SENTIMENT & PATTERN ANALYSIS FOR IDENTIFYING NATURE OF THE CONTENT HOSTED IN THE DARK WEB

Ch A S Murty

Information Security Services, C-DAC
Hyderabad, Telangana 501510 India
chasmurty@cdac.in
http://www.cdac.in

Parag H Rughani

National Forensic Sciences University,
Gandhinagar, Gujarat 382007, India
parag.rughani@nfsu.ac.in
http://www.nfsu.ac.in

Abstract

Analysis of web content and sentiment nowadays are hot topics among researchers. The dark web is that integral part of WWW, which provides freedom of content hosting. The dark web is majorly used to do illegal activities; although accessing the dark web is legal in most countries, its usage can arouse suspicion with the law. The dark web's hosted websites provide various services in multiple Categories like Adult, Counterfeits, illegal markets, Drugs, weapons are prevalent. Sentiment's analysis attempts to identify emotions and opinions based on the transactions and text communications. Pattern Detection is a process of recognizing patterns in a dataset using a machine-learning algorithm. This study provides inputs to a framework for automating Dark Web Scraping and analysis of its hosted content. During the analysis, we also find some exciting patterns in hosted content of the dark web.

***Keywords*: Information Security, Dark Web, Data Mining, Machine Learning and Artificial Intelligence, Sentimental Analysis, Cyber Security, Deep Web, Pattern Analysis, Data Analysis, Topic Modling**

1. Introduction

Sentiment analysis aims to identify subjective information components from textual data and determine their sentiment polarity. The dark web allows users to participate in various transactions such as selling and purchasing from Armor Tank to Drugs to the Pirated Software. Content on the unindexed Internet is referred to as the Dark Web. This unindexed portion of WWW is intentionally hidden and inaccessible by standard Browsers referred to as Dark Web.

Dark Web is a place where the service providers have freedom of content hosting. According to the TOR Metrics project, more than 200K registered onion-based addresses as of May'2020, and on average, 2 million users use TOR daily [1]. The Internet Protocol (IP) address is required to access the Internet, provided by the Internet Service Provider (ISP). This IP address is essential because it provides a unique identity over the Internet, and at the same time, it helps find the geolocation of any device connected to the Internet. Onion Routers Provides anonymity to the dark web. These Routers encrypt and bounce communication through a network of relays run by volunteers around the world. The tor circuit mainly has three types of nodes:

- Entry node: The very first node, which receives the Incoming traffic.
- Intermediate node: This node passes the data from one node to another line by line.
- Exit node: The last node delivers the traffic to the open Internet.

Nowadays, Law Enforcement Agencies (LEA) show their interest in automatically identifying the TOR service domains and their activities. Sentimental analysis is a strategy to explore whether a gathered content is positive, negative, or neutral. Essentially, it involves examining the emotions related to a piece of writing for any topic. Sentiment analysis checks individuals' opinions, tastes, views, and interests by seeing diverse perspectives, such as celebrities, politicians, foods, places, or other topics [2].

Pattern Discovery and Topic Modelling is a Process of finding Interesting Patterns from the text using Machine Learning Algorithms. Topic Modelling (I) is an unsupervised machine learning technique that detects words and patterns between then clusters the words groups with similar characteristics. Both of the above methods can find and classify an activity, and that is what we implemented in the research for analyzing various activities happening on the dark web.

Web mining research incorporates approaches from different fields of study such as database management, information retrieval, artificial intelligence, and natural language processing [3]. Web mining [4] is a subset of data mining [5], the process of understanding information from hosted web page content that may consist of text, image, audio, or video data on the web. This study mined the hosted text data of dark web websites providing various services like adult con-tent hosting, Drugs, Counterfeits documents, illegal guns, and arms.

Web mining and analysis like content analysis and structural analysis can help discover and avert terrorist threats all across the world [6]. Understandings of emotions using Sentiment Analysis and text analysis are some tools that are helpful in the examination of trending topics, exciting pat-terns of the dark web.

This paper performed various text analysis algorithms such as sentiment analysis, topic modeling, association finding [7] on hosted text data of various dark web websites. We use Latent Dirichlet Allocation (LDA) [8] to examine a vast, diverse text corpus from the hosted website of different categories of the dark web. We can analyze different topics among the categories of websites, and we can identify some exciting patterns among the top frequent terms in each corpus of text data of different categories of the website.

A brief description of this research paper comes along with an abstract and an introduction to data and web mining, dark Web, and Tor Network. The 2nd section discusses an overview of the literature's literary works from earlier days to recent years and system architecture in section 3.0. The research Methodology is briefly described in section 3.0 and section 4.0, respectively. Section 5 describes the data and results. Section 5 concludes the paper.

