can easily trick a person unaware of drug terminology. For example, *nuggets* is a street name used for crack cocaine. The participants may also use the commercial names of the drugs (like brand names) instead of the generic names. Moreover, the drug names may also be

represented in numerical forms like *501, 77* that motivate us in keeping the two and three digits numeric tokens in the data preprocessing step.

DNR Approaches

There are four categories of DNR approaches: dictionary-based, rule-based, machine learning-based and hybrid approach [139].

Dictionary-Based: This approach matches the given text against a particular drug dictionary to identify drug names. A drug dictionary is an automatically or manually collected database of drug names from a single or multiple resources. The text can be matched exactly or approximately to obtain optimal performance.

Rule-Based: The rule-based approach use predefined rules that reflect the patterns or naming convention of drug names. A rule-based approach can identify drugs that are named after adopting certain standard rules laid down by the concerned organization or agencies. This approach cannot identify drug names that are generated without following any existing patterns or rules.

Machine learning-Based: The machine learning approach considers the DNR as a classification problem where a text token has some distinct attributes or features that are used by the classification model to predict the class of the token. The performance of this approach is dependent on the choice of the learning model to be used and the quality of the feature set used by the model.

Hybrid Approach: The hybrid approach integrates multiple other approaches to garner their advantages while minimizing the limitations of the individual approaches. The final outcome of a hybrid approach could be better than that of the individual approaches.

In our methodology, we shall be applying the dictionary-based approach to identify the name and slangs of illicit drugs as this approach is best suited in the context of the problem under study. As already mentioned, slang terms are mostly used by the participants of the illicit drug business with no standard nomenclature or convention. Therefore, the rule-based approach may be ineffective in the present scenario, while the dictionary-based approach could be more effective provided a good dictionary with a comprehensive collection of slang/street names of drugs for matching with the text. Needless to say, the drug dictionary should be updated frequently to cope with the ever-changing environment of drug trafficking. The United States Drug Enforcement Administration provides the dictionary of slang/street names of the drugs and is regularly updated. We shall use the same dictionary in our work to extract the drug names.

To identify the drug's name and slang consisting of two or three words, the bigrams and trigrams are created from the token list of product listings. Thus, for each of the HS, a final list of tokens containing unigrams, bigrams and trigrams is generated. For each of the HS, an associative array with a *key-value* format is created to store the types of drugs and their frequency that are available on a particular HS. The *key* field of the associative array contains the drug type and the *value* field holds the corresponding frequency. The *value* field is initialized to zero. The procedure for creating the associative array is given by Algorithm 1.

The final list of tokens is matched against the drug dictionary to extract the drug names. Since the product listings may contain typos; it can affect the matching of tokens in the dictionary. The solution to this problem is to calculate the Levenshtein distance [140] of a token to measure its similarity with the drug dictionary. The Levenshtein distance between the two strings is the measure of the number of the characters that need to be changed to transform one string into another.

Algorithm 1: Building an associative array of the HS

Input: D: drug dictionary
T: list of n tokens
L: list of m available product listings
$@lev$: function that computes the Levenshtein ratio of the two strings
Output: A: Associative array

begin algorithm
for $i = 1$ to n **do**
    $flag = 0$
    for *dname* in D **do**
        Apply $@lev$ to compute the Levenshtein ratio R_i of token $T[i]$ and drug name *dname*
        if $R_i < 0.25$ **then**
            $k = dname[class]$
            $flag = 1$
            break
    end for
    for $j = 1$ to m **do**
        if $T[i]$ in $L[j]$ and $flag = 1$, **then**
            $A[k] = A[k] + 1$
            Remove $L[j]$
    end for
    $flag = 0$
end for
return A
end algorithm

In our case, we consider two tokens to be representing the same entity if their Levenshtein distance is not greater than 24 percent. The corresponding drug type is retrieved from the dictionary if the token is matched exactly or the Levenshtein distance is below 25 percent with the drug dictionary. Once the token is matched in the dictionary, it is then

searched in the available product listings extracted from the HS, to identify other listings of the same drug from different vendors. The number of listings that are matched with the token is counted and saved in the *value* field of the associative array which shall be indicating the frequency of the corresponding drug type of the matched token.

The listings that get matched with the token are eliminated from the set of available product listings of the HS. This process is repeated for each of the items in the token list that matched with the drug dictionary. However, if the token is matched with the drug type already existing in the associative array, then the frequency of that drug is incremented based on the number of matches found. After all the items in the token lists are matched with the drug dictionary, the set of available product listings is discarded. The elements of the associative array having zero in their *value* field are deleted.

At the end of this step, an associative array for each of the HS is obtained reflecting the drug types and their frequency. This final associative array is passed on to the third step for computing the harm score of the HS. Figure 4.2 shows one such example of the associative array thus created of an HS that sells *cannabis, cocaine, ecstasy and LSD.*

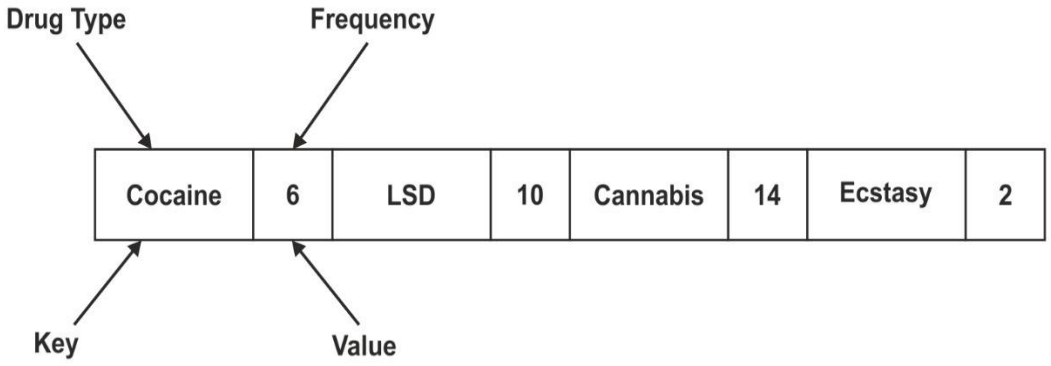

Figure 4.2: An example of the associative array.

4.3.3 Harm Score of Hidden Services

The serious health concerns that drug abuse causes on person and society have mandated the formulation of drug control and scheduling policies by the concerned agencies. Drug abuse can have several consequences on different aspects like personal, social, environmental and economic damage. The harm caused by the controlled drugs should be estimated in such a way that it accounts for all the aspects of damage. One such study was conducted to assess the harm of the controlled drugs in terms of multiple factors to aid the policymakers in devising better policies [141].

The authors of the study have formed a committee to estimate the harm of twenty drugs based on sixteen factors using the multi-criteria decision analysis technique. The committee members were renowned experts of drugs from the United Kingdom. The criteria or factors were divided into two major categories: criteria affecting the users and criteria affecting the others. The major categories were further divided into several sub-categories. The list of criteria affecting users were: Drug-specific mortality, Drug-related mortality, Drug-specific damage, Drug-related damage, Dependence, Drug-specific impairment of mental functioning, Drug-related impairment of mental functioning, Loss of tangibles and Loss of relationships.

The lists of criteria affecting others were: Injury (to others), Crime, Environmental damage, Family adversities, International damage, Economic cost and Community. The committee has assigned an overall score based on sixteen criteria to each of the twenty drugs on a scale of 0 to 100. A score of 0 means the drug is least harmful while a score of 100 indicates that the drug is most harmful. Table 4.2 shows the name of the drugs and their corresponding harm score estimated by the group of experts based on sixteen factors.

Table 4.2: Drug types and their harm score.

S.No (k)	Drug Type (d_k)	Individual Harm Score $t(d_k)$
1.	Heroin	55
2.	Crack Cocaine	54
3.	Methamphetamine	33
4.	Cocaine	27
5.	Amphetamines	23
6.	Cannabis	20
7.	Gamma-Hydroxybutyric Acid (GHB)	19
8.	Benzodiazepines	15
9.	Ketamines	15
10.	Methadone	14
11.	Mephedrone	13
12.	Butane	11
13.	Anabolic Steroids	10
14.	Khat	9
15.	Ecstasy	9
16.	Lysergic Acid Diethylamide (LSD)	7
17.	Buprenorphine	7
18.	Mushrooms	6

The overall harm score of the HS is calculated using the harm score of the individual drugs presented in Table 4.2 [141]. From the twenty drugs used in the above study, we skip the *alcohol* and *tobacco* from the study as these two substances are not controlled and have no relevance in the illicit drug trade.

Let A_i be the associative array of the i^{th} HS H_i, a drug vector V_i of dimension $n\,(n = 18)$ is created for H_i. The elements $x_i^k\ (k = 1,2, \dots 18)$ of V_i are obtained using Equation (4.1).

$$x_i^k = \begin{cases} t(d_k) * f(d_k),\ if\ d_k\ \in\ A_i \\ 0,\ otherwise \end{cases} \tag{4.1}$$

d_k and $t(d_k)$ are the drug type and its individual harm score obtained from Table 4.2, $f(d_k)$ is the frequency of d_k in the associative array A_i.

A harm score $\tau(H_i)$ is assigned to H_i using logarithm function given by Equation (4.2).

$$\tau(H_i) = log_{10}\left(1 + \frac{|V_i|}{v_o}\right) \tag{4.2}$$

where $|V_i| = \sum_k x_i^k$ and $v_o = \min[\, t(d_k)]$

A large HS like cryptomarket may offer numerous products with multiple listings from vendors, $|V_i|$ may get a very large numeric value. Therefore, the logarithm function is used to calculate the harm score to conveniently express the large values. Since the logarithm function is undefined for zero, we have added one in the argument of logarithm function in Equation (4.2) for conditions when $|V_i| = 0$. A HS which is not involved in illegal drug trafficking will have $|V| = 0$ and subsequently, an overall zero harm score reflecting that it does not have any negative effects on its users. The HS are ranked once the harm score for each of the hidden services is obtained.

4.3.4 Ranking

The ranking of the HS is needed to identify the most severe one among other HS as there would cases where some HS offers potentially harmful drugs while others may deal in much less harmful drugs. The HS is ranked in the decreasing order of their harm score with the top rank being occupied by the HS possessing the highest harm score. If two HS have equal harm scores then the HS containing drug with the highest individual score is placed at the higher rank. For example, consider two HS A and B, A offers *crack cocaine* (individual harm score of 54) from a single vendor while B offers *cocaine* (individual harm score of 27) from two different vendors. The harm score of both A and B shall be the same but A shall get a higher rank than B because A offers *crack cocaine* whose individual harm score is 54.

4.4 Experimental Setup

This section describes the implementation of the proposed ranking technique and evaluates the results on the standard metrics for the problem of rankings. The rankings generated by our proposed methodology need to be compared with some benchmark or ground truth rankings of HS, however, to the best of our knowledge, no such study exists that put the content based ranking of HS. Therefore, to assess the accuracy of the rankings produced by the proposed methodology, the ranking of the HS from the three experts is being taken as the ground truth for benchmarking.

4.4.1 Dataset

The dark web text dataset is used in this work that was presented in the chapter three. The dataset consists of 4,102 labeled instances divided into 31 categories including drugs. The individual instance represents a Tor HS with its root page in the HTML format. We have picked up 179 HS samples from the dataset that were related to the illicit drug trades. The data from the HS are extracted and processed for further steps. The content of the HS that shall be used in the implementation of the experiments was in English.

4.4.2 Expert Ranking-Ground Truth

A group of three experts was formed and the dataset of HS was presented before them to generate the ground truth ranking. The three experts belonged to different domains related to the area of drug abuse and its impact. The experts were a psychologist, a professional medical doctor and an academician. The experts independently ranked each of the HS in the dataset based on the three criteria i.e. i) presence of illicit drugs on the HS, ii) the severity of the present drugs and iii) the number of listings of each individual drug. The three separate rankings are obtained from the individual expert.

Now, the three rankings need to be merged together to obtain a single ranking for the benchmarking purpose. Since, all the three rankings would vary given the subjective nature of the task and obvious differences in opinion regarding the drug harms among the experts. Therefore, rank-based aggregation (RBA) [142] shall be used to combine the individual rankings to obtain final rankings. RBA removes the biases in rankings by compensating the noise that appeared by the other experts during aggregation. The final rankings generated after applying RBA shall be served as ground truth for assessing the accuracy of our proposed ranking methodology.

4.4.3 Evaluation Metrics

The following standard metrics were employed to judge the accuracy of the rankings generated by our proposed methodology.

Kendall's tau

Kendall's tau is a widely used metric to compare the pair of rankings in the field of information retrieval [143]. Given the two rankings P and Q of length n, Kendall's tau of P and Q is given by Equation (4.3).

$$R(P,Q) = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{4.3}$$

where n_c and n_d be the number of concordant and discordant pairs respectively, the pair of ranks are called concordant if $p_i < p_j$ and $q_i < q_j$ or $p_i > p_j$ and $q_i > q_j$ and discordant if $p_i < p_j$ and $q_i > q_j$ or $p_i > p_j$ and $q_i < q_j$ for each pair of rank (p_i, q_i), $(p_j q_j) \ldots\ldots\ldots (p_n, q_n)$.

Kendall's tau has been selected because it is commonly used in the literature [144-147] and is more robust than other standard metrics [148].

Rank-Biased Overlap

Rank-biased overlap (RBO) works by assigning different weights to different ranks in the list with higher ranks getting more weightage [149]. Thus, RBO can be used to measure the accuracy in top ranks by giving more importance to the higher ranks. The RBO of two rankings P and Q of length n is given by Equation (4.4).

$$RBO(P,Q,p) = (1-p) \times \sum_{d=1}^{r} p^{d-1} A(P,Q,d) \quad (4.4)$$

where r is the number of unique ranks, p is a configurable parameter in (0,1) such that the smaller value of p implies that the metric is more top-weighted and $A(P,Q,d)$ is the overlap value between two rankings P and Q up to rank d, given by Equation (4.5)

$$A(P,Q,d) = \frac{|P_{1:d} \cap Q_{1:d}|}{|P_{1:d} \cup Q_{1:d}|} \quad (4.5)$$

RBO can measure the accuracy of the proposed methodology in high ranks similar to other metrics [150], therefore it has been used for evaluation purposes as the top-ranked HS would be more important in terms of law enforcement perspective. Also, rank-biased overlap is efficient than other evaluation metrics in handling the non-conjointness in the rankings [149].

4.5 Results and Discussion

The experiment was implemented using Python v3.6 [151] on a Windows 8.1 operating system running on an Intel i5 machine supported by 4 GB of RAM. The intermediate steps of the data preprocessing stage including the elimination of stop words was achieved by the Python Natural Language Toolkit (NLTK) package. Equations (4.1) and (4.2) were used to compute the harm score of the HS.

Table 4.3 shows the top ten ranked HS from the ground truth rankings and the one generated by the proposed method respectively. The English alphabets are used to denote the HS instead of their onion URL for the sake of simplicity. The data in Table 4.3 clearly reflects the performance of the proposed method as the highest-ranked HS obtained is the same as obtained from the ground truth of experts rankings. This HS was involved in the trading of eleven controlled drugs.

Table 4.3: The top ten HS retrieved from the ground truth and the proposed ranking methodology respectively.

HS	Retrieved Rank	Ground Truth	Difference
A	1	1	0
B	2	2	0
C	3	3	0
D	4	4	0
E	5	5	0
F	6	6	0
G	7	7	0
H	8	8	0
I	9	10	1
J	10	9	1

Now we calculate Kendall's tau between the two rankings to measure the correctness of ranking allotted to each of the HS. Since the nature of our problem demands that the top-ranked HS are of greater importance to the law enforcement agencies; hence we compute Kendall's tau up to the rank 50 of the list with the interval of 10 i.e. find Kendall's tau up to rank k where $k \in \{10,20,30,40,50\}$.

Table 4.4 shows the value of Kendall's tau for different values of k. The data in Table 4.4 indicates the effectiveness of the proposed ranking methodology when compared with the ground truth. The near to one value of Kendall's tau indicates the strong relationship between the two rankings, though a slight decrease was observed in Kendall's tau as k

increases. The top ten ranked HS are the dealer of potentially harmful drugs and may hold a key space in the drug trade. The law enforcement agencies should put their resources in investigating these HS.

Table 4.4: Kendall's tau between the two ranked lists at different values of *k*.

k	Kendall's tau
10	0.9556
20	0.9474
30	0.9218
40	0.9120
50	0.8784

To evaluate the consistency of the proposed method, we randomly selected the six samples of the rank pairs from the complete list and computed their Kendall's tau. The sample size is 25. Table 4.5 shows Kendall's tau of the different samples. Once again, the high values of Kendall's tau confirm that the proposed methodology is close to the ground truth throughout the entire ranking list.

Table 4.5: Kendall's tau value for randomly selected samples from the ranked list.

Sample #	Kendall's tau
1	0.9318
2	0.9121
3	0.8872
4	0.9058
5	0.9255
6	0.9485

Now we check the accuracy of the proposed method in the high ranks by calculating the RBO of the two ranking lists. Figure 4.3 represents the RBO between the two lists. The curve of the graph touches to one which shows the high accuracy of the proposed method in top ranks.

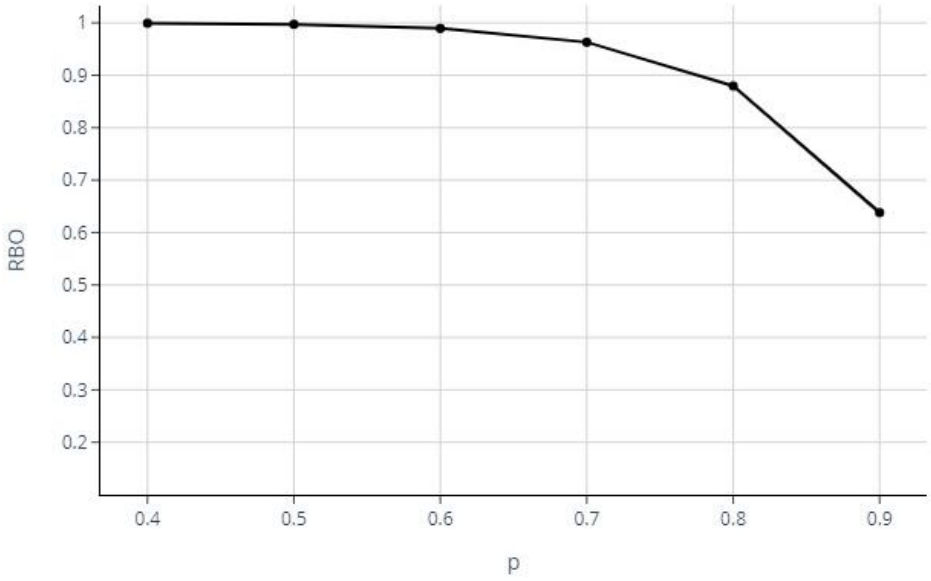

Figure 4.3: The RBO curve for the two ranking lists.

CHAPTER 5

CONTENT BASED IDENTIFICATION PREDICTING FIREARM LISTINGS

5.1 Background

The probable use of the dark web for the procurement of illegal firearms has been a debatable topic for a long until recently when several studies have uncovered the presence of firearms trafficking on the anonymous platform [82-84]. This dark corner got the spotlight following the Paris and Munich terrorist attacks in 2015 and 2016 respectively after several media houses reported the linking of the weapons used to the dark web platform. There were also reported cases of firearms being bought for personal use from the dark web [152].

The firearms on the dark web could be used both for organized terror attacks as well as for private use with the customer being either an individual or a group. Moreover, a study reported that small firearms like handguns and rifles were used in the majority of mass shootings [153]. For instance, *Glock 17*, a type of pistol was used in the Munich shooting while assault rifles were used in the Paris attacks. The popularity of these arms is also complemented by the fact that the most commonly available firearms on the Tor dark web cryptomarkets include pistols and rifles [82], [83].

The goal of this study is to provide law enforcement, monitoring and investigating agencies with an automatic tool to detect the firearm listings of handguns and rifles using text analysis and machine learning techniques. This may help the concerned agencies in easy and quick detection of hidden services involved in the illegal firearm trafficking and subsequent actions to be taken. The identification of specific types of firearms could also aid in identifying the vendors of the firearms that

operate across several hidden services on the Tor dark web. To achieve this task, an ensemble machine learning approach combined with a Part-of-Speech tagged and N-grams features is applied to detect the listings of pistols and rifles.

5.2 Overview of the Firearms available on the Tor Network

The *Silk Road* marketplace has been regarded as the first ever hidden service to offer firearms including weapons of mass destruction [18]. Later on, *The Armory* has come up that exclusively provides a variety of firearms and related products with worldwide delivery services. Here we briefly describe the firearms and other weapon-related products that are commonly sold on the Tor dark web.

Firearms

A firearm is a barreled weapon capable of discharging a bullet from a distance with ability to kill the person being fired upon. The firearm can be described in several different ways like by their caliber indicating the diameter of the bore of the gun or the mode of action like the pump, revolver, muzzleloader, semi-automatic, fully-automatic, etc. The firearm may be designed to be used as a hand-held or in a mounted position. A firearm is designed for individual use and comes in different sizes as defined as follows:

Handguns: Handguns are the smallest of firearms having short-length barrels specifically designed to be used single-handedly. Their small size makes them easily portable and concealable from law enforcement agencies. They are mainly of two types: semi-automatic pistols and revolvers. The semi-automatic pistols have a single chamber and barrel. At each trigger pressed, the handgun fires bullet from the chamber, throws out the bullet casing and loads a new bullet from the magazine. Thus the handguns fire only a single round for pressing the trigger once. Revolvers

on the other hand have multiple chambers fitted in a cylinder. At every trigger pressed, the cylinder rotates to align the next chamber in line of bore to fire the shot. Revolver can shot a bullet multiple times without reloading. For example, the Glock pistol series like *Glock 17, Glock 26, Colt Revolver,* etc.

Shotguns: Shotguns have a larger barrel length as compared to handguns and require both arms to properly handle and fire the shot. The shotgun can fire either a shell that contains multiple small pellets or a single solid bullet called a *slug*. Shotguns can have one or two barrels and can either be loaded automatically or manually. The manual loading of the shotguns can be performed by either of the following methods: pump-action, breech-loading, revolving actions and lever action. They also have much powerful action than handguns. The automatic loading of shotguns functions in the same way as that of handguns. Shotguns are commonly used in sport and hunting activities and by peoples for home security. For example, *Mossberg* shotgun.

Rifles: Rifles also known as long guns have the longest barrel length. The barrels have spiral grooves cut into the inner lining that makes the bullet spin in the barrel to get enhanced precision. The walls of the barrels are thick to withstand high pressures and are generally fired at stationary targets. The bore of the rifle is designed to accommodate ammunition of specified caliber only. The rifles can be distinguished much like the shotguns into the manual or automatic category. The manual loading follows similar methods as for shotguns except for an additional bolt action loading. The automatic loading rifles are also known as assault rifles for their high firing rate with manageable size and portability. Rifles are known for their versatility that comes with good range and accuracy. For example, *Garand M 1841, Ruger M77, G36 Assault Rifle* are some of the models of rifle available on the Tor network.

Digital Products

Digital products are the popular products only next to the firearms in weapon trafficking on the Tor dark web [82]. It includes e-books and guides that provide step by step method to create explosives using home items, manufacturing and customizing firearms and accessories as per specific requirements, manufacturing firearms with 3D printing technology. The advancement in the field of 3D printers has eased the manufacturing of homemade firearms thereby posing a major security concern.

Accessories and other weapons

Firearm accessories like magazines, cartridges, gun mounters, armors, ballistic vests, etc can be bought from the dark web. Other weapons include ammunition like bullets and grenades, non-lethal weapons like tasers, stun guns, batons, pepper sprays, etc are also available. Melee weapons like knives, knuckle dusters, blades can also be bought.

In the present work, we shall be proposing a methodology to detect the listings of handguns (pistols) and rifles on the Tor dark web. Henceforth, the term firearm shall refer to pistols and rifles in the subsequent sections.

5.3 Methodology

The existing work on the Tor dark web firearm trafficking was of exploratory nature that intends to estimate the market value of firearm trafficking, identifying the variety of firearm and associated products available and figuring out the vendor profiles and their characteristics. In this work, we shall present the methodology for the automatic detection of firearm listings on the Tor hidden services using the ensemble machine learning models.

The effectiveness of the proposed solution is evaluated on the public datasets using standard evaluation metrics. The proposed methodology for detecting the firearm listings is composed of three steps: i) Data Preprocessing, ii) Feature Construction and iii) Detection of Firearm Listings. The diagrammatic representation of the proposed methodology is presented in Figure 5.1.

5.3.1 Data Preprocessing

The data preprocessing step is an essential task whose objective is to eliminate the irrelevant, redundant and noisy data before moving on to the subsequent stages. It helps in reducing the computational time of the model and enhancement in the overall performance. The preprocessing of the data is achieved by the following three tasks: Tokenization, Stop Words Removal and Noise Removal.

Tokenization

The textual data may consist of sentences and phrases that may not be suitable for direct application to the model. Therefore, tokenization is performed on the text data that breaks them into individual words and numbers called tokens. The white space characters are used as a delimiter for breaking the text into individual tokens.

Stop Words Removal

In the English language (or any other spoken language), there are certain words like *is, am, are, the, of*, etc whose removal from the text does not change the meaning or context of the text, such words are called the stop words. The stop words frequently appear in textual data, therefore their elimination from the text may result in a considerable reduction in the size of the data which may, in turn, improve the efficiency of the classification model. The stop words are eliminated with the help of the predefined list of stop words for a specific language.

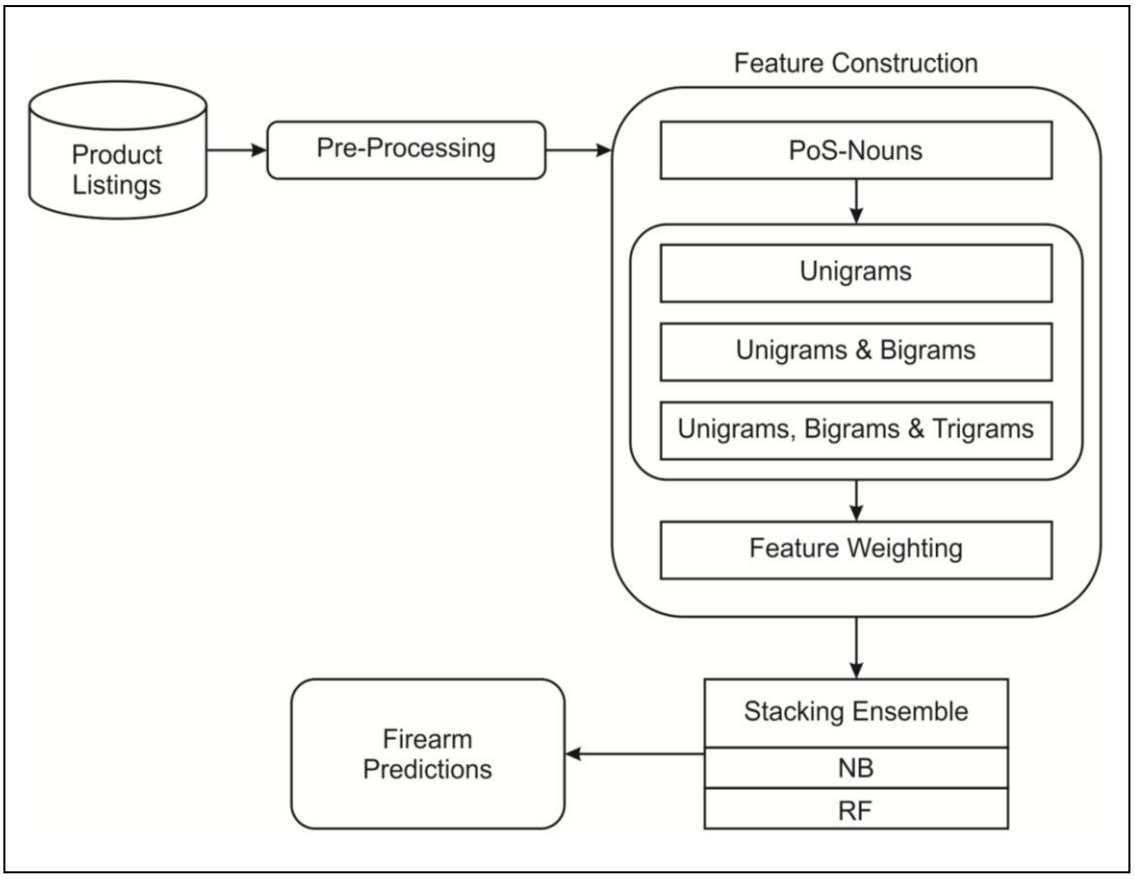

Figure 5.1: The proposed methodology.

Noise Removal

The noise in the data refers to the components that do not bear any relation to the data under study. In our data of product listings, the noisy parts are special characters (like $ or #), punctuation marks (like a full stop, comma, exclamation mark), URLs and e-mail addresses. The noise does not convey any meaning to the model, therefore, it should be removed from the data. Moreover, noise removal reduces the dimension of the feature set.

After the application of the above three tasks, the resulting text is transformed into the lower case to ensure uniformity across the dataset. The preprocessed data is then passed onto the next stages for further action.

5.3.2 Feature Construction

The data received from the preceding step contains unstructured text that needs to be transformed into an appropriate feature vector that can be supplied to the classification model. A variety of the feature construction methods like the *Bag of Words (BoW)* model, *Part of Speech (PoS)* tagged features, N-grams, heuristic patterns, contextual and semantic features and parse trees exist that have been employed in the text classification literature. In this methodology, the following steps are undertaken for adequate feature selection and representation.

Part of Speech (PoS) Tagging

Part of Speech (PoS) tagging is the process by which individual words or tokens are assigned a category of part of speech i.e. noun, verb, adverb, etc. By this process, each of the tokens in the text is given a tag that identifies whether it is a noun, adjective and so on. Here in our work, we have selected the non-numeric tokens that have been tagged as nouns and all other numeric and alpha-numeric tokens. This is done as the listings of weapons include the characteristics of a firearm like its caliber, action type, finesse, the manufacturer of the firearm, model variant, etc. which are all nouns. Therefore, all the noun-tagged tokens were extracted for further steps to observe the effect of selecting such tokens on the result. A sample of a pistol listing from the Tor marketplace is given in the next paragraph. Table 5.1 shows the PoS tagging of some of the tokens of the pistol listing.

"*The Glock 17 Gen4, in 9x19, introduces revolutionary design changes to the world's most popular pistol. The Modular Back Strap design lets you instantly customize its grip to adapt to an individual shooter's hand size. The surface of the frame employs the new scientifically designed, real-world-tested, Gen4 rough textured technology. Internally, the new Glock dual recoil spring assembly*

substantially increases the life of the system. A reversible enlarged magazine catch, changeable in seconds, accommodates left or right-handed operators. The Tundra is a dry suppressor, but can also be used with a small amount of coolant for even greater flash and sound reduction. Its light weight and relatively small size makes for a pleasant, accurate shooting experience as compared to nose-heavy suppressors."

Table 5.1: An example of PoS tagging of a pistol listing.

Token	PoS Tag
The	Determiner
Glock	Proper Noun
17	Cardinal Number
pistol	Noun
of	Preposition
frame	Noun
recoil	Noun
spring	Noun
substantially	Adverb
used	Verb
and	Conjunction
suppressors	Noun, plural

Extraction of N-grams

N-gram is a string of consecutive tokens from the text where N indicates the number of tokens in the string. The most commonly generated N-grams are unigram (N=1), bigram (N=2) and trigram (N=3). The N-gram is a popularly used technique and has been applied in different fields [154]. In our work, the product listings in the dataset contain the product name, their brand and other identifying attributes (like *fully automatic*) that may be composed of two or three words. Hence, we shall use the combination of N-grams (N=1, 2, 3) to adequately

represent the feature vector. The example of an N-gram model is shown in Table 5.2.

Table 5.2: An example of the N-gram models.

N-gram model	Beretta M9 is semi automatic pistol
Unigram	Beretta, M9, is, semi, automatic, pistol
Bigram	Beretta M9, M9 is, is semi, semi automatic, automatic pistol
Trigram	Beretta M9 is, M9 is semi, is semi automatic, semi automatic pistol

Feature Representation and Weighting

The Bag of Words (BoW) scheme is used to represent the feature vector. In the BoW model, the vocabulary consisting of the distinct features from the dataset is created. The element of the feature vector in a BoW model only indicate whether the corresponding feature is present in the document or not and it does not count the frequency of feature in the document. Therefore, the Term Frequency-Inverse Document Frequency (TF-IDF) [113] is used to assign weight to each of the features reflecting its corresponding frequency in the document. As discussed in Chapter three, a high weight is assigned by the TF-IDF to the feature which frequently occurs in a document but infrequently in the other documents of the dataset.

5.3.3 Detection of Firearm Listings

The classification models and the ensemble model for detecting the firearm listing are discussed in this section as follows.

To predict the firearm listings, we have utilized the three machine learning algorithms that are: Naïve Bayes (NB), Random Forest (RF) and Logistic Regression (LR). These classifiers have been selected because of their track record of better classification performance. The description of the base classifiers and techniques to build their ensemble is given as follows:

Naïve Bayes

Naïve Bayes (NB) is a probabilistic supervised classifier based on the Bayes theorem. It makes a naïve assumption that the features are conditionally independent of each other [128]. NB predicts the probability of an event by calculating the joint probability with respect to the occurrence of the other event. NB is among one of the popular classification algorithms and has been used in a number of studies with good classification performance [132], [133].

Random Forest

Random Forest (RF) classifier is based on the decision tree algorithms [155]. It is an ensemble of multiple individual decision trees that form a forest. Multiple samples from the training dataset are chosen to generate each decision tree. A final class value is chosen from the trees based on the majority voting scheme. RF can effectively manage the missing values and good parameter tuning can prevent them from over fitting [156].

Logistic Regression

Logistic Regression (LR) is a linear classification algorithm borrowed from the field of statistics. It passes the input data through the sigmoid function that returns a probability value which can be used to classify a particular data point to its corresponding class. A decision boundary or a threshold is computed for mapping the data points to their respective classes depending on the value of the sigmoid function. It is based upon the assumption that the probability of mapping a data point to a particular class can be obtained from the linear combination of the features of that point [128].

The ensemble technique leverages the power of multiple individual classifiers called base classifiers in order to obtain improved

classification performance as compared to a single base classifier. It combines the weak classifier with a strong classifier to get an overall better classification output. The base classifier can be combined in two different ways to form the ensemble: *Majority Vote* and *Stacking* [128].

Majority Vote

The majority voting ensemble classifier combines the predictions of the several base classifiers when they are applied in a parallel scheme. The final outcome is obtained by selecting the prediction that gets the highest number of votes among the predictions of several multiple classifiers. The final outcome shall be used to classify the instance to the target class.

Stacking

Stacking is similar to the majority voting ensemble except that the base classifiers are executed in a sequential manner in stacking instead of the parallel manner as in the majority vote. In stacking, the base classifiers are placed one over the other resembling stack data structure such that each of the base classifiers passes its prediction to the classifier above it. The multiple base classifiers are combined and are trained on a single data. The predictions of the base classifiers are used as the input for the final estimator called meta-classifier whose output is considered as the final prediction of the stacking ensemble.

The stacking ensemble is more suitable than the majority voting as the former is set apart by the model learning that occur at the final level by the meta-classifier, which is absent in the voting scheme [157]. Also, the second classifier in the stacking ensemble can be trained to possibly learn the error committed by the first classifier. Consequently, the final predictions can be improved upon incorporating the error estimates to the predictions of the first classifier [158]. Therefore, we shall apply the

stacking method to build our ensemble model with NB and RF as base classifier and LR being the meta-classifier. Figure 5.2 illustrates the concept of a stacking ensemble model.

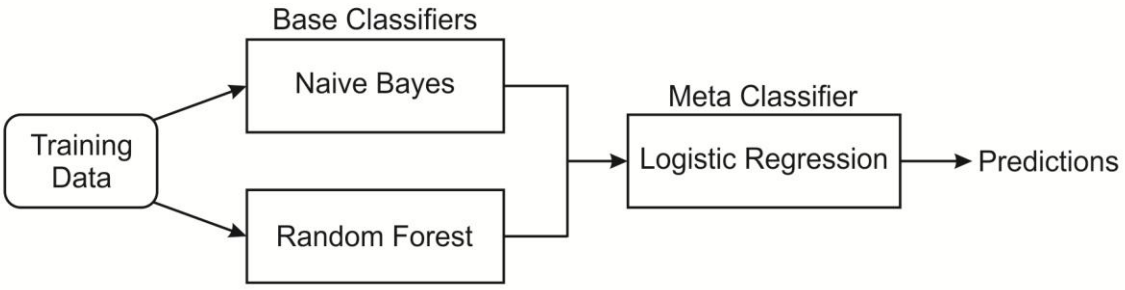

Figure 5.2: The Stacking Ensemble model.