2. Motivational and Related Work

Nowadays, the research community has raised its interest in recognition of the dark web and its activities. Natural Language Processing, Machine Learning is one of the emerging Fields. Sentiment analysis is the process that automates the mining of opinions, views, and emotions from the text using the database source through NLP. Sentiment Analysis involves classifying opin-ions/emotions like positive and negative, whereas we also used an approach to find the various patterns in the website's similar category. We use Topic modeling to find the top topic for the analysis of the activity.

In a Research "Temporal Analysis of Radical Dark Web Forum Users", researchers are able a novel way of analyzing the Dark Web forums through temporal analyses of forums and users and sentiment analysis to discover trends and patterns of sentiment scores over time-correlated to real-world terrorist events [9].

In a Research "BiSAL a bilingual sentiment analysis lexicon to analyze Dark Web forums for cybersecurity", researchers are ably presented a bilingual lexical resource (BiSAL) for sentiment analysis over English and Arabic texts related to cyber threats, radicalism, and conflicts. The BiSAL consists of two different lexical resources, namely SentiLEN and SentiLAR. SentiLEN contains a list of 279 sentiments representing English words related to cyber threats, radicalism, and conflicts, along with their morphological variants and sentiment polarity [10].

In a Research "Scalable Sentiment Classification across Multiple Dark Web Forums", researchers can examine several approaches to sentiment classification in the DWFP and opportunities to transfer classifiers and text features across multiple forums to improve scalability and performance [11].

In a Research "Sentiment and Affect Analysis of Dark Web Forums: Measuring radicalization on the Internet, "researchers can use an automated approach to the sentiment and affect the analysis of radical Jihadist web forums. The approach utilized a rich feature set to represent forum communications and machine learning techniques to identify and measure the sentiment polarity and intensities of four effects in forum postings [12].

Since Deep Web is a relatively huge place, then surface web. Researcher around the globe explored this place in every aspect but since we do not know where it ends, many researches are still going on. In this study, we majorly focused on the text mining and analysis of dark web sites.

In DW analysis, researchers have worked a lot on web classification algorithms. Researchers around the globe use various techniques to classify the hosted content on both types of data, such as structural data and content data. Existing Studies [13, 14, 15, and 16] use various methodology to classify the dark web content.

M. W. A. Nabki [17] et al. (2017) discussed various supervised algorithms, especially on the Logistic Regression-based classifier for multiple activities of the TOR network and their illegality based on the dark web content. With 10-Fold cross-validation, an accuracy of 96.6% and 93.7% of F-Score values achieved with the Darknet Usage Text Addresses (DUTA) Dataset. In research of Classification of Illegal Activities on Dark Web, Al Nabki proposed a classification method that uses 'Federal Code of United States of America' as training data to their model, which gave them the accuracy of 0.935.

Andrew J. et al. [18] analyse the Dark Web forums, users, and sentiment analysis to discover trends and patterns of sentiment scores over time-correlated to real-world terrorist events Xuan Zhang. et al. [19] presented a framework for analyzing the crime and criminals happening on the Dark Web and tracing the criminals.
Using text Analyse, various techniques like LDA [8], which is one of the best ways to analyse topics among the text corpus. Few studies are there which shown effective results using LDA on dark web text [20, 21]. Researchers use LDA in numerous ways like L'huillier et al. used LDA to extract and analyse the top members of Dark Web forums. Rios et al. used network filtering like techniques, and they discovered sub-communities of interest whose main topic could be a possible homeland security threat.

Using text Analyse, various techniques like LDA [8], which is one of the best ways to analyse topics among the text corpus. Few studies are there which shown effective results using LDA on dark web text [20, 21]. Researchers use LDA in numerous ways like L'huillier et al. used LDA to extract and analyse the top members of Dark Web forums. Rios et al. used network filtering like techniques, and they discovered sub-communities of interest whose main topic could be a possible homeland security threat.

3. Systems Architecture

In the first module, we enabled data mining techniques such as scraping, extraction of keywords from hosted pages of the dark web. Two approaches are integrated as one is an automated col-lection of text-based keywords directly from web pages themselves and, another is dumping the website by using tools and extracting keywords. Scraping the hosted content in the dark Web, extracting text from the content, and performing text cleaning in which we remove all numbers, punctuation, stop words, and special characters. This process will reduce all irrelevant text, which ultimately reduces the complexity of text processing and analysis.