5.4 Experimental Setup

This section describes the experimental settings required to implement the proposed ensemble classification model followed by the standard evaluation metrics for testing the effectiveness of the proposed model.

5.4.1 Dataset

The effective evaluation of the proposed method requires a dataset with an ample number of training and testing instances. Therefore, the experiment was performed on a public dataset collected by an independent researcher Gwern Branwen [159] and has been used in various existing peer-reviewed literature [36], [160], [161]. The dataset contains the digital traces of the 87 marketplaces and their related 37 discussion forums that were collected over a period of two years. For our work, we have selected the subset of the dataset comprising of three marketplaces namely: *Alphabay*, *Armory* and *Dreammarket* that have been identified as the major hub of firearm trafficking on the Tor dark web [82], [84]. The selected subset contains listings of several other products along with the firearm listings, hence, we manually extracted the firearm listings and put

them into two classes: *Pistol* and *Rifle*. We also created *Others* category that contains listings of ammunition, digital products, other arms and accessories, and other products that are available on the marketplaces like drugs, counterfeits, etc. A Python library for text parsing called BeutifulSoup[10] was used to parse the listings to extract the relevant content and stored it in text files. These files of individual listings shall be used in the experiments. The detail of selected subset of the dataset is given in Table 5.3.

Table 5.3: Description of the Dataset.

Class	**Number of Listings**
Pistol	392
Rifle	378
Others	2230
Total	**3000**

5.4.2 Evaluation Metrics

There are different evaluation metrics available to assess the performance of the machine learning models like accuracy, precision, recall and a combination of the former two called *f*-score. However, judging the performance in terms of precision and recall separately does not lead to an effective evaluation in the true sense, as the higher values of precision are obtained at the cost of low recall values [137]. Therefore, the *f*-score shall be used to evaluate the performance of the proposed method. In our context of work, the following terms are used to define the evaluation metrics.

True Positive (TP): Positive class correctly predicted
False Positive (FP): Negative class incorrectly predicted as positive
True Negative (TN): Negative class correctly predicted
False Negative (FN): Positive class incorrectly predicted as negative

[10] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Here, the positive class indicates the category of firearm listings while the negative class denotes the category of *Others*. A false negative outcome would mean that the model has predicted the product listing to be of the *Others* category while it actually was a firearm listing. Now, the evaluation metrics are given by Equations (5.1), (5.2) and (5.3) respectively:

$$f - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.1}$$

$$Precision = \frac{|TP|}{|TP|+|FP|} \tag{5.2}$$

$$Recall = \frac{|TP|}{|TP|+|FN|} \tag{5.3}$$

5.4.3 Validation and Parameter Settings

The *K*-fold cross-validation method is used to validate the experimental results. In this method, samples are split into *K* mutually exclusive parts known as folds and the classification algorithm is executed for *K* rounds. At each round, the *K*-1 folds are used for training the classifier and the remaining part for testing. The final classification results are obtained by merging the individual results of each round. In this method, all the samples from the dataset get the chance to be in the training set and each sample is employed once to test the model. Therefore, we apply the 10- fold cross-validation scheme.

While constructing the feature vector with TF-IDF, all those features that have a frequency below three were skipped. To obtain the optimal parameters for the Random Forest classifier, the grid search has been applied that performs the exhaustive search over the grid of parameter value and their possible combinations. The values of the parameters for the classifiers were set as given in Table 5.4.

Table 5.4: Configuration of the parameters of different classifiers.

Classifier	Parameters
Random Forest	max_depth: 70 max_features: 'auto' n_estimators: 400
Naïve Bayes	Default Settings
Logistic Regression	Default Settings
Stacking Classifier	stack_method: ‘auto’

5.5 Results and Discussion

The three classification algorithms: NB, RF and LR were employed to predict the firearm listings. The logistic regression classifier was used as the final estimator. Initially, the individual performances of the base models were obtained on the dataset. The *K*-fold cross validation scheme with *K*=10 is applied on the three approaches. All the experiments were implemented in the Python v3.6 [151], the associated NLP package NLTK was used for the elimination of the stop word. The PoS tagging of features was also implemented with the help of the NLTK library and the tags were provided by the Penn Treebank [162].

Scikit-learn library was used for implementing the classification algorithms and their ensemble. The experiments were carried out on an Intel Core i5 1.6 GHz machine running on Windows 8.1 operating system with 4GB of RAM.

The proposed methodology has been evaluated using precision, recall and *f*-score. The NB and RF act as the base estimators while LR did the final estimator task. The performance of the individual classifiers is shown in Table 5.5. The individual performance of the RF classifier is better than the NB, however RF is computationally slower than the NB classifier. The stacking of the NB and RF in an ensemble has achieved better performance than the individual classifiers alone. With close to 88% of *f*-score, the NB+RF ensemble could accurately predict the

specified firearm listings. The logistic regression classifier has performed well in integrating the output of NB and RF. However, the construction of the ensemble model took longer time than the individual model which is common in such settings. Figure 5.3 shows the comparison of the time required (in seconds) to fit the NB, RF and the stacking ensemble (NB+RF) on the training samples.

Table 5.5: Comparison of the individual classifiers.

Classifier	Precision	Recall	*f*-score
NB	78.88	80.49	79.68
RF	85.20	83.73	84.46
Stacking (NB,RF)	87.58	88.37	87.97

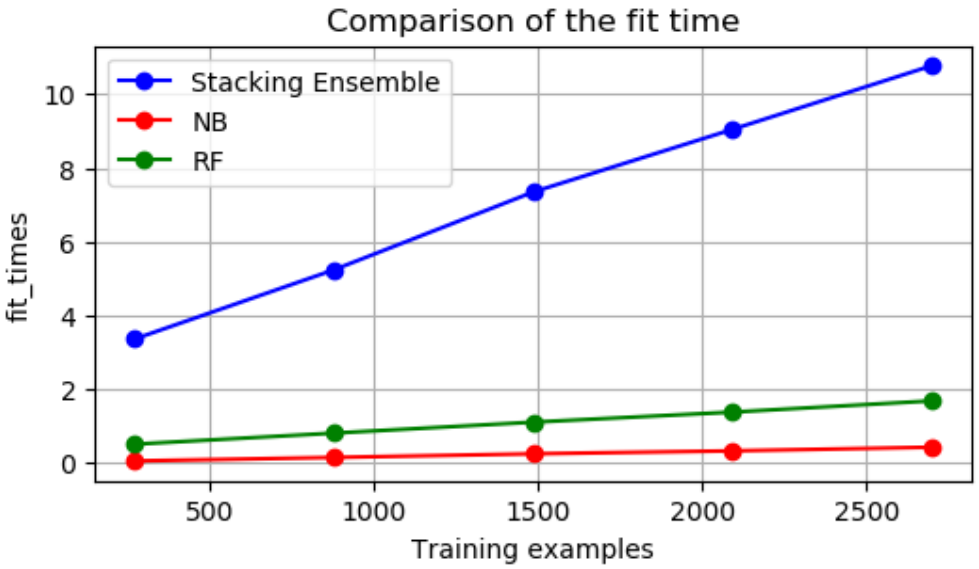

Figure 5.3: Comparison of the fit time of the three models.

The result of selecting the PoS tagged features and their combination of N-grams (N=1, 2, 3) on the classification performance for the individual classifier and their ensemble can be seen in the Table 5.6 and Table 5.7 respectively. The comparison of the size of feature space when different combinations of PoS tagged and N-grams features are utilized is shown in Table 5.8.

Table 5.6: Comparison of the base classifiers with PoS tagged features and N-grams.

Classifier	N-grams	Precision	Recall	*f*-score
NB	uni	89.42	79.52	84.07
	uni+bi	90.54	79.02	84.39
	uni+bi+tri	91.28	79.02	84.32
RF	uni	88.90	87.26	88.07
	uni+bi	94.83	94.65	94.74
	uni+bi+tri	92.31	91.79	92.05

Table 5.7: Performance of the Stacking Ensemble with PoS tagged features and N-grams.

Classifier	N-grams	Precision	Recall	*f*-score
Stacking (NB, RF)	uni	91.43	92.15	91.79
	uni+bi	97.15	96.84	96.99
	uni+bi+tri	93.46	92.75	93.10

Table 5.8: Size of the Feature Space.

Type of Feature Set	Size
Full	2003
PoS Tagged with Unigrams	1063
PoS Tagged with Unigrams and Bigrams	2470
PoS Tagged with Unigrams, Bigrams & Trigrams	3874

The application of feature set with only PoS tagged features has considerably increased the *f*-score of the individual base classifiers. The combination of unigrams and bigrams with PoS tagged features has achieved the best performance. However, the inclusion of unigrams and bigrams has resulted in small increase in the size of the feature set. The product listings on the marketplaces contain product name and specification like *M1 Garand, Glock 22, Ruger M77, breech action*, *lever action* etc which are bigrams. The inclusion of PoS tagged features and their N-grams (N=1, 2) could have helped in distinguishing the firearm listings from other products which subsequently improved the results.

A significant improvement in the classification performance of the stacking ensemble model is achieved upon the application of the PoS tagged features. The stacking ensemble also achieved the best results when unigrams and bigrams were used in the feature space. Figure 5.4 shows the learning curve of the stacking ensemble with PoS tagged unigram and bigram features. The ensemble model starts to generalize as the training samples are increased. The gap between the two curves is almost negligible representing low bias in the model. Finally, the two curves converge at the similar scores. Thus, the proposed ensemble model of the NB and RF classifier could accurately predict the listings of firearms.

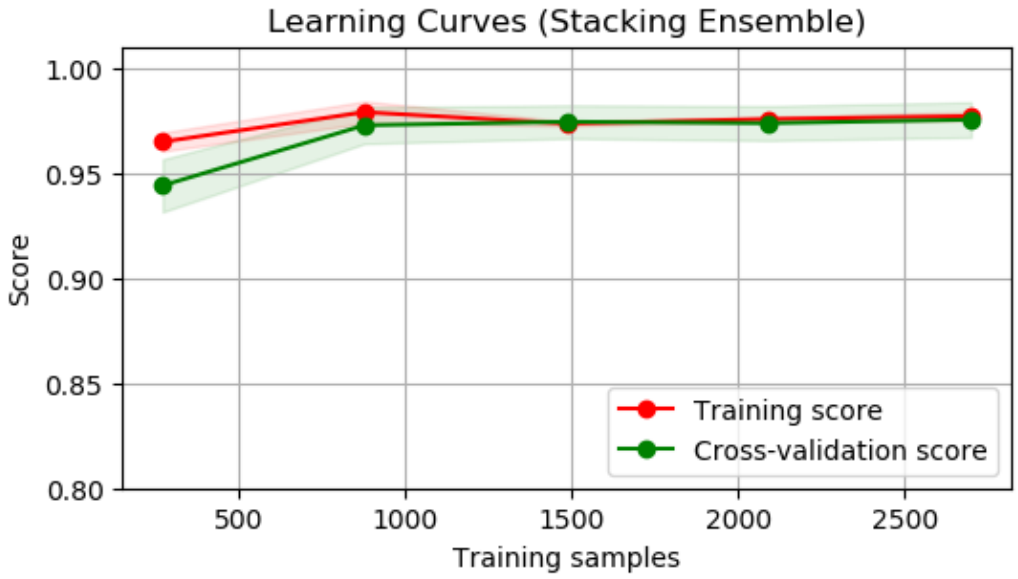

Figure 5.4: Learning Curves of the Stacking Ensemble.

CHAPTER 6

LINK BASED IDENTIFICATION

6.1 Background

The researchers have recently managed to expose the topological characteristics of a range of large and complex networks such as online social networks, the World Wide Web, biological networks and the Internet. Though these networks differ greatly in their functionality, size and components, they surprisingly have similarity in their topology that indicates that such complex networks are followed by some self-organizing rules that form the basis of their structure. The analysis of complex networks helps in understanding the real world phenomenon based on such networks.

The infamous dark web, though hidden from regular view is capable of posing serious negative effects on any society and nation in many ways. The knowledge about the dark web topology is not much known given its limited accessibility and dynamic nature. It is unknown whether the dark web shares network characteristics with other types of complex networks including the self-organizing rules. Moreover, it is largely unknown how the dark web can resist the constant monitoring from law enforcement agencies and withstand cyber attacks.

This chapter shall analyze the topological properties of the Tor dark web network by studying its corresponding web graphs and associated attributes. In addition to finding the general understanding of the Tor dark web, it would also be provided with insights into the disruptive strategies to the law enforcement agencies. Moreover, a hyperlink based ranking algorithm is proposed for identifying the influential hidden services in the Tor network.

6.2 Topological Properties of the Tor Network

Topological analysis is the methodology of studying large-scale networks by analyzing the statistical properties of the network structure [163]. There are three types of large complex networks: random, small-world and scale-free. Each type of complex network is identified by several statistics like the degree distribution, average path length, and the connected components. Table 6.1 outlines the key terminology of the graph theory that shall be used in the topological analysis of the Tor dark web.

Table 6.1: The terminology of the network topology.

Terminology	Description
Graph	The set of vertices or nodes and the set of edges or links are collectively called undirected graph or simply graph. Graph with specified direction between nodes represented by edges is called directed graph or digraph.
Path	A path $v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow \ldots\ldots\ldots v_n$ is a sequence of edges between two nodes v_1 and v_n such that $v_i \rightarrow v_j$ represents an edge from the node v_i to v_j.
Degree	The degree of the node is the total number of edges that are incident on it.
Connected Component	A connected component S in an undirected graph is a collection of nodes such that each node in S is reachable from every other node in S.
Centrality	The centrality of the node is the quantitative metric to define the relative importance of that node in the graph.

The two nodes in a random network are connected with a certain probability due to which each node has more or less the same degree. The random networks have a small average path length between any two nodes due to which one node can reach out to another node in a small number of steps. Random networks generally do not contain clusters of nodes and are

characterized by Poisson distributions [163]. However, it was found in studies that most complex networks are of a small-world and scale-free type than random networks.

Contrary to random networks, the small-world and scale-free networks show different characteristics. A small-world network contains relatively large clusters than the random networks while sustaining a small average path length [164]. On the other hand, scale-free networks follow the power-law in the degree distribution where the majority of the nodes in the network have a small degree value while the remaining few nodes have relatively very large degree value [163]. The scale-free networks have been thought of being self-organized networks where the nodes with a high degree attract the new connections and are preferred over other nodes leading to the appearance of the power-law distribution.

The topological analysis of the complex networks could give insights into the understanding of their functionality. The network topology may greatly affect the functionality of a complex network. For example, the easy accessibility and quick navigation of the Internet are due to its small average path length where on average a web page can be reached from any other web page with 19 clicks only. Moreover, scale-free networks are highly immune to failures but largely vulnerable to organized attacks [165].

The study of online networks like the dark web, and surface web requires the construction of its corresponding web graph. A web graph can be thought of as a collection of edges and vertices where a web page or website serves as a vertex and the hyperlinks among the websites forms the edges of the graph. By analyzing the web graph, the underlying properties of online networks are uncovered which can be helpful in the development of effective data mining and web crawling strategies.

6.2.1 Methods

The dark web text dataset has been extended to be used for the topological analysis of the dark web. The newly found links in the dark web text dataset were scraped to obtain new links. The newly found links were again scraped to obtain more links. This process is repeated two more times successively until no new links are found. Finally, the crawler was able to collect 48,174 onion domains that shall form the corresponding web graph. The nodes or vertices of the graph are represented by the individual hidden services and the hyperlinks between the hidden services constitute the edges of the graph.

A Python script is employed to extract all the hyperlinks from the hidden service by searching for the HTML <*a*> tags. The self pointing links (i.e. loops) and parallel links were ignored. The adjacency list has been constructed once the hyperlinks of all the nodes have been extracted. The directed graph is generated from the adjacency list using the Python NetworkX[11] library. The equivalent undirected graph is also obtained from the directed graph. The Tor web graph is generated at the domain level, therefore any individual web pages of a hidden service at the sub-domain level are clubbed together. Thus, each node in the web graph represents a hidden service along with its sub-domains. Similarly, edges of a node denote the hyperlinks from the domain including those from the sub-domains as well.

Each node of the graph and its incident edges represent the unique hidden service and its connections to other services respectively. All other statistical properties of the graph discussed in the subsequent sections are calculated by the in-built function of the NetworkX library. The generated Tor web graph contains 48,174 vertices and 103,526 edges.

[11] https://networkx.org/

6.2.2 Results

In this section, several graph theoretic properties of the Tor web graph are reported followed by a discussion on the possible insights into the network structure of the Tor dark web.

Degree Distributions

Figure 6.1 and Figure 6.3 show the in-degree and out-degree distribution of the nodes respectively. Around 7% of nodes have zero in-degree i.e they are the *source,* followed by other 39% nodes with only single incoming links. Overall, around 97% of nodes have in-degree below 10, a clear indication that the majority of the hidden services are hard to found with only seven hidden services have in-degree above 100 and the highest in-degree value being 184. On the other hand, the out-degree also shows extremities in the distribution, 75% of nodes have zero out-degree i.e. they are *sink,* 98% have out-degree below 10. Around 31 hidden services have above 100 values of which 9 nodes have above 1000 out-degree.

The highest out-degree node has 2846 outgoing hyperlinks that could cover a large portion of the graph. The top ten highest out-degree nodes were the Wiki/directory services. Though they may have a relatively large out-degree, they are quite difficult to find due to their small in-degree (<10). Another key finding was the presence of some isolated nodes that makes the graph disconnected. The isolated nodes have zero in-degree and out-degree values. Around 3006 such nodes were found in the graph. Figure 6.2 and Figure 6.4 show the in-degree and out-degree distribution on the log-log scale. The in-degree distribution shows several spikes at the tail whereas the points are scattered irregularly in the out-degree distribution. The *p*-value of the Kolmogorov-Smirnov goodness of fit test was more than 0.1 which suggests that the in-degree distribution follows the power law [189].

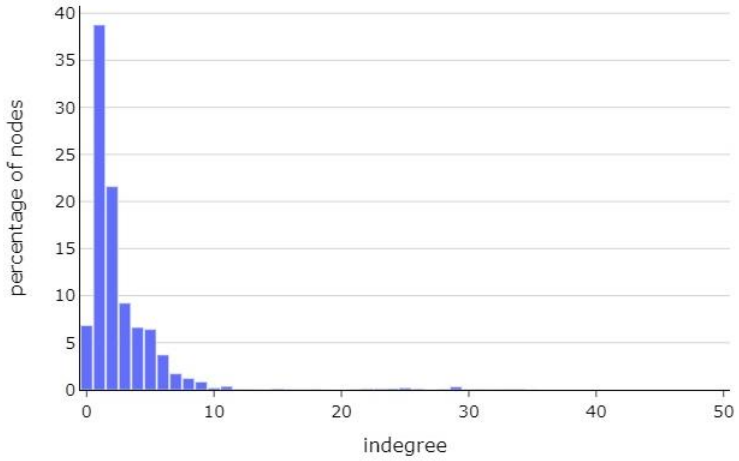

Figure 6.1: In-degree distribution.

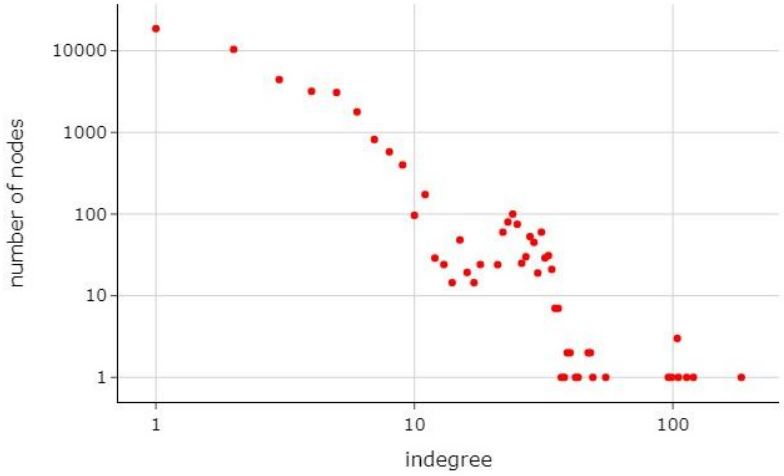

Figure 6.2: In-degree distribution (log-log scale).

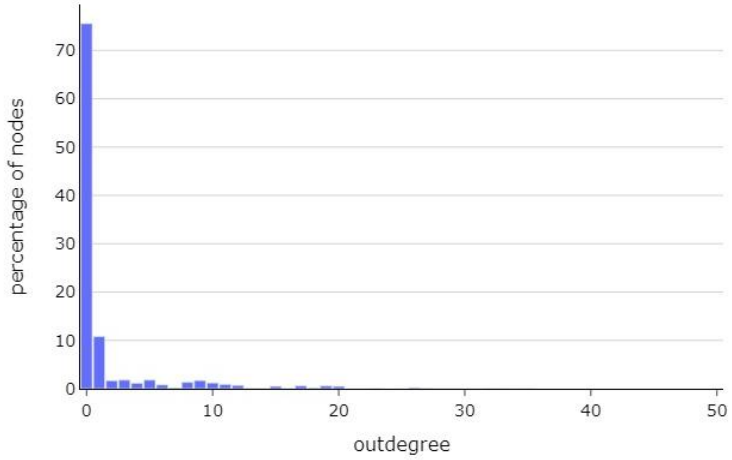

Figure 6.3: Out-degree distribution.

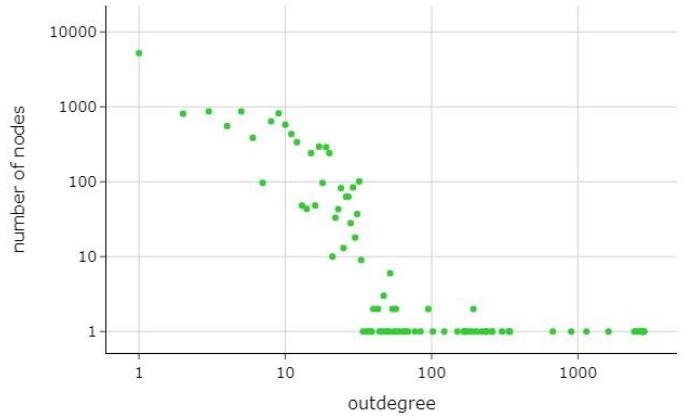

Figure 6.4: Out-degree distribution (log-log scale).

Overall, the graph is said to have sparse nature as more than 95% of nodes have the out-degree and in-degree values below ten. The edges are seen to be clustered around some high out-degree nodes. However, the in-degree and out-degree nodes largely differ in their size unlike in the surface web where they are of comparable size [69]. The out-degree node is around fifteen times bigger than the largest in-degree node which subsequently leads to the relatively slow decay of the in-degree distribution as compared to the out-degree distribution.

Centrality Measures

Centrality metrics in the topological analysis are used to identify the prominent nodes in the network. In the case of the Tor network, the central nodes may be the significant hidden services that could hold the structure of the network. Here, PageRank and eigenvector centrality is used to identify the key nodes in the graph. The PageRank was originally developed to identify the central web page in the World Wide Web [166]. It assigns a probability to each node in the graph. A higher probability value indicates a greater chance of the node being accessed from any other random node in the network. Table 6.2 shows the top four nodes with the PageRank value and Figure 6.5 shows the PageRank distribution. The PageRank and in-degree are highly correlated with correlation coefficient of 0.94 as the top four nodes are the same in both the

PageRank and the in-degree distribution. This correlation is also found in the analysis of the pay level domain (PLD) graph of the surface web [69].

Table 6.2: The top four PageRank nodes and their description.

S.No.	PageRank	In-degree	Description
1.	0.037	120	100x Your Coins in 24 Hours - Officially Hidden Service Anonymous
2.	0.024	184	Dream Market Forum
3.	0.012	113	Dream Market Login - Featured anonymous marketplace
4.	0.009	105	Dream Market Login - Featured anonymous marketplace (Mirror)

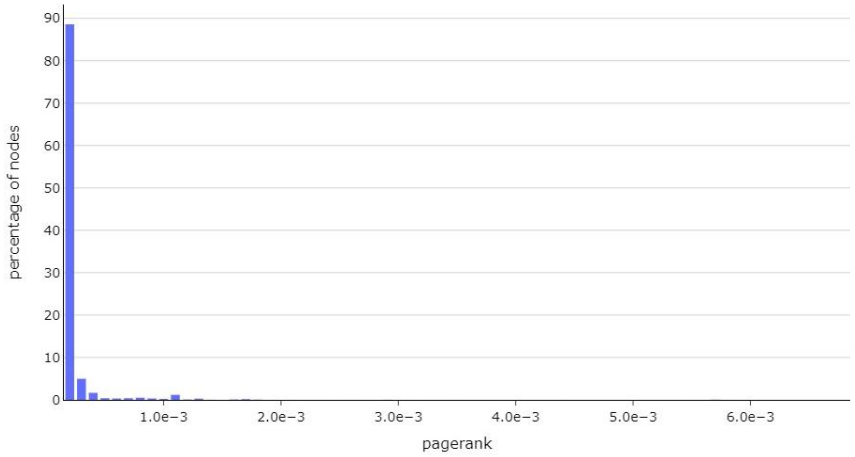

Figure 6.5: PageRank distribution.

In Table 6.2, despite having less in-degree than node #2, node #1 has a higher PageRank value because it is a Bitcoin-related service that may get incoming hyperlinks from other important services like the marketplaces having good PageRank that ultimately improves the PageRank of node #1. For node #2, the marketplace forum is highly popular among the users to discuss their issues and opinions regarding a product and service hence it get a high PageRank value. Moreover, *Dream Market* is among the popular marketplaces on the Tor dark web [167]. On the other hand, the out-degree does not have any impact on the PageRank value.

The eigenvector centrality identifies the central nodes by assigning a score to each node in the graph based on its connectivity to the other nodes. A node with connections to other high-scoring nodes is assigned a higher score than another node having the same number of connections to other nodes but having low scores. Overall, a node is important if its surrounding nodes are important. The eigenvector centrality distribution is shown in Figure 6.6. Table 6.3 shows the four nodes with the highest eigenvector centrality score.

Table 6.3: The top four eigenvector centrality nodes and their description.

S.No.	EGV	Out-degree	Description
1.	4.713E-01	2846	CB3ROB Tactical Data Services - TOR Darknet site listing
2.	4.50E-01	2731	The Hidden Directory
3.	4.182E-01	2474	The onion crate-Tor hidden service index
4.	3.787E-01	2150	Jack's Tor Hidden Links

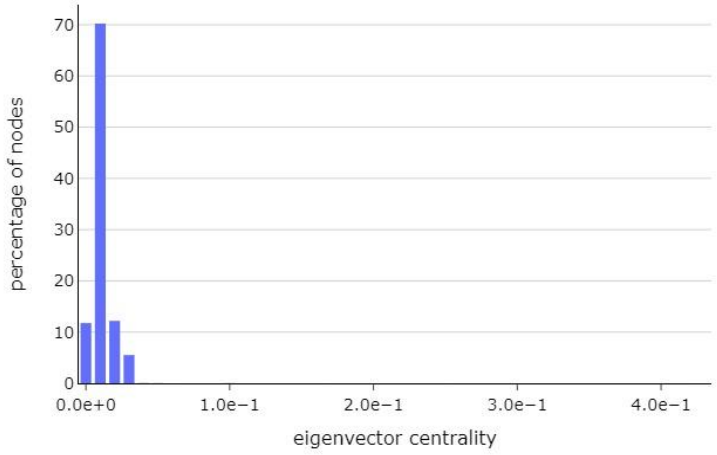

Figure 6.6: Eigenvector centrality distribution.

The eigenvector centrality, unlike the PageRank, is unaffected by the in-degree of the node and it has a positive correlation with the out-degree of the node. The out-degree and the eigenvector centrality have a linear relationship with a correlation coefficient of 0.92.

Distances

The pair of nodes having a direct connection between them was only 2.3 % among all the possible pairs in the graph. The sparse nature of the dark web is again reflected by the presence of only a handful of connected pairs of nodes. The shortest path length of the connected pairs falls in the range of 1 to 12 with the average length being 4.32. Thus a user can reach from one hidden service to another in a path by clicking at most three times on average. Despite being poorly connected with other nodes, the average path length of the dark web is roughly the same as that of the surface web which is 4.27. Though the percentage of connected pairs in the surface web was well above forty [69]. The distribution of distances between connected pairs is shown in Figure 6.7.

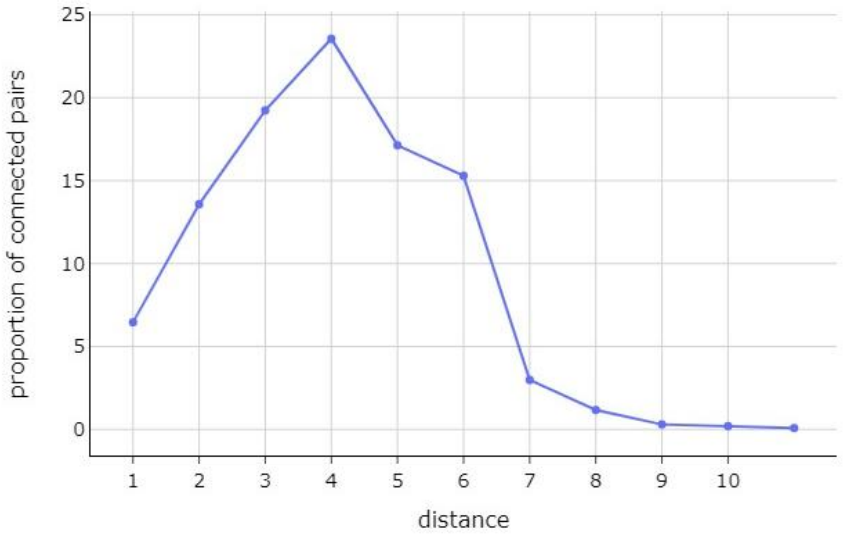

Figure 6.7: Distance distribution of connected pairs.

Connected Components

The presence of the isolated nodes has rendered the graph largely disconnected. However, the connectivity of the remaining portion of the graph is examined by eliminating the isolated nodes from the graph and then identifying the connected components of the directed and the undirected graph. The undirected version of the graph remains disconnected even after skipping the isolated nodes. A total of eight

connected components were present. It contains a single fairly large connected component with 45089 nodes followed by another connected component of size 57. All the remaining components contain less than ten nodes. Thus the largest component covers almost entirely all the nodes of the graph excluding the isolated ones.

The access to any of the nodes of the large component would ensure a complete traversal of the graph. This also holds for the surface web to much extent where the largest component comprises more than 90% of the websites [69]. The removal of the top five out-degree vertices from the graph significantly reduced the size of the largest component such that it could only cover half of the web graph. However, the internal connectivity of the graph was unaffected by the removal of the top five out-degree nodes. This reflects the robustness of the graph structure that does not disintegrate upon removal of the top out-degree nodes.

Moving to the directed graph, four strongly connected components (SCC) of size 1796, 83, 27 and 16 respectively were identified while other SCCs has size less than 10.

Bow-Tie Decomposition

Broder *et al.*, 2000 [68] presented the bow-tie like structure of the World Wide Web formed by six components. The components are the mutually disjoint sets whose descriptions are given as follows:

LSCC: The Largest Strongly Connected Component of the graph also called CORE.
IN: The set of nodes excluding those in LSCC and are reachable to CORE.
OUT: The set of nodes excluding those in LSCC and are reachable from CORE.

TUBES: The set of nodes excluding those in LSCC, IN and OUT such that they lie in between the directed path from IN to OUT.
TENDRILS: The set of nodes excluding all the above-listed nodes such that they are reachable from IN or can reach OUT.
DISCONNECTED: The set of all the remaining nodes.

The LSCC is of size 1796 and the remaining five components are obtained using LSCC. Figure 6.8 shows the bow-tie structure of the Tor dark web with its different components. The OUT is comparatively bigger than the IN due to the presence of large out-degree nodes in the CORE that connects to most of the non CORE nodes. The DISCONNECTED contains all the isolated nodes of the graph. The 20.15% out of total 22.73% of TENDRILS belongs to subset INTENDRILS that are nodes reachable from IN while the remaining forms the subset OUTTENDRILS having nodes that can reach OUT.

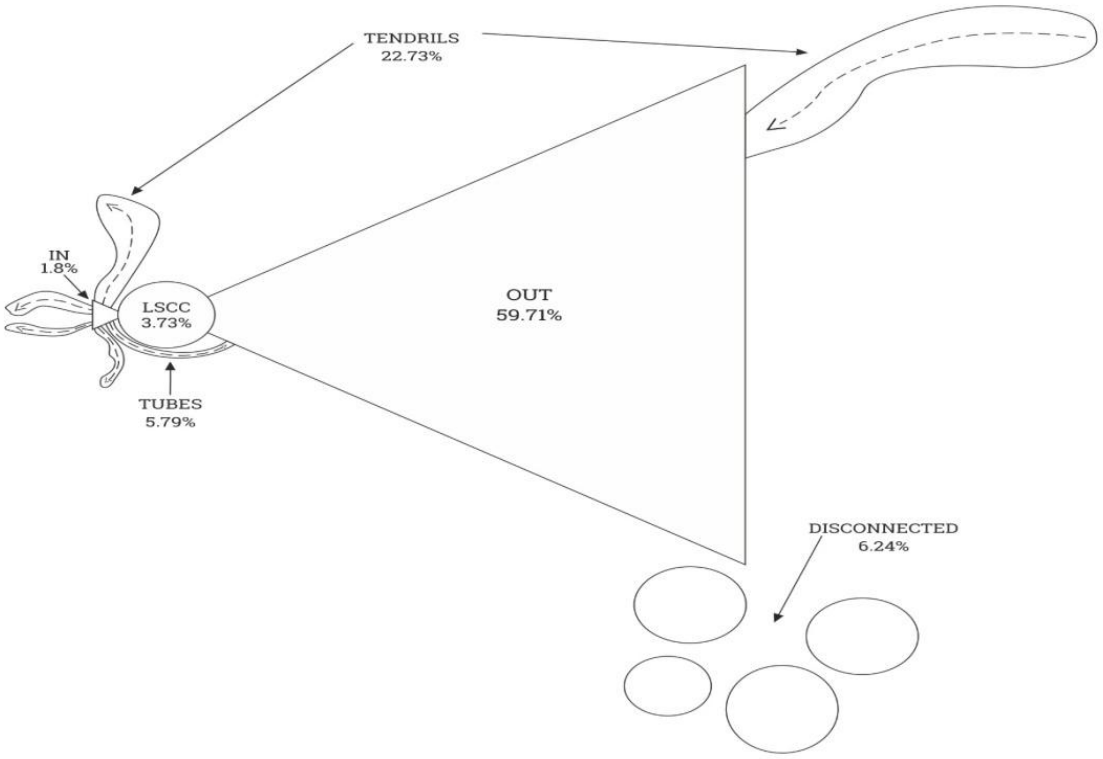

Figure 6.8: Bow-tie decomposition of Dark Web.

Now, we compare the size of different components of the bow-tie structure of the Tor dark web graph to their counterparts in the surface web as presented in the literature [69]. Table 6.4 shows the comparison where we can see that due to the weak connectivity, the Tor has a comparatively smaller LSCC than that of the surface web. However, the OUT component is twice the size of its surface web counterpart containing a large chunk of nodes of the graph than any other component. The DISCONNECTED is nearly the same size in both the Tor dark web and surface web.

Table 6.4: Comparison of size of bow-tie components of the Tor Dark Web and the Surface Web.

Bow-Tie Component	Dark Web	Surface Web
LSCC	3.73	51.94
IN	1.8	7.65
OUT	59.71	30.98
TUBES	5.79	0.04
TENDRILS	22.73	1.2
DISCONNECTED	6.24	8.2

Small-World and Scale-Free Characteristics

In the Tor web graph, the majority of the nodes have small out-degree except few nodes that have a very large out-degree value. The web graph also resists the removal of the highest out-degree nodes and does not break down completely. The above two features reflect the scale-free characteristic of the Tor network [168]. The small-world property of a network is recognized by the presence of a small average shortest path length of distance six or less [169]. The average distance between the two nodes in the graph was 4.32 which confirms the small-world type of the Tor network. The small-world property suggests that a user on any hidden service can navigate to other services through few clicks only. The surface web and the Tor dark web seem to be sharing the scale-free and small-world characteristics [70], [170].

Connectivity to the Surface Web

Around 19386 hyperlinks were found to be originating from the Tor dark web towards the surface web resources and of which 5406 hyperlinks were of distinct URLs. Additionally, 954 more URLs were found with the *.i2p* suffix that represents the websites present over the I2P dark web network. Figure 6.9 shows a diagrammatic representation of the surface web connections originating from the Tor hidden services. For clear presentation, only specific websites that have in-degree (of hyperlinks from the dark web) above ten and the commonly known websites on the surface web are shown. In Figure 6.9, the weight assigned to each of the connection arrow from the dark web indicates the total of the incoming hyperlinks received by the corresponding node from the Tor hidden services.