In the second module, we used techniques like feature selection, feature analysis to categorize the keywords that correspond to a particular category based on an opinion of experts and by analyzing their occurrence and density percentage to form a dataset. The dataset made such that the independent (classes) and dependent features (Legal/illegal) for the classification problems for the dark web. Based on the percentage of keywords matched for legal/illegal against to categories of keywords assigned.

In Module 3, the corpus of each category is given to Sentiment Analysis, in which each word is checked for the sentiment and will the added into a category, i.e. ['Anger',' Anticipation',' Dis-gust',' Fear',' Joy',' Negative',' Positive',' Sadness',' Surprise',' Trust']. Then after the same corpus is processed for the Finding Patterns and TopicModelling.

4. Dataset Description

4.1. *Data Collection*

Our System Architecture consists of all required elements for collecting the data for sentiment and pattern analysis of the dark web. During the development phase, there are several challenges: the continuous availability of hosted URLs, security implementations, anti-scraping methods, and scraping multiple onion websites simultaneously. We solved those problems by integrating and developing a customized script in our first module.

The text data scraped between the months of January-May 2021. Initially, we have an extensive data set and many duplicities after pre-processing the data we selected from 2000 websites. Da-ta has Three Columns URL, CORPUS, and CATEGORY. URL is the address of the onion website from where we will extract hosted text data. Corpus is the text what we scrape from dark web websites. The category is in what text data is, for example, Adult, Drugs, Weapons, and other categories.

The data collection process is implemented through various tools and technologies, including TOR Browser, Python Programming Language, various libraries, and customized scripts. The libraries such as Requests, Beautiful Soup, Proxy Broker (for Socks5 Proxy).