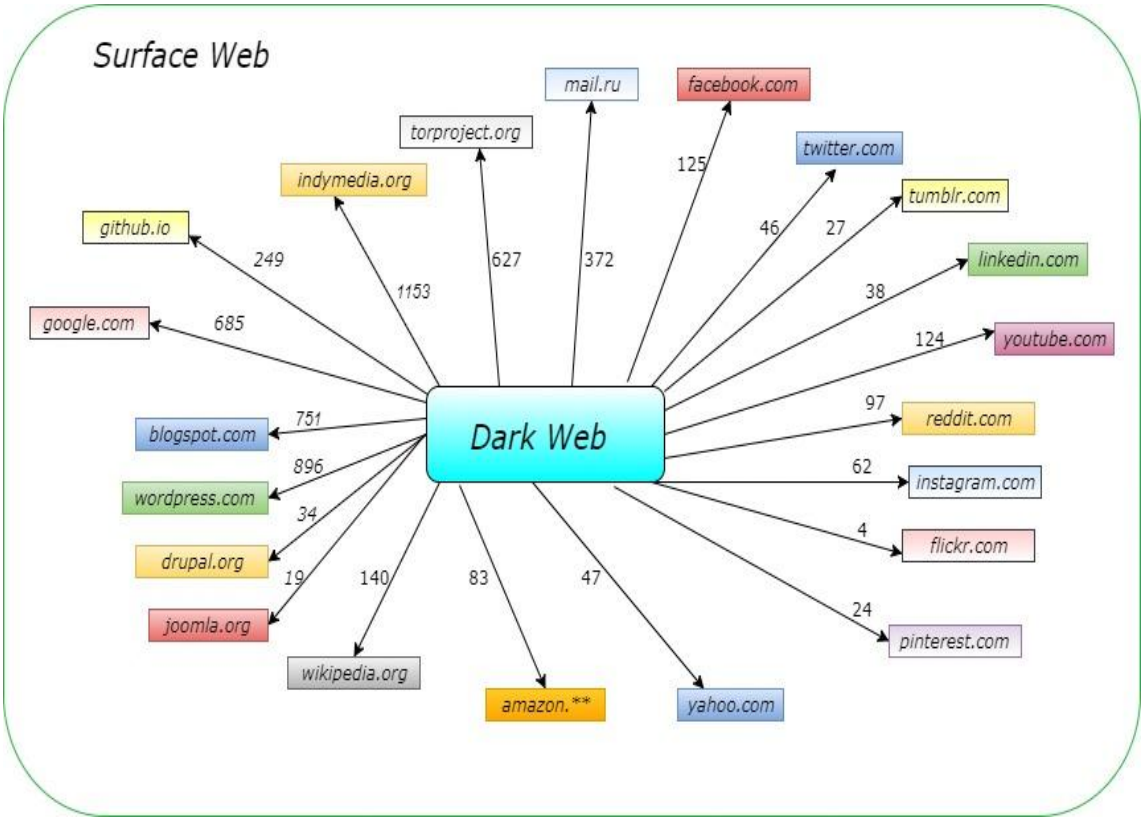

Figure 6.9: Pictorial representation of connectivity of Tor Dark Web to the Surface Web.

The website *indymedia.org* on the surface web receives the maximum number of incoming hyperlinks from the Tor dark web. The dark web platform has been proactively used to propagate uncensored information by the news journalists and whistleblower [171], hence the popularity of *indymedia.org* on the dark web may be attributed to the services offered by it to the group of journalists and whistleblowers.

The considerable number of hyperlinks to content management and blogging services like Blogspot, Joomla and WordPress reflect the preference of the Tor community in using the predefined templates to publish and share their content and expressions. The popular social network sites on the surface web also have many connections from the Tor dark web. The adult content services also get 124 hyperlinks from the Tor hidden services.

Connectivity to the Top Level Domain (TLD) on the Surface Web

The 11 Top-Level Domain (TLD) suffixes that get the most number of links from the Tor are identified. Table 6.5 represents the suffixes identified and their corresponding number of links from the Tor dark web. The remaining TLDs are placed under a separate category called the *Others*.

Table 6.5: TLD wise distribution of incoming hyperlinks.

TLD	Count	TLD	Count
com	3982	*io*	165
org	3676	*info*	153
net	1584	*eu*	102
de	1247	*edu*	96
ru	816	*uk*	88
fr	361	*Others*	7116

The suffixes shown in Table 6.5 together form nearly 63% of the hyperlinks from the Tor dark web, of which *.com* and *.org* combined have

approximately half of the total links. Moving to the country-wise distribution of TLDs, *.ru (Russia)* and *.fr (France)* takes first and second place respectively in the list. As already discussed in Chapter three, the linguistic distribution of the dark web text dataset has found that the French, Russian and German languages are the top non-English language with content on the Tor. In all of the *.ru* TLDs, 372 links were of *mail.ru*, the Russian Internet giant that has reach over 86% percent of the Russian Internet users [172].

One contrasting feature in the above findings is the significant number of hyperlinks to the *.edu* domains which diverge from the bad perception of the dark web of hosting controversial and unethical content [38-40]. The hyperlinks with *.edu* TLD are of premier academic institutes like Stanford, Harvard, MIT, Princeton, etc. However, upon closer inspection of the hyperlinks, it was found that most of the *.edu* links were originating from a single hidden service that provides access to scholarly and research articles free of cost by evading the subscription and paywalls.

6.2.3 Discussion

The graph theoretic properties of the Tor web graph have uncovered interesting characteristics of the Tor network. The low values of the degree of the hidden services differentiate them from the surface websites. The majority of the nodes with small in-degree value indicate that the hidden services abstain themselves from their advertisement. The hidden services may have managed to maintain a dedicated and trusted user base that is well aware of their web address thereby do away with the requirement to publish their URL. The hidden services may also not care about advertising their web locations owing to their short life cycle [35]. In this way, the hidden services may avoid the interventions from the law enforcement agencies to some degree.

However, hiding their identity from the outsiders gives rise to a new question, how the dedicated user bases get to know the location of their desired hidden service? One answer is that there may be some discussion forums and other online platforms available on the surface web that act as a bridge between the user and the hidden service by regularly giving updates about the current location of the service. This point is also strengthened by the fact that many hidden services have connections to popular social networks and blogging websites on the surface web. For instance, a popular drug marketplace on the Tor could have a dedicated discussion forum on the surface web that updates its loyal customers about their new stock and dark web location in a disguised way. The isolated nodes may also apply the same *modus operandi* to get in touch with their users. However, more research is needed to get the actual picture behind this.

While the hidden services do have low in-degree, they are also not interested in providing hyperlinks to other services in the network given their low out-degree values. They prefer not to provide out-going links to other resources. The short lifespan of the hidden services may again be the reason behind their divergent attitude. The dynamic environment of the dark web where a service may go down within hours of coming alive does not ensure that an out-going link will always lead to an active and running resource. Moreover, the dark web ecosystem is highly competitive for doing business [173] doubled with the chance of law enforcement interventions. In such a harsh scenario, any out-going link would increase the probability of a user switching to other services as well as an invitation to the investigating agencies.

The Tor network structure is resilient to the elimination of key nodes from the graph. The removal of the highest out-degree node from the largest connected component could only disintegrate fifty percent of

the nodes. Therefore, law enforcement actions may not prove to be much effective in disrupting the network. Moreover, the law enforcement action to shut down even a single hidden service is a costly affair that requires collaboration from multiple agencies and hence is an ineffective step [34]. However, the highest out-degree nodes like Wiki/Directory services may act as a gateway to all other hidden services in the connected component and thus they control the entry of first-time users to other services in the largest connected component. The removal of such nodes may restrict many new users from accessing a significantly large portion of the dark web. Though, previous users may not be affected if they still have access to any of the other services.

The introverted nature of the hidden services and a short average path length between any two services allows them to maintain a strongly connected component of similar content. This type of bonding among the hidden services could prove advantageous for the law enforcement agencies to simultaneously bring down many services. For example, starting with a single marketplace and following successive links could help track down all the mirror services and associated forums in the strongly connected component. This mechanism has the potential of disrupting a large user group since the top three nodes with the highest out-degree connect to nearly 93% of the remaining nodes in the network. The topmost out-degree nodes could serve the purpose of entry gate to the Tor dark web network for the law enforcement authorities to undertake appropriate actions.

The out-going hyperlinks to the I2P network from the hidden services could be an indication of the expansionist attitude of the hidden service administrators. It is advantageous to them in two ways: first, it may help increase the user base of the service and secondly it serves as an alternative avenue in case of disruption of the service in the Tor network.

6.3 Link Analysis Algorithm for Identifying Influential Hidden Services

The high degree of anonymity granted by the Tor network has subsequently led to the growth of unlawful activities on the dark web. While there are numerous hidden services on Tor that deals in the business of illegal products and services, there exist some prominent or key websites with an exclusive range of illegal products and have a dedicated customer base. These hidden services are crucial in terms of law enforcement perspective as their identification and shutdown could lead to an upheaval in the network.

The existing works have explored the Tor hidden services by scraping their content followed by their analysis to draw possible conclusions on the characteristics of the hidden services. The scraped content could also be utilized to identify the influential hidden services using the graph theory and hyperlink based algorithms. The concept of influential domains in the Tor network is similar to that in the surface web. A domain or hidden service is considered influential if a user arrives at it after surfing the network moving from one domain to another through the hyperlink connectivity among them [37]. Here, a hyperlink based algorithm is proposed for ranking the hidden services in the Tor network to identify the influential ones among them.

The law enforcement agencies can concentrate their operations on the top hidden services ranked by the proposed algorithm. However, the other low-ranking hidden services should not be completely ignored instead the agencies should work out plans to adequately manage their efforts in monitoring the illegal content. The ranking algorithm could identify the clone of an influential domain that was previously brought down by the concerned authorities if it appears on the network and gains popularity again.

6.3.1 Methodology

The ranking algorithm presented here needs the construction of a web graph representing all the hidden services of the dataset under study. The ranking algorithm is implemented on the web graph to generate the rankings of the individual hidden services. The main stages of the proposed ranking algorithm are described as follows.

Construction of Web Graph

The web graph corresponding to the hidden services is composed of a collection of nodes (or vertices) and edges. The nodes of the web graph represent the individual hidden service whereas the directed edge between the two nodes in the graph depicts the hyperlink between the corresponding hidden services in the dataset. The self-loops (a self-loop is an edge with the same start and end node) and parallel edges (two edges are parallel if they have a common start and end node) were eliminated from the web graph. All those edges were removed from the web graph that points to the node that does not exist in the web graph.

The graph obtained after applying the preceding actions is used for further steps. Moreover, for each of the nodes, the number of hyperlinks directed towards the surface websites was saved. If any node has multiple out-going hyperlinks to the same website on the surface web, then only a single count is taken for that website to the total.

Ranking Procedure

The proposed ranking algorithm calculates the overall ranking of a node based on three separate components: number of hyperlinks to the surface web, central location of node within the Tor and influence of the adjacent nodes. The three components measure the overall influence of the hidden services in the graph. Table 6.6 contains the definition of the various symbols used in the subsequent sections.

Table 6.6: Definition of the symbols used.

Symbol	Definition
$V = \{v_1, v_2, v_3, \ldots\ldots v_n\}$	set of nodes in the Tor web graph
$E = \{e_1, e_2, e_3, \ldots\ldots e_m\}$	set of edges in the Tor web graph
$surf(v_i)$	number of surface web hyperlinks of a node v_i
$deg(v_i)$	degree centrality of a node v_i
$btw(v_i)$	betweenness centrality of a node v_i
$cls(v_i)$	closeness centrality of a node v_i
$r(v_i)$	the overall influence of the node v_i

Influence of Surface Web Hyperlinks: The proposed ranking algorithm takes into consideration the importance of the connections to the surface web to the influential nature of a hidden service in the dataset. A hidden service having good connectivity to the surface websites may reflect its willingness to publicize its goods and services on the regular web where a relatively larger number of users exists. A study has found that more than 90 percent of the Tor hidden services have hyperlinks to the surface web. Moreover, several Tor directories and other services with adult content, media and news content are involved in the communication and information transfer from dark web to the surface web [190]. The significance of the surface web links needs to be quantified to incorporate their weightage in the overall rankings.

For this purpose, $\tau(v_i)$ defined by Equation (6.1) measures the influence of the connectivity to the surface web of a node v_i. $\tau(v_i)$ is the ratio of the total of the surface web hyperlinks $surf(v_i)$ to the degree centrality of the node v_i. In the case of nodes with zero degree centrality, $\tau(v)$ would be undefined, hence one is added to the denominator of the fraction to evade the indeterminate forms. Therefore, $\tau(v)$ for the isolated nodes will be equal to $surf(v)$.

$$\tau(v_i) = \frac{surf\,(v_i)}{deg\,(v_i)+1} \tag{6.1}$$

Influence of Node Connectivity: The connectivity of the node in the graph influences the importance of the node and is governed by its location in the network. If a user surfs to a node with a good location in the graph then it will have a better chance of reaching other parts of the network via the neighboring vertices, their adjacent nodes and so on until the user arrives at the intended node. Therefore, such strategically located nodes are influential in the Tor network and their detection may be beneficial in monitoring the entire network by the concerned agencies.

The various graph theory metrics used to measure the centrality of the node show the relative importance of the nodes in the graph. Such metrics quantify the importance of the location of a node in the graph. The value of the metrics reflects the capacity of the node to provide passage to many users via multiple paths through that node [174]. The three common centrality metrics used are closeness centrality, betweenness centrality and degree centrality.

- **Closeness Centrality:** Closeness centrality tells whether a node is closely located to every other node in the graph or not [175]. The shortest path length between two nodes is defined as the distance between the two nodes. The closeness centrality of a node is the reciprocal of the sum of its distance to all the other nodes in the graph. If the sum of the distances to other nodes is small, then the closeness centrality of a node would be large and vice versa. A node with high closeness centrality value would indicate its closer relationship with other nodes in the network.

- **Betweenness Centrality:** Betweenness centrality indicates the extent that the node appears on the shortest path between other pairs of nodes in the graph. It measures the capacity of a node to act as a bridge in the graph. In simple words, it calculates the

fraction of the shortest paths passing via a node [176]. Though the calculation of betweenness centrality is comparatively complex, it can be used to compute the ability of the node to control the movement of users through the network.

- **Degree Centrality:** Degree centrality is the simplest centrality metric and is defined as the total number of edges incident upon it i.e. the total number of edges that a node has [175]. In case of a directed graph, two different metrics of degree centrality are used that are in-degree and out-degree. In-degree of a node is the total count of all the edges that end at that node while out-degree is the total number of edges that start from the node. Therefore, for a directed graph, degree centrality of a node is the sum of its out-degree and in-degree. Hence, the more edges a node has, the greater is its degree centrality.

In a web graph, a node that has a higher degree centrality value would allow a higher number of users to move through it than the node with a lower degree centrality [177]. A node with high closeness centrality would accelerate the spread of users to the other nodes in the network. Such nodes can be influential in the network as they can be easily reached to other nodes and would be effective in facilitating the movement of users [178]. The greater value of the betweenness centrality of a node would enable it to have more control over the movement of the users as it sits on several paths in the graph. Therefore, more users would depend on a high betweenness centrality node to move to the other nodes in the graph [179]. The influence of the connectivity of a node v_i in the network is represented by $\mu(v_i)$ and is defined by Equation (6.2).

$$\mu(v_i) = deg(v_i) + cls(v_i) + btw(v_i) \tag{6.2}$$

Calculating Overall Influence: The proposed ranking algorithm assigns an overall influence score to a node based on the individual influence score derived from its connectivity to the surface websites and connectivity within the web graph. The overall influence score measured by the influence metric varies from the other centrality metric in such a way that it takes into account, the ability of the node to control the movement of users and their speed in the Tor network. Since each of the individual nodes in the graph differs in their ability to facilitate the movement of the users, each of them has a different individual influence score. The influence score of a node v_i is given by $\delta(v_i)$ and is defined by Equation (6.3).

$$\delta(v_i) = \tau(v_i) + \mu(v_i) \tag{6.3}$$

The hidden services in the Tor web graph have different influences and importance when compared with the other services in the graph depending on the connectivity of each domain. Therefore, the process of identifying and detecting the influential hidden services can be regarded as the problem of ranking wherein the most influential hidden services would be the top-ranked in the list among others. The influence of a particular hidden service also depends on the other hidden services to which it has outgoing hyperlinks in addition to its attributes. A hidden service that has outgoing hyperlinks to other influential hidden services in the network, then its influence would be increased. In such scenarios, the PageRank algorithm can effectively be applied to incorporate the influence of the adjacent nodes [144], [180], [181].

The proposed ranking algorithm is based on a modified PageRank algorithm. An initial influence score $r(v_i)$ is assigned to each of the node v_i in the web graph which is updated iteratively according to Equation (6.4) until the convergence is achieved.

$$r(v_i) = (1 - a) + a \sum_{v_j \in Q_i} \log\{r(v_j) * \delta(v_i) + 1\} \quad (6.4)$$

Q_i denotes the set of all the nodes in the graph that have an incoming hyperlink from the node v_i and a $\in [0,1]$ is the damping factor set to the value 0.85 [166]. The cumulative influence of the node v_i and its adjacent nodes is obtained by taking the logarithm of the product of $r(v_j)$ and $\delta(v_i)$. Since the logarithm function is undefined when its argument is zero, one is added to the product to avoid such condition when $\delta(v_i)$ is zero.

The final rank of the node v_i depends on two factors: i) the individual influence score $\delta(v_i)$ and ii) the connectivity to other influential nodes represented by $r(v_j)$. If $Q_i = \{\emptyset\}$ for a node v_i, then $r(v_i)$ would be 0.15. This means the node v_i is influential only if it has outgoing hyperlinks to other nodes in the network. In other words, a node is considered to be influential only if it allows the users to pass through it to the other nodes in the network. Thus, the elimination of a node with zero or low out-degree would not cause much disruption in the network. Moreover, the shutdown of even a single hidden service from the Tor network is a costly and time-consuming operation [34]. Therefore, the proposed ranking algorithm would identify the top influential nodes whose removal may bring upheaval in the Tor network.

6.3.2 Experimental Setup

The experimental setup required to implement the proposed ranking algorithm that includes the dataset description, evaluation metrics and graph robustness metrics are discussed in this section.

Dataset

The dark web text dataset has been used to implement the proposed ranking algorithm. The selected subset contains 4041 samples in the dataset. Since the ranking algorithm is purely based on the hyperlinks, all

other textual content was discarded from the hidden services and only hyperlinks were extracted with the help of a parser. The hyperlinks to the surface web were extracted by a regular expression based parser. All hyperlinks to the other dark web networks like I2P, sub-domains and Internet Relay Chats (IRC) were ignored.

Evaluation Metrics

The results of the proposed ranking algorithms are compared with the results of other link analysis algorithm to assess the performance. The performance is compared with that of the PageRank [166] and ToRank [37] algorithms. Following the procedure used in the existing literature [101], [182], [183], various graph theory metrics that evaluate the robustness of the graph will be used to assess the performance of the proposed ranking algorithm. The description of the graph metrics is given as follows.

Graph Density: Graph density measures the extent to which the graph is connected. A high value of the graph density indicates the strong connectivity in the graph. The graph density (GD) is derived from Equation (6.5).

$$GD = \frac{e}{n(n-1)} \tag{6.5}$$

Clustering Coefficient: The clustering coefficient measures the extent to which the neighbors of a particular node are connected to each other [164]. Its value lies between zero and one.

Average Shortest Path Length: The average shortest path length is the average of the shortest path length of all the possible pairs of nodes in the graph [185].

Diameter: Of all the shortest path lengths between any two nodes in the graph, the longest path among them is the diameter of the graph.

Giant Component: A strongly connected component of the graph that contains the highest proportion of all the nodes in the graph is referred to as the giant component of the graph.

The robustness of the graph could be measured from the graph density curve [184]. The top-ranked nodes returned by the ranking algorithm are repeatedly removed from the graph along with its associated edges and the graph density is computed at each elimination. The process of iteratively removing the nodes is stopped once the graph density becomes zero.

In line with the method of the previous work [37], the ranking algorithm that spans the smallest area under the graph density curve shall be the best performer among all the algorithms in detecting the influential nodes in the graph. If the influential nodes are correctly ranked by the ranking algorithm, then the most influential node would be at the top of the list followed by the other nodes with decreasing influence. The removal of the top nodes would result in a drastic reduction in the graph density given the good connectivity and location of the influential nodes in the graph structure.

The robustness of the graph structure could also be measured from the size of its giant component and the clustering coefficient. If the removal of top-ranked nodes from the graph causes a significant reduction in the clustering coefficient and size of the giant component, then the ranking algorithm has correctly ranked the influential nodes. On the other hand, the removal of the correctly ranked nodes would ultimately increase the average shortest path length and the diameter of the graph which

subsequently reflects the better performance of the ranking algorithm [186-188].

The repeated computation of the four evaluation metrics after the elimination of each node of the graph is a computationally expensive task, therefore the above four metrics are calculated only at the removal of the top 1^{st}, 5^{th}, 10^{th} and 20^{th} percentile of the nodes from the ranked list.

Parameter Settings

The web graph corresponding to the Tor hidden services in the graph is constructed using the Python NetworkX library. The influence score of each node in the graph is iteratively computed by the ranking algorithm until convergence is achieved. The proposed ranking algorithm shall suppose to converge when the error difference in the influence score of a node between the two consecutive iterations is less than 0.0001. The values of the parameters of the PageRank and ToRank algorithm are set as per the previous study [37]. Each node is assigned an initial influence score which is equal to one divided by the total number of nodes in the graph. The experiments were implemented on a system running on an Intel i5 processor with 4 GB of RAM and Windows 8.1 operating system.

6.3.3 Results

The web graph generated from the Tor hidden services contains 4041 nodes and 14059 edges. The nodes are ranked in descending order by their corresponding influence score returned by the ranking algorithm. The top-ranked nodes are then removed one by one and the graph density is calculated at each removal. The graph density is also calculated for the ranked nodes returned by the PageRank and ToRank algorithm. Figure 6.10 shows the graph density curve of the proposed ranking algorithm, PageRank and the ToRank algorithm. The smallest area under the curve is 0.013 which is obtained by the proposed ranking algorithm followed by

the ToRank algorithm whose area under the curve is 0.015. The PageRank, however, covers a relatively larger area under the graph density curve.

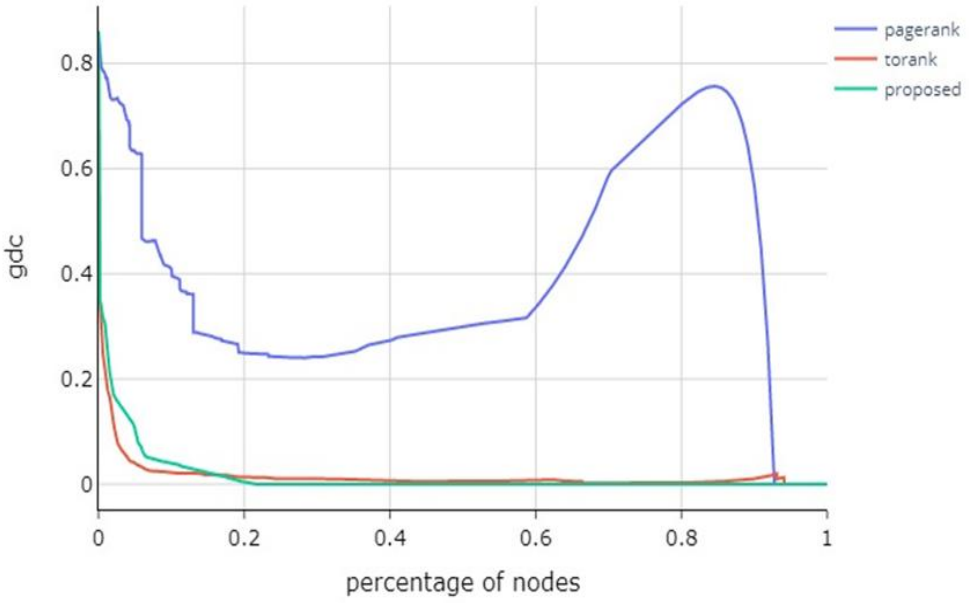

Figure. 6.10: The graph density curve of the three ranking approaches.

Table 6.7 shows the performance of the three ranking algorithms in terms of the four graph robustness metrics. The values of the four metrics for the full graph without any node removal are shown along with the values obtained after removing the top 1^{st}, 5^{th}, 10^{th} and 20^{th} percentile of nodes. The average shortest path and the diameter experiences a significant increase for the proposed ranking algorithm. In fact, the diameter of the graph increased to 26 after removing the top 20% of nodes. The proposed algorithm achieves comparatively better results than the other two approaches on all four metrics.

Table 6.7: Comparison of the proposed ranking technique with the other algorithms.

Algorithms	Nodes Removal	Clustering Coefficient	Average Shortest Path	Giant Component	Diameter
Full Graph	--	0.199	4.37	4761	12
PageRank	Top 1%	0.288	4.42	4689	12
	Top 5%	0.248	4.49	4521	12
	Top 10%	0.221	4.53	3873	13
	Top 20%	0.07	4.55	3325	13
ToRank	Top 1%	0.041	4.85	2997	17
	Top 5%	0.02	5.43	2030	20
	Top 10%	0.011	6.07	752	23
	Top 20%	0.006	7.12	131	25
Proposed	Top 1%	0.051	4.83	3160	19
	Top 5%	0.018	5.78	1606	23
	Top 10%	0.007	6.24	537	25
	Top 20%	0.0	7.69	10	26

The proposed ranking algorithm is an iterative algorithm that iteratively computes the influence score of the nodes based on their connectivity with their adjacent nodes. Initially, a small value of influence score is allocated to each of the nodes, once the algorithm converges, the node with high connectivity to other influential nodes will have a greater individual influence score in the network. Since the proposed ranking algorithm is hyperlink based, the main limitation is its inability to assess the influence score of the isolated nodes in the graph.

CHAPTER 7
CONCLUSIONS

7.1 Summary

The increasing amount of illegal activities on the Tor dark web has made it possible for people to access digital goods and services that could be associated with unlawful activities. The Tor network has been shown to contain a significant number of hidden services with illegal content, including illegal drugs and firearms trafficking, child abuse content and many more things. Such kind of content should be taken down or obstructed to provide a safe online ecosystem for users. Consequently, the law enforcement agencies attempt to confront these activities by every possible means, including automated and intelligent techniques and mechanisms.

To the best of our knowledge, most of the approaches used by the enforcement agencies are either manual or keyword-based filter approaches which are not efficient and effective enough to combat the dynamic environment of the Tor dark web. Hence in this thesis, new methods and techniques using machine learning, natural language processing and graph theory are proposed to monitor the content on the Tor dark web.

Specifically saying, the current work could be presented into an integrated framework consisting of three components: i) Hidden Service Classification, to classify hidden services into one of the predefined categories; ii) Content Based Identification, to identify the most harmful hidden services involved in drug trafficking and those involved in the firearm trade and iii) Link Based Identification, to identify and rank the influential hidden services using their hyperlink connectivity in the

network. The methodology and algorithms proposed in this thesis could help benefits the law enforcement agencies in automatically identifying the outlawed ventures on the Tor dark web.

7.2 Research Contributions

In this work, we put forward a framework consisting of three components to identify illegal content and to detect the most influential hidden services using machine learning techniques. Since the machine learning techniques are purely dependent on the data provided as the input, the presented tools and techniques are not limited only to the Tor network. These techniques could also be extended to be used in other similar areas of concern like the social media platforms and the surface web content. Now, we summarize the conclusions to show the ability of the presented work in monitoring the unlawful activities on the Tor dark web. The main contributions are briefly described as follows:

- **A labeled dataset of the Tor hidden services is constructed:** We constructed a dataset of hidden services; the data is collected by a customized web crawler designed in Python that exclusively connects to the Tor network via SOCKS proxy. The instances of the dataset are labeled manually into one of the 31 categories identified by us. The dataset consists of 4102 samples where each sample represents a unique onion domain.

- **Performed the content analysis of the collected dataset:** The collected dataset was analyzed for the legality of the content present in it. We identified 31 different categories for the content of the dataset of which the categories representing the Bitcoin, drugs trafficking and CC dumps are among the largest. Overall, 38 percent of the total content in the dataset was illegal and the remaining content was legal.

- **Performed the language-wise categorization of the content:** In a first of its kind, the language-wise categorization of the hidden services is performed. We have identified 31 different languages in which the content was present. The 29 percent of the non-English content was in the Russian language followed by German, French and Spanish. Overall, the European languages make up the majority of content. Most of the categories of illegal content identified in non-English were the same as in English content. Russian also accounts for the majority of the illegal content found in the non-English content. Overall, the illegal content in non-English was a bit less than in English content.

- **A text-based supervised machine learning model is proposed for classifying five categories of illegal activities on the Tor dark web:** We explored the key factors that influence the performance of the machine learning classification model. The method of feature engineering may greatly affect the outcome of the text classifier. The three classification models that were explored are Naïve Bayes (NB), Logistic Regression (LR) and the Support Vector Machines (SVM). The application of Term Frequency-Inverse Document Frequency (TF-IDF) for feature representation in combination with the LR model gives the best performance in terms of *f*-score.

- **A two-step dimensionality reduction scheme is proposed for optimal classification results:** The use of an optimal feature set could greatly improve the performance of the classifiers. A two-step dimensionality reduction (DR) scheme is proposed that can produce a feature set with a significant reduction in its size without compromising the performance of the model. In the first step of DR, Mutual Information (MI) is used to select the feature from the original feature set. The MI score of each of the features is computed and ranked from highest to lowest. The top 16 percent of the features are selected for the second step of DR

resulting in a substantial reduction in the size of the feature set. In the second step of DR, the Linear Discriminant Analysis (LDA) is applied to the reduced feature set to transform it into a new space with smaller dimensions. At the end of the two-step DR scheme, the resulting feature space contains n-1 features where n is the total number of classes. The effectiveness of the proposed DR scheme was tested on the dark web text dataset and the Reuters-21,578 dataset. The experimental results have shown that the proposed DR technique has brought significant improvement in the classification performance. The LR classifier was once again the best performer than the NB and SVM classifier.

- **A metric is proposed to estimate the harm level of hidden services involved in drug trafficking on the Tor network:** Some of the hidden services on the Tor may deal in potentially harmful drugs like cocaine and heroin while others may sell less harmful drugs like *khat* or *mushrooms*. We proposed a metric that calculates the harm score of each hidden service based on the type and toxicity of the drug available. The Drug Name Recognition (DNR), a specific type of Natural Language Processing (NLP) task was employed to extract the names of drugs (both common and street names). The harmful effect of the drugs was quantified based on the ratings assigned to them in the existing literature. Then the metric is proposed that can estimate the overall harm of the hidden service based on the harm score of the individual drugs present in it.

- **Ranking of hidden services to identify the most harmful domains:** The above-proposed metric was utilized to rank the hidden services involved in the drug trafficking from the most harmful to least harmful. In case of a tie in rankings, specific criteria were laid down to break the tie between two or more hidden services. The law enforcement agencies may use the ranking list to identify the most harmful hidden

services as they pose a greater risk to the general public than others. The ranking methodology is purely based on the content of the hidden services so that it can easily identify the harmful hidden services if they are being run on other domains with the same content. The ranking methodology could help the enforcement agencies in focusing their efforts on the dangerous hidden services.

- **Generation of ground truth for testing the effectiveness of the ranking methodology:** There is no gold standard or baseline to judge the correctness of the rankings. Therefore, a group of three experts was formed who independently ranked each of the hidden services in the dataset based on three factors. The independent ranking lists were then merged using rank-based aggregation method to obtain a single list free of any bias. The ranked list of the proposed ranking methodology was evaluated against the ground truth generated by the three experts. The ranking lists produced by our methodology give good performance when evaluated using Kendall's tau and rank biased overlap metrics.

- **An ensemble classification model is proposed for detecting the specific firearm listings on the Tor network:** Other than drug trafficking, the trade of illegal weapons on the Tor dark web has become a cause of concern for the law enforcement agencies. The literature review of the firearm trafficking on the Tor and recent terror attacks indicated that the firearms of type pistols and rifles are preferred by the criminals. To detect the pistols and rifle listings on the hidden services, we proposed the ensemble classification models. The classifiers that were used for building ensemble were the Naïve Bayes (NB), Random Forests (RF) and the Logistic Regression (LR). The ensemble was built using the stacking technique. Several feature representation and weighting techniques were explored to identify the

best combinations that could accurately predict the specific firearm listings.

- **We explored the effect of using part of speech tagged features on the performance of the classification model:** Part-of-speech (PoS) tagging is used to assign different tags to the tokens that identified the class of part of speech like noun, verb, adjective, etc. The features that were tagged as nouns were selected from the feature set as many of the attributes and characteristics of the firearm are represented in nouns like its action type, caliber, make of the firearm, variant etc. The noun-tagged features were then used to create the unigrams, bigrams and trigrams and assigned weights using the term frequency-inverse document frequency. The experiment was performed on a subset of a publicly available dataset of Tor cryptomarkets. The PoS tagged features when used in combination with unigrams and bigrams produced the best results on NB and RF classifiers. The stacking ensemble also outperforms the other two classifiers when supplied with PoS tagged feature set and unigrams and bigram. The learning curve of the stacking ensemble classifier shows that it gradually generalizes once the training samples are increased. However, the stacking ensemble requires relatively more time for fitting which is common in such scenarios.

- **We explored the topological properties of the Tor web graph:** A web graph corresponding to the dark web dataset has been constructed where the nodes represent the individual hidden services and edges represent the hyperlinks between the domains. A number of standard graph metrics have been computed for the Tor web graph. The majority of the nodes have out-degree and in-degree less than ten. The in-degree distribution follows the power law which was statistically confirmed while the out-degree distribution does not follow the power law. The average shortest path length of the Tor web graph was 4.32 which is

nearly the same as of surface web. The undirected version of the Tor web graph was disconnected with eight connected components. The largest component contains nearly all the nodes in the graph. The directed graph contains the four strongly connected components of considerable size while the remaining components were less than ten in size. The bow-tie decomposition of the Tor web graph reveals that it differs from the typical structure as found in the surface web. Most of the components of the bow-tie structure in the Tor web graph are either very large or very small in comparison to the surface web. However, both the Tor dark web and the surface web contain nearly the same percentage of the isolated nodes. The connectivity to the surface websites uncovers that the Tor network has outgoing links to popular surface websites like Amazon, Google, Twitter etc. Most of the links were going to social media sites, news, content management sites and adult content. The top-level domain (TLD) with the *.com* extension has the largest number of links followed by *.org*. The German (.de) and Russian (.ru) were at the top in country-wise TLD.

- **We proposed a link based ranking algorithm for identifying the influential Tor hidden services:** The content based identification techniques could only identify the hidden services dealing in the illicit drugs and firearm trade. However, the link based algorithms could detect any type of hidden services irrespective of its content. A hyperlink based iterative algorithm is proposed to detect the influential Tor hidden services. The influence of a particular hidden service is governed by its location in the Tor network and connectivity within the network as well as to the surface web. The connectivity of the node to other influential services may also affect the overall influence. These factors could be embedded into an algorithm to calculate the overall influence of the hidden service. The proposed ranking algorithm is based on the modified PageRank algorithm and uses the Tor web graph

to calculate the influence of each of the hidden services. The performance of the proposed algorithm is compared with the PageRank and ToRank algorithms on various graph robustness metrics. The proposed algorithm outperforms the other two algorithms on all the standard metrics. The link based identification could be used in combination with the content based technique to detect suspicious activities on the Tor network.

7.3 Future Directions

In this section, we present some research areas that come out which could be addressed in future work or taken up by other researchers.

- **Exploring other dark web networks:** In the present work, we have explored and addressed only the Tor dark web network, however, there are other dark web networks like Freenet and I2P which may be contain suspicious activities and needs to be monitored.

- **Utilizing the non-textual content along with the textual content to enhance the classification performance:** The hidden services classification component proposed in this thesis only uses the text-based features. The incorporation of non-textual features like the images and other visual content could help improve the performance of the classifier.

- **Modifying the classifier for classifying the multi-label hidden services:** A large number of cryptomarkets on the Tor dark web deal in multiple illegal products and services. The current classification model does not account for such cryptomarkets in the training data, therefore the extension of the classifier to include these samples would greatly enhance its usability.

- **Exploring the nature inspired algorithms for creating an optimal feature for the classifier:** The nature inspired and evolutionary algorithms have recently gained popularity because of their ability to optimize the objective function. They can also be used for feature selection problems to generate the optimal feature space while improving the overall classification performance.

- **Extending the content based identification for other illegal activities:** The content based identification component was designed specifically for illicit drugs and firearm trafficking. However, there are other illegal activities like trafficking of exotic wildlife products, counterfeits, etc that also need attention. The content based identification can be extended to identify the illegal activities apart from drugs and firearms.

- **Enhancing the link based ranking algorithms with the content based features:** Currently, the link based ranking algorithm for identifying the influential hidden services is purely based on the hyperlink connectivity between the domains. The presence of multimedia content like images, audio and video could be utilized with the existing link based algorithm to further improve the rankings.

REFERENCES

A. Contributions Arising from the Research Reported

1.) M. Faizan and R. A. Khan, “Exploring and analyzing the dark web: A new alchemy,” *First Monday*, vol. 24, no. 5, 2019.

2.) M. Faizan and R. A. Khan, “A Two-Step Dimensionality Reduction Scheme for Dark Web Text Classification,” in *Proc. Recent Advancement in Computer, Communication and Computational Sciences*, Ajmer, India, 2019, pp. 303-312.

3.) M. Faizan, R. A. Khan and A. Agrawal, “Ranking potentially harmful Tor hidden services: Illicit drugs perspective,” *Applied Computing and Informatics*, 2020.

4.) M. Faizan and R. A. Khan, “A Survey of the Illegal Activities on the Tor Dark Web,” in *Proc. Cyber Crime and Ethical Hacking*, New Delhi, India, 2020.

5.) M. Faizan *et al.*, “Exploring the Topological Properties of the Tor Dark Web,” *IEEE Access*, vol. 9, pp. 21746-21758, 2021.

6.) M. Faizan *et al.*, “An Ensemble Approach to Identify Firearm Listing on Tor Hidden Services,” *Computer Systems Science and Engineering*, vol. 38, no. 2, pp. 141-149, 2021.

7.) M. Faizan *et al.*, “A Link Analysis Algorithm for Identification of Key Hidden Services,” *Computers, Materials & Continua*, vol. 68, no. 1, pp. 877-886, 2021.

B. Main References

[1] M. Pannu, I. Kay and D. Harris, "Using Dark We Crawler to Uncover Suspicious and Malicious Websites," in *Proc. International Conference on Human Factors in Cybersecurity*, Orlando, FL, USA, 2018, pp. 108-115.

[2] D. Moore and T. Rid, "Cryptopolitik and the Darknet," *Survival*, vol. 58, no. 1, pp. 7–38, 2016.

[3] M.G. Reed, P.F. Syverson and D,M, Goldschlag, "Anonymous connections and onion routing," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 4, pp. 482-494, 1998.

[4] R. Graham and B. Pitman, "Freedom in the Wilderness: A Study of a Darknet space," *Convergence*, vol. 26, no. 3, pp. 593–619, June 2020.

[5] R. Dingledine, N. Mathewson and P. Syverson, "Tor: The Second-Generation Onion Router," in *Proc. 13th USENIX Security Symposium*, 2004, pp. 21.

[6] C. Quintin. "7 Things You Should Know About Tor." Electronic Frontier Foundation. https://www.eff.org/deeplinks/2014/07/7-things-you-should-know-about-tor (accessed March 20, 2021).

[7] G. Greenwald, E. MacAskill and L. Poitras. "Edward Snowden: the whistleblower behind the NSA surveillance revelations." The Guardian. https://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance (accessed March 20, 2021).

[8] M. Brown. "Edward Snowden: the true story behind his NSA leaks." TheTelegraph. https://www.telegraph.co.uk/culture/film/11185627/EdwardSnowden-the-true-story-behind-his-NSA-leaks.html (accessed March 20, 2021).

[9] E. Macaskill and G. Dance. "NSA Files: Decoded." The Guardian. https://www.theguardian.com/world/interactive/2013/nov/01/snowde

n-nsa-files-surveillance-revelations-decoded#section/1 (accessed March 20, 2021).

[10] E. Jardine, "The dark Web dilemma: Tor anonymity and online policing," *SSRN Electron. J.*, vol. 21, pp. 1-24, Dec. 2015, [online] Available: https://www.cigionline.org/sites/default/files/no.21.pdf

[11] M. Chertoff and T. Simon, "The impact of the dark Web on Internet governance and cyber security," *Global Commission Internet Governance*, vol. 6, pp. 1-18, May 2015, [online] Available: https://www.cigionline.org/sites/default/files/gcig_paper_no6.pdf.

[12] E. Stacey, *Combating Internet-Enabled Terrorism: Emerging Research and Opportunities*. Hershey, PA, USA: IGI Global, 2017.

[13] A. Greenberg. "Hacker lexicon: What is the dark Web?" Wired. https://www.wired.com/2014/11/hacker-lexicon-whats-dark-web/ (accessed March 25, 2021).

[14] I. Sanchez-Rola, D. Balzarotti and I. Santos, "The onions have eyes: A comprehensive structure and privacy analysis of Tor hidden services," in *Proc. 26th International Conference on World Wide Web*, pp. 1,251–1,260, 2017.

[15] N. Bertrand. "ISIS is taking full advantage of the darkest corners of the Internet," Business Insider. https://www.businessinsider.com/isis-is-using-the-dark-web-2015-7 (accessed March 27, 2021).

[16] E. Jardine, A. M. Lindner and G. Owenson, "The potential harms of the Tor anonymity network cluster disproportionately in free countries," *Proceedings of the National Academy of Sciences*, vol. 117, no. 50, pp. 31716-31721, 2020.

[17] P.-H. Meland, Y.F.F. Bayoumy and G. Sindre, "The Ransomware-as-a-Service economy within the darknet," *Computers &Security*, vol. 92, 2020.

[18] N. Christin, "Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace," in *Proc. 22nd international conference on World Wide Web*, 2013, pp. 213-224.

[19] K. Soska and N. Christin, “Measuring the longitudinal evolution of the online anonymous marketplace ecosystem,” in *Proc. 24th USENIX Secur. Symp.*, 2015, pp. 33-48.

[20] T. Moore and N. Christin, “Beware the middleman: Empirical analysis of Bitcoin-exchange risk,” in *Proc. International Conference on Financial Cryptography and Data Security*, Okinawa, Japan, 2013, pp. 25-33.

[21] M.C. Van Hout and T. Bingham, “Responsible vendors, intelligent consumers: Silk Road, the online revolution in drug trading,” *International Journal of Drug Policy*, vol. 25, no. 2, pp. 183-189, 2014.

[22] A. Phelps and A. Watt, “I shop online-recreationally! Internet anonymity and Silk Road enabling drug use in Australia,” *Digital Investigation*, vol. 11, no. 4, pp. 261-272, 2014.

[23] M.C. Van Hout and T. Bingham, “Surfing the Silk Road: A study of users’ experiences,” *International Journal of Drug Policy*, vol. 24, no. 6, pp. 524-529, 2013.

[24] M.J. Barratt, J.A. Ferris and A.R. Winstock, “Use of Silk Road, the online drug marketplace, in the United Kingdom, Australia and the United States,” *Addiction*, vol. 109, no. 5, pp. 774-783, 2014.

[25] E. Ormsby, “Silk Road: insights from interviews with users and vendors,” in *The Internet and drug markets*. Lisbon, Portugal: EMCDDA, 2016, pp.61-68.

[26] M.J. Barratt, J.A. Ferris and A.R. Winstock , “Safer scoring? Cryptomarkets, social supply and drug market violence,” *International Journal of Drug Policy*, vol. 35, pp. 24-31, 2016.

[27] K. Zetter. “How the Feds Took Down the Silk Road Drug Wonderland.” Wired. https://www.wired.com/2013/11/silk-road/#:~:text=The%20informant%20told%20DHS%20investigators,stirrers%20or%20underwear%20on%20Amazon (accessed March 20, 2021).

[28] U.S. Attorney's Office. (2015). Ross Ulbricht, AKA Dread Pirate Roberts, sentenced in Manhattan Federal Court to life in prison. [Online]. Available: https://www.justice.gov/usao-sdny/pr/ross-ulbricht-aka-dread-pirate-roberts-sentenced-manhattan-federal-court-life-prison

[29] A. Greenberg, A. “Global web crackdown arrests 17, seizes hundreds of darknet domains.” Wired. https://www.wired.com/2014/11/operation-onymous-dark-web-arrests/ (accessed March 25, 2021).

[30] U.S. Attorney's Office. (2014). Operator of Silk Road 2.0 website charged in Manhattan Federal Court. [Online]. https://www.fbi.gov/

[31] E. Harfenist and M. Turgeman, M. “Dutch police open dark net site to spook vendors and buyers.” Vocative. https://www.vocativ.com/372640/dutch-police-open-dark-net-site-to-spook-vendors-and-buyers/index.html (accessed March 25, 2021).

[32] Europol. Massive blow to criminal dark web activities after globally coordinated operation. Europol. https://www.europol.europa.eu/newsroom/news/massive-blow-to-criminal-dark-web-activities-after-globally-coordinated-operation (accessed March 25, 2021).

[33] J. Aldridge and D. Décary-Hétu, “A response to Dolliver's “Evaluating drug trafficking on the Tor Network: Silk Road 2, the sequel,”” *International Journal of Drug Policy*, vol. 26, no. 11, pp. 1124-1125, 2015.

[34] D. Décary-Hétu and L. Giommoni, “Do police crackdowns disrupt drug cryptomarkets? A longitudinal analysis of the effects of Operation Onymous,” Crime, Law and Social Change, vol. 67, pp. 55–75, 2017.

[35] G. Owenson, S. Cortes and A. Lewman, “The Darknet’s smaller than we thought: The life cycle of Tor hidden services,” *Digit. Invest.*, vol. 27, pp. 17-22, 2018.

[36] S. Foley, J. Karlsen and T.J. Putnins, “Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies?” *Review of Financial Studies*, 2018.

[37] M.W. Al-Nabki, E. Fidalgo, E. Alegre and L. Fernandez-Robles, “ToRank: Identifying the Most Influential Suspicious Domains in the Tor Network,” *Expert Systems with Applications*, 2019.

[38] C. Guitton, “A review of the available content on Tor hidden services: The case against further development,” *Computers in Human Behavior*, vol. 29, no. 6, pp. 2805–2815, 2013.

[39] A. Biryukov, I. Pustogarov, F. Thill and R.-P. Weinmann, “Content and popularity analysis of Tor hidden services,” in *Proc. 34th International Conference on Distributed Computing Systems Workshops*, 2014, pp. 188–193.

[40] M. Spitters, S. Verbruggen and M. Staalduinen, “Toward a comprehensive insight into the thematic organization of Tor hidden services,” in *Proc. 2014 IEEE Joint Intelligence and Security Informatics Conference*, 2014, pp. 220–223.

[41] G. Owen and N. Savage, “Empirical analysis of Tor Hidden Services,” *IET Inf. Secur.*, vol. 10, pp. 113-118, 2015.

[42] “Deeplight: Shining a light on the dark Web,” Intelliagg, 2016. [Online]. Available: https://media.scmagazine.com/documents/224/deeplight_(1)_55856.pdf

[43] M.W. Al Nabki, E. Fidalgo, E. Alegre and I. de Paz, “Classifying illegal activities on Tor network based on Web textual contents,” in *Proc.15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 35–43.

[44] M. Mohamad and A. Selamat, “An evaluation on the efficiency of hybrid feature selection in spam email classification,” in *Proc. International Conference on Computer, Communications, and Control Technology (I4CT)*, IEEE, 2015, pp. 227–231.

[45] H. Chen, S. Mckeever and S.J. Delany, “Harnessing the power of text mining for the detection of abusive content in social media,” in *Advances in Computational Intelligence Systems*, Cham: Springer, 2017, pp. 187–205.

[46] B. Pang, *et al.*, “Opinion mining and sentiment analysis,” *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp.1–135, 2008.

[47] B. Ohana, S.J. Delany and B. and Tierney, “A Case-Based Approach to Cross Domain Sentiment Classification,” in *Proc. International Conference on Case-Based Reasoning*, Springer, Berlin, Heidelberg 2012, pp. 284–296.

[48] Y. Li, A. Tripathi and A. Srinivasan, “Challenges in short text classification: The case of online auction disclosure,” in *Proc. Mediterranean Conference On Information Systems* (MCIS, 2016), 2016.

[49] A. Sun, E. Lim and W. Ng, “Web classification using support vector machine,” in *Proc. 4th international workshop on Web information and data management*, 2002, pp. 96–99.

[50] M. Kan and H. Thi, “Fast webpage classification using url features,” in *Proc.14th ACM international conference on Information and knowledge management*, 2005, pp. 325– 326.

[51] M. Graczyk and K. Kinningham, “Automatic product categorization for anonymous Marketplaces,” 2015.

[52] T. Sabbah, A. Selamat, M.H. Selamat, R. Ibrahim and H. Fujita, “Hybridized term weighting method for dark web classification,” *Neurocomputing*, vol.1 73, pp. 1908 – 1926, 2016.

[53] G. Durrett, *et al.*, “Identifying products in online cybercrime marketplaces: A dataset for fine-grained domain adaptation,” in *Proc. 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 2598-2607.

[54] J. Dalins, C. Wilson and M. Carman, “Criminal motivation on the dark web: A categorisation model for law enforcement,” *Digital Investigation*, vol. 24, pp. 62-71, 2018.

[55] J. Park, H. Mun and Y. Lee, “Improving tor hidden service crawler performance,” in *Proc. 2018 IEEE Conference on Dependable and Secure Computing (DSC)*, 2018, pp 1–8.

[56] S. He, Y. He and M. Li, “Classification of illegal activities on the dark web,” in *Proc. 2019 2Nd International Conference on Information Science and Systems*, ACM, New York, NY, USA, 2019, pp. 73–78.

[57] M. Perry, “Torflow: Tor network analysis,” 2009. [Online]. Available: https://research.torproject.org/techreports/torflow-2009-08-07.pdf

[58] R. Snader *et al.*, “A Tune-up for Tor: Improving Security and Performance in the Tor Network,” 2008. [Online]. Available: https://www.internetsociety.org/doc/tune-tor-improving-security-and-per formance-tor-network-paper

[59] Z. Weinberg, *et al.*, “StegoTorus: A camouflage proxy for the tor anonymity system,” in *Proc ACM Conference on Computer and Communications Security (CCS’12)*, ACM, New York, 2012, pp. 109–120.

[60] D. Arp, F. Yamaguchi and K. Rieck, “Torben: A practical side-channel attack for deanonymizing Tor communication,” in *Proc. 10th ACM Symposium on Information,Computer and Communications Security (ASIACCS’15)*, ACM, New York, 2015, pp. 597–602.

[61] V. Griffith, Y. Xu and C. Ratti, “Graph Theoretic Properties of the Darkweb,” *arXiv preprint*, 2017.

[62] M. Bernaschi, A. Celestini, S. Guarino and F. Lombardi, “Exploring and Analyzing the Tor Hidden Services Graph,” *ACM Trans. Web*, vol. 11, no. 4, pp. 24:1-24:26, 2017.

[63] M. Bernaschi, A. Celestini, S. Guarino, F. Lombardi and E. Mastrostefano, “Spiders like Onions: on the Network of Tor Hidden Services,” in *Proc. The World Wide Web Conference (WWW ’19)*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 105–115.

[64] B. Monk, J. Mitchell, R. Frank and G. Davies, G, “Uncovering Tor: An examination of the network structure,” *Security and Communication Networks*, 2018.

[65] M. Zamani, *et al.*, “Differences in structure and dynamics of networks retrieved from dark and public web forums,” *Physica A: Statistical Mechanics and its Applications*, vol. 525, pp. 326-336, 2019.

[66] R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, “Trawling the Web for emerging cyber-communities,” *Computer Networks*, vol. 31, no. 11-16, pp. 1481-1493, 1999.

[67] J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan and A.S. Tomkins, “The Web as a Graph: Measurements, Models, and Methods,” in *Proc. Computing and Combinatorics. COCOON 1999*, Springer, Berlin, Heidelberg, 1999, pp. 1-17.

[68] A. Broder, *et al.*, “Graph structure in the Web,”, *Computer Networks*, vol. 33, no. 1-6, pp. 309-320, 2000.

[69] R. Meusel, S. Vigna, O. Lehmberg and C. Bizer, “The Graph Structure in the Web – Analyzed on Different Aggregation Levels,” *The Journal of Web Science*, vol. 1, no. 1, pp. 33-47, 2015.

[70] A.-L. Barabási, R. Albert and H. Jeong, “Scale-free characteristics of random networks: The topology of the world-wide Web,” *Phys. A Stat. Mech. Appl.*, vol. 281, no. 1, pp. 69-77, Jun. 2000.

[71] O. Lehmberg, R. Meusel and C. Bizer, “Graph structure in the Web: Aggregated by pay-level domain,” in *Proc. ACM Conf. Web Sci.*, 2014, pp. 119-128.

[72] M. Serrano, A. Maguitman, M. Boguñá, S. Fortunato and A. Vespignani, "Decoding the Structure of the WWW: A Comparative Analysis of Web Crawls," *ACM Transactions on the Web*, vol. 1, no. 2, pp. 1-25, 2007.

[73] J. Broséus, *et al.*, "Studying illicit drug tracking on Darknet markets: Structure and organisation from a Canadian perspective," *Forensic Science International*, vol. 264, pp.7-14, 2016.

[74] J. Broséus, D. Rhumorbarbe, M. Morelato, L. Staehli and Q. Rossy, "A geographical analysis of tracking on a popular darknet market," *Forensic Science International*, vol. 277, pp. 88-102, 2017.

[75] J. Van Buskirk, A. Roxburgh, M. Farrell and L. Burns, "The closure of the Silk Road: What has this meant for online drug trading?" *Addiction*, vol. 109, no. 4, pp. 517-518, 2014.

[76] J. Aldridge and D. Décary-Hétu, "Not an 'Ebay for Drugs': The Cryptomarket 'Silk Road' as a Paradigm Shifting Criminal Innovation," 2014.

[77] R.A. Hardy and J.R. Norgaard, "Reputation in the Internet black market: an empirical and theoretical analysis of the Deep Web," *Journal of Institutional Economics*, vol.12, no. 3, pp. 515-539, 2016.

[78] J. Nurmi, T. Kaskela, J. Perälä and A. Oksanen, "Seller's reputation and capacity on the illicit drug markets: 11-month study on the Finnish version of the Silk Road," *Drug and Alcohol Dependence*, vol. 178, pp. 201-207, 2017.

[79] J. Van Buskirk, S. Naicker, A. Roxburgh, R. Bruno and L. Burns, L, "Who sells what? Country specific differences in substance availability on the Agora cryptomarket," *International Journal of Drug Policy*, vol. 35, pp. 16-23, 2016.

[80] D.S. Dolliver and J.L. Kenney, "Characteristics of Drug Vendors on the Tor Network: A Cryptomarket Comparison," *Victims & Offenders*, vol. 11, no. 4, pp. 600-620, 2016

[81] M. Dittus, J. Wright and M. Graham, “Platform Criminalism: The `Last-Mile' Geography of the Darknet Market Supply Chain,” in *Proc. World Wide Web Conference (WWW '18)*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, pp. 277–286..

[82] G.P. Paoli, et al., “Behind the Curtain: The Illicit Trade of Firearms, Explosives and Ammunition on the Dark Web,” RAND Corporation, California, USA, RR-2091-PACCS, 2017. Accessed: Mar. 30, 2021. [Online]. Available: https://www.rand.org/pubs/research_reports/RR2091.html

[83] C. Copeland, M. Wallin and T.J. Holt, “Assessing the Practices and Products of Darkweb Firearm Vendors,” *Deviant Behavior*, vol. 41, no. 8, 2019.

[84] D. Rhumorbarbe, *et al.*, “Characterising the online weapons trafficking on cryptomarkets,” *Forensic Science International*, vol. 283, pp. 16–20, 2018.

[85] J.K Saini and D. Bansal, “A Comparative Study and Automated Detection of Illegal Weapon Procurement over Dark Web,” *Cybernetics and Systems*, vol. 50, no. 5, pp. 405-416, 2019.

[86] Van Wegberg, *et al.*, “Plug and prey? measuring the commoditization of cybercrime via online anonymous markets,” in *Proc. 27th USENIX Conference on Security Symposium (SEC'18)*, USENIX Association, USA, 2018, pp. 1009–1026.

[87] F. Ferrara. “Sanctioned Suicide, il Deep Web più frequentato dagli aspirant suicide.” Fidelity House. https://news.fidelityhouse.eu/web/sancioned-suicide-il-deep-web-piu-frequentato-dagli-aspiranti-suicidi-185027.html 5–8 (accessed March 27, 2021).

[88] Franceschi-Bicchierai. “The deep web suicide site.” Vice. https://www.vice.com/ (accessed March 27, 2021).

[89] J. Bartlett, *The Dark Net: Inside the Digital Underworld*. Brooklyn, New York, USA: Melville House, 2016.

[90] C. Mörch, *et al.*, "The Darknet and Suicide," *Journal of Affective Disorders*, 2018.

[91] INTERPOL. "Research identifies illegal wildlife trade on the Darknet". https://www.interpol.int/ (accessed April 03, 2021).

[92] A. Lavorgna, "Wildlife trafficking in the Internet age," *Crime Science*, vol. 3, no. 5 2014.

[93] D.S. Dolliver, "Evaluating drug trafficking on the tor network: Silk road 2, the sequel," *International Journal of Drug Policy*, vol. 26, no. 11, pp. 1113–1123, 2015.

[94] M.J. Barratt and J. Aldridge, "Everything you always wanted to know about drug cryptomarkets* (*but were afraid to ask)," *International Journal of Drug Policy*, vol. 35, pp. 1–6, 2016.

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

[96] E. Nunes, *et al.*, "Darknet and deepnet mining for proactive cybersecurity threat intelligence," in *Proc.2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, 2016, pp. 7–12.

[97] A. Bidoki, P. Ghodsnia, N. Yazdani and F. Oroumchian, "A3CRank: An adaptive ranking method based on connectivity, content and click-through data," *Information Processing & Management*, vol. 46, no. 2, pp. 159 – 169, 2010.

[98] V. Derhami, E. Khodadadian, M. Ghasemzadeh and A. Bidoki, "Applying reinforcement learning for web pages ranking algorithms," *Applied Soft Computing*, vol. 13, no. 4, pp. 1686–1692, 2013.

[99] T. Anwar and M. Abulaish, "Ranking radically influential web forum users," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 6, pp. 1289–1298, 2015.

[100] X. Xu, C. Zhou and Z. Wang, "Credit scoring algorithm based on link analysis ranking with support vector machine," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2625–2632, 2009.

[101] A. F. Colladon and E. Remondi, “Using social network analysis to prevent money laundering,” *Expert Systems with Applications*, vol. 67, pp. 49–58, 2017.

[102] H. Chen, *Dark web: Exploring and data mining the dark side of the web*. New York, USA: Springer-Verlag, 2012.

[103] Y. Zhang, Y. Bao, S. Zhao, J. Chen and J. Tang, “Identifying node importance by combining betweenness centrality and katz centrality,” in *Proc. International Conference on Cloud Computing and Big Data (CCBD)*, 2015, pp. 354–357.

[104] M. Nouh and J.R. Nurse, “Identifying Key-Players in Online Activist Groups on the Facebook Social Network,” in *IEEE International Conference on Data Mining Workshop (ICDMW)*, Atlantic City, NJ, USA, 2015, pp. 969–978.

[105] Y. Hu, J. Zhang, X. Bai, S. Yu and Z. Yang, Z, “Influence analysis of Github repositories,” *SpringerPlus*, vol. 5, pp. 1268. 2016.

[106] Y. Hu, S. Wang, Y. Ren, and K. Choo, “User influence analysis for Github developer social networks,” *Expert Systems with Applications*, vol. 108, pp. 108–118, 2018.

[107] A.B. Eliacik and N. Erdogan, “Influential user weighted sentiment analysis on topic based microblogging community,” *Expert Systems with Applications*, vol. 92, pp. 403–418, 2018.

[108] I. Anger and C. Kittl, “Measuring influence on twitter,” in *Proc. 11th International Conference on Knowledge Management and Knowledge Technologies*, ACM, New York, NY, USA, 2011, pp 1-4.

[109] H. Chen, *et al.*, “Uncovering the dark Web: A case study of Jihad on the Web,” *J. Am. Soc. Inf. Sci.*, vol. 59, pp. 1347–1359, 2008.

[110] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Mårtenson and P. Svenson, “Analysis of Weak Signals for Detecting Lone Wolf Terrorists,”in *Proc. 2012 European Intelligence and Security Informatics Conference*, Odense, Denmark, 2012, pp. 197-204.

[111] G. Weimann, “Going Dark: Terrorism on the Dark Web,” *Stud. Confl. Terror*, vol. 39, pp. 195–20, 2016.

[112] J. Cossu, N. Dugué and V. Labatut, “Detecting Real-World Influence through Twitter,” in *Proc. 2015 Second European Network Intelligence Conference*, Karlskrona, Sweden, 2015, pp. 83-90.

[113] F. Sebastiani, “Machine learning in automated text categorization,” *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.

[114] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 5, pp. 537–550, 1994.

[115] W. Shang, *et al.*, “A novel feature selection algorithm for text categorization,” *Expert Systems with Applications*, vol. 33, pp. 1–5, 2007.

[116] I. Kononenko, “Estimating attributes: Analysis and extensions of relief,” in *Proc. European conference on machine Learning*, 1994, pp. 171–182.

[117] Y. Li, C. Luo and S. Chung, “Text clustering with feature selection by using statistical data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, pp. 641–652, 2008.

[118] L. Liu, *et al.*, “A comparative study on unsupervised feature selection methods for text clustering,” in *Proc. IEEE international conference on natural language processing and knowledge engineering*, China, 2005, pp. 597–601.

[119] T. Mitchel, *Machine learning*. New York, USA: McGraw-Hill, 1997.

[120] X. He, D. Cai and P. Niyogi, “Laplacian score for feature selection,” in *Proc. Neural Information Processing Systems*, 2005, pp. 505–512.

[121] A. Ferreira and M. Figueired, “Efficient feature selection filters for high-dimensional data,” *Pattern Recognition Letters*, vol. 33, pp. 1794–1804, 2012.

[122] T. Jolliffe, *Principal component analysis*. New York, USA: Springer-Verlag, 2002.

[123] W. Song and S. Park, “Genetic algorithm for text clustering based on latent semantic indexing, *Computers and Mathematics with Applications*, vol. 57, pp. 1901–1907, 2009.

[124] R.A. Fisher, “The statistical utilization of multiple measurements,” *Annals of Human Genetics*, vol. 8, no. 4, pp. 376–386, 1938.

[125] D. Koblas and M.R. Koblas, “SOCKS,” in *Proc. Usenix Security Symposium*, 1992, pp. 77-83.

[126] S. Grigonis, “EU in the face of migrant crisis: Reasons for ineffective human rights protection,” International Comparative Jurisprudence, vol. 2, no. 2, pp. 93-98, 2016.

[127] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. New Jersey, USA: Wiley-Interscience, 2006.

[128] T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009.

[129] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” in *Proc. ECML*, 1998, pp. 137-142.

[130] S. Tong and D. Koller, “Support Vector Machine Active Learning with Applications to Text Classification,” *Journal of Machine Learning Research*, vol. 2, pp. 45-66, 2002.

[131] D. Chen, H. Bourlard and J.P. Thiran, “Text identification in complex background using SVM,” in *Proc. International Conference on Computer Vision and Pattern Recognition*, 2001, pp. 621-626.

[132] D. Koller and M. Sahami, “Hierarchically classifying documents using very few words,” in *Proc. Fourteenth International Conference on Machine Learning*, 1997, pp. 170– 178.

[133] D.D. Lewis and W.A. Gale, “A sequential algorithm for training text classifiers,” in: *Proc. 17th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 3–12.

[134] D.D. Lewis, Reuters-21578 Text Categorization Collection, distribution 1.0. 1997. [Online]. Available: http://kdd.ics.uci.edu/ databases/reuters21578/reuters21578.html/

[135] J. Yang, *et al.*, “A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization,” *Information Processing & Management*, vol. 48, pp. 741-754, 2012.

[136] R.H.W. Pinheiro, *et al.*, “A global-ranking local feature selection method for text categorization,” *Expert Systems with Applications*, vol. 39, pp. 12851-12857, 2012.

[137] L. Man, *et al.*, “Supervised and traditional term weighting methods for automatic text categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 721–735, 2009.

[138] C.J. Van Rijsbergen, *Information Retrieval*. MA, USA: Butterworth-Heinemann, 1979.

[139] S. Liu, B. Tang, Q. Chen and X. Wang, “Drug Name Recognition: Approaches and Resources,” *Information*, vol. 6, no. 4, pp. 790-810, 2015.

[140] G. Navarro, “A guided tour to approximate string matching,” *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31-88, 2001.

[141] D.J. Nutt, L.A. King and L.D. Phillips, “Drug harms in the UK: a multicriteria decision analysis,” *Lancet*, vol. 376, pp. 1558–1565, 2010.

[142] H.D. Kim, C. Zhai and J. Han, "Aggregation of multiple judgments for evaluating ordered lists," in *Proc. European Conference on Information Retrieval*, Milton Keynes, United Kingdom, 2010, pp. 166-178.

[143] M.G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, pp. 239–251, 1945.

[144] C.C. Yang, X. Tang and B.M. Thuraisingham, "An analysis of user influence ranking algorithms on dark web forums," in *Proc. ISI-KDD*, 2010, pp. 10:1–10:7.

[145] Q. Wang, Y. Jin, S. Cheng and T. Yang, "ConformRank: a conformity-based rank for finding top-k influential users," *Phys. A*, 2016.

[146] J. Dai *et al.*, "Identifying influential nodes in complex networks based on local neighbor contribution," *IEEE Access*, vol. 7, pp. 131719–131731, 2019.

[147] N. Wang, Q. Sun, Y. Zhou and S. Shen, "A study on influential user identification in online social networks," *Chin. J. Electron.*, vol. 25, no. 3, pp. 467–473, 2016.

[148] C. Croux and C. Dehon, "Influence functions of the Spearman and Kendall correlation measures," *Stat. Methods Appl.*, vol. 19, pp. 497–515, 2010.

[149] W. Webber, A. Moffat and J. Zobel, "A similarity m0easure for indefinite rankings," *ACM Trans. Inf.* Syst., vol. 28, no. 4, pp. 1-38, 2010.

[150] G.S. Shieh, "A weighted Kendall's tau statistic," *Statist. Probabil. Lett.*, vol. 39, no. 1, pp. 17-27,1998.

[151] *Python*. (2016), Python Software Foundation. Accessed: Feb. 12, 2021. [Online]. Available: https://www.python.org/downloads/release/python-360/

[152] S. Morris. "Teenager obsessed with mass shootings jailed for buying gun online." The Guardian. https://www.theguardian.com/

[153] J. A. Fox and M. J. Delateur, “Mass shootings in America: moving beyond newtown,” *Homicide Studies*, vol. 18, no. 1, pp. 125–145, 2014.

[154] M. Karthik and M. Davis, “Search using n-gram technique based statistical analysis for knowledge extraction in case based reasoning systems,” 2004. [Online]. Available: https://arxiv.org/abs/cs/0407009

[155] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[156] G. Biau and E. Scornet, “A random forest guided tour,” *TEST*, vol. 25, pp. 197–227, 2016.

[157] S. Džeroski and B. Ženko, “Is Combining Classifiers with Stacking Better than Selecting the Best One?” *Machine Learning*, vol. 54, pp. 255–273, 2004.

[158] S Elayidom, “Design and development of data mining models for the prediction of manpower placement in the technical domain,” Ph.D. dissertation, Deptt. Computer Science, Cochin University of Science and Technology, Kochi, Kerala, India, 2012. [Online]. Available: https://shodhganga.inflibnet.ac.in/bitstream/10603/7989/15/15_chapter%206.pdf

[159] G., Branwen, Dark Net Market archives, 2011-2015. 2015. [Online]. Available: https://www.gwern.net/DNM-archives/

[160] O. Cherqi, G. Mezzour, M. Ghogho and M. Elkoutbi, “Analysis of hacking related trade in the darkweb,” in *Proc. IEEE International Conference on Intelligence and Security Informatics (ISI)*, Miami, FL, USA, 2018, pp. 79-84.

[161] A., Berman and C. L. Paul, “Making Sense of Darknet Markets: Automatic Inference of Semantic Classifications from Unconventional Multimedia Datasets,” in *Proc. International*

Conference on Human-Computer Interaction, Orlando, FL, USA, 2019, pp. 230-248.

[162] A. Abeill´e, *Treebanks: Building and Using Parsed Corpora*. Netherlands: Springer, 2003.

[163] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[164] D.J. Watts and S.H. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, pp. 440–442, 1998.

[165] P. Holme, B.J. Kim, C.N. Yoon and S.K. Han, "Attack vulnerability of complex networks," *Physical Review E*, vol. 65, no. 5, pp. 56-109, 2002.

[166] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422/.

[167] A. Sarkhel. "The Deep Dark Side of Web! How People Are Getting Drugs Guns Delivered at Doorstep. The Economic Times. https://economictimes.indiatimes.com/tech/internet/the-deep-dark-side-of-web-how-people-are-getting-drugs-guns-delivered-at-doorstep/articleshow/53407720.cms/.

[168] A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Sci. Amer.*, vol. 288, no. 5, pp. 60-69, 2003.

[169] S. Milgram, "The small world problem," *Psychol. Today*, vol. 1, no. 1, pp. 61-67, May 1967.

[170] A. Barabási, "The physics of the Web," *Phys. World*, vol. 14, no. 7, pp. 33-38, 2001.

[171] M. Chertoff, "A public policy perspective of the Dark Web," *J. Cyber Policy*, vol. 2, no. 1, pp. 26-38, 2017.

[172] *Mail*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Mail.Ru/.

[173] M. Paquet-Clouston, D. Décary-Hétu and C. Morselli, “Assessing market competition and vendors’ size and scope on AlphaBay,” *Int. J. Drug Policy*, vol. 54, pp. 87-98, 2018.

[174] R. A. Hanneman and M. Riddle. *Introduction to Social Network Methods*. Riverside, CA, USA: Univ. California Riverside Press, 2005.

[175] W. McKnight. *Information Management*. MA, United States: Morgan Kaufmann, 2014.

[176] Sunil Kumar Raghavan Unnithan, Balakrishnan Kannan, Madambi Jathavedan, "Betweenness Centrality in Some Classes of Graphs", International Journal of Combinatorics, vol. 2014, Article ID 241723, 12 pages, 2014. https://doi.org/10.1155/2014/241723

[177] J. Powell and M. Hopkins. *A Librarian's Guide to Graphs, Data and the Semantic Web*. Netherlands: Elsevier, 2015.

[178] P. Parau, C. Lemnaru, M. Dinsoreanu and R. Potolea, “Opinion Leader Detection,” in *Sentiment Analysis in Social Networks*. United States: Morgan Kaufmann, 2017, ch. 10, pp. 157-170.

[179] D. L. Hansen, B. Shneiderman, M. A. Smith and I. Himelboim, “Twitter: Information flows, influencers, and organic communities,” in *Analyzing Social Media Networks with NodeXL*, 2nd ed. United States: Morgan Kaufmann, 2020, ch. 11, pp. 161-178.

[180] A. Java, P. Kolari, T. Finin and T. Oates, “Modeling the spread of influence on the blogosphere,” in *Proc. of the WWW workshop*, 2006.

[181] J. Zhang, M. S. Ackerman and L. Adamic, “Expertise networks in online communities: structure and algorithms,” in *Proc. of the WWW*, 2007, pp. 221–230.

[182] L.B. Booker, The effects of observation errors on the attack vulnerability of complex networks. Tech. rep., MITRE CORP MCLEAN VA, 2012.

[183] A. Fronzetti Colladon and P.A. Gloor, “Measuring the impact of spammers on e-mail and twitter networks,” *International Journal of Information Management*, 2018.

[184] Y. Wang, *et al.*, “Reproducibility and robustness of graph measures of the associative semantic network,” *PloS one*, vol. 9, no. 12, 2014.

[185] G. Mao and N. Zhang, “Fast approximation of average shortest path length of directed BA networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 466, pp. 243–248, 2017.

[186] R. Cohen and S. Havlin. *Complex networks: structure, robustness and function*. Cambridge University Press, 2010.

[187] V. Chang, “A cybernetics social cloud,” *Journal of Systems and Software*, vol. 124, pp. 195–211, 2017.

[188] S. Iyer, T. Killingback, B. Sundaram and Z. Wang, “Attack robustness and centrality of complex networks,” *PloS one*, vol. 8, no. 4, 2013.

[189] A. Clauset, C. R. Shalizi and M. E. J. Newman, “Power-law distributions in empirical data,” *SIAM Rev.*, vol. 51, no. 4, pp. 661-703, 2009.

[190] M. Zabihimayvan and D. Doran, “A First Look at References from the Dark to Surface Web World,” *arXiv preprint*, 2019.