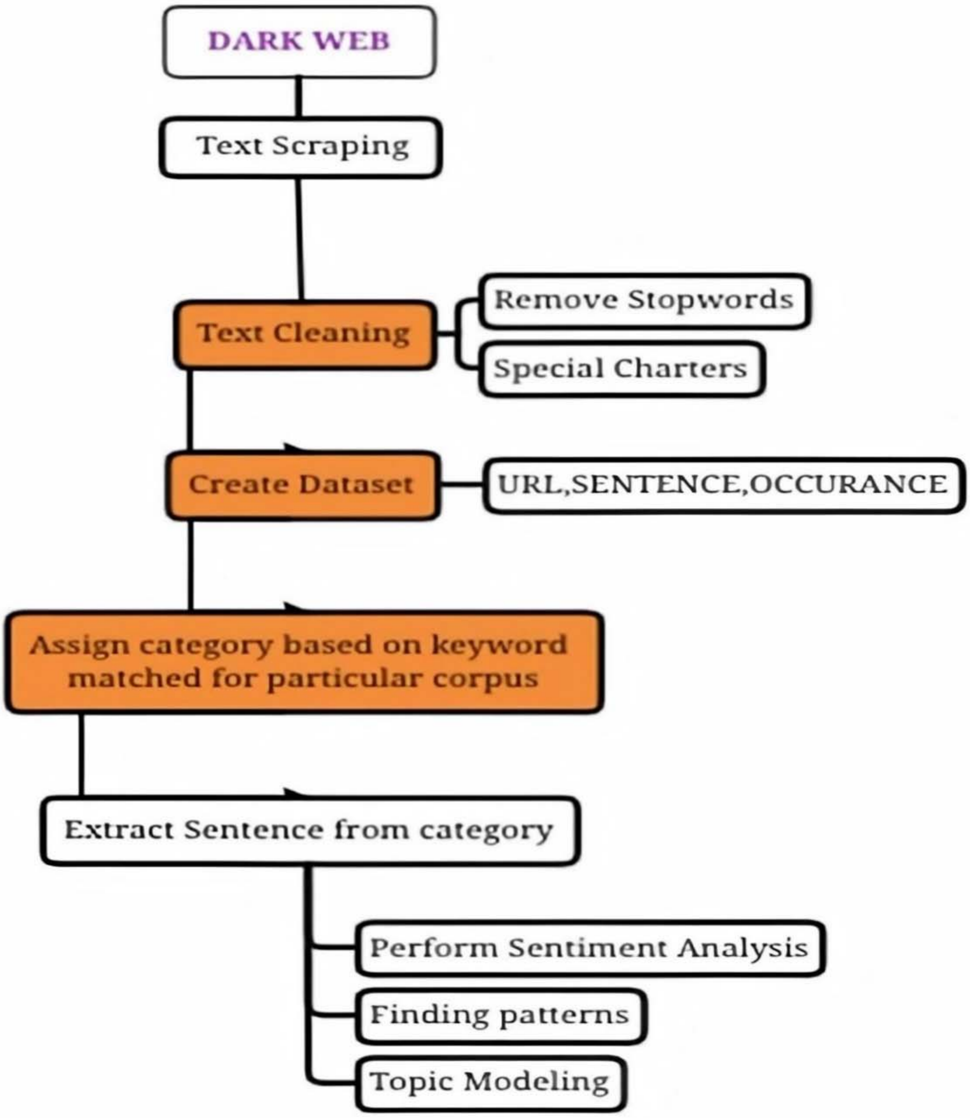


Figure 1: Systems Architecture Which Contains Modules From Scraping of Text Data to Sentiment Analysis.

4.2. *Data Description*

Data has Three Columns URL, CORPUS, and CATEGORY. URL is the address of the onion website from where we will extract hosted text data. Corpus is the text what we scrape from dark web websites. The category is in what text data is, for example, Adult, Drugs, Weapons, and other categories.

URL	CORPUS	CATEGORY
http://222t2volor5km5ao.onion/	automated PayPal and credit … email pro	Cryptocurrency
. . .	. . .	. . .
http://65e7rpzocon2sarx5jf4rgguljps5ckit3z3n5vt6u6h43zatirah6yd.onion/	seeds cannabis drugs shopper …	Drugs

Table 1: Data Description

The Data Set Contains around 1932 Entries of URLs and their hosted text data, which fall into Categories Like ["ADULT"." CRYPTO"," COUNTERFEIT"," DRUG"," MARKET"," SERVICE"," WEAPON"].

5. Implementation

The Sentiment Analysis and Pattern Matching comprised seven main stages and defined as URLs Mining, Data Scraping, Data Cleaning, Feature Extraction, Sentiment Analysis, Pattern Discovery, and finally Topic Modelling.

5.1. *URL Mining:*

Dark Web is an unindexed web, and the format of the Dark web URL is very complex, which uses a stronger crypto algorithm V3 URLs with 56 bytes long. Remembering the URLs is very difficult and needs to depend on dark web search engines. These properties make the mining process difficult, so we created a customized script that mines URLs by visiting different search engines and producing remarkable results.

5.2. *Data Scraping:*

Web Scraping, aka harvesting, means data extraction from the websites. It is an automated process implemented using a bot or crawler. This process includes targeted websites, Collecting URLs, Request URL is to get the HTML of the webpage, can save data into CSV, JSON format.

5.3. *Data Cleaning:*

During the data cleaning process, the data is allocated for cleaning before analyzing it, such as removing duplicity and cleaning text. While dealing with text, it is required to clean all the irrelevant text, which does not add any weights to the result set. In addition, Text Cleaning Reduces the Complexity.

5.4. *Feature Extraction:*

The pre-processed dataset has many distinctive properties (II) as in the feature extraction method; we extract the aspects from the processed dataset. Later this aspect is used to compute the positive and negative polarity in a sentence, which helps determine the individuals' forex: Words and their frequencies: unigrams, bigrams, and n-gram models with their frequency count as features. There has been more research on using word presence rather than frequencies to describe this feature better.

5.5. *Sentiment Analysis:*

Sentiment analysis is an approach for determining whether data is positive, negative, or neutral. Sentiment analysis is often performed on collected textual data as it is studying the emotions associated with writing on any subject. Sentiment analysis examines people's opinions, tastes, points of view, and interests from various viewpoints.

5.6. *Pattern Discovery:*

Pattern Discovery aims to identify the text data pattern by achieving word association. The words are chosen based on the top frequent terms of a category.

5.7. *Frequent Term Extraction:*

Frequent Term Extraction is the process of Data Exploration; frequent terms in the same paragraph are found using the document term matrix, ref [3.1] of Methodology to under-stand the process. Reference with table [2, 3, 4, 5, 6, 7, and 8] were the frequent terms of each category of the website mentioned. Scraped data is in the natural language we need to convert it into the précised format for the analysis. The method used in the pre-processing is known as document term matrix. The Document Term Matrix (DTM) is a mathematical matrix that shows the frequency of words in a set of documents. The column in the DTM relates to a single word inside that document.

Each row in DTM relates to a single document in the collection. In natural language processing,
Such matrices are commonly utilized, as there are two advantages of converting a corpus into DTM. i.e., first, it presents each document by the count of unique terms; it also helps in further text analysis. The Basic Notation for counting the Term Frequency is:

$$Tf(t) = (Number\ of\ times\ term\ apppears\ in\ a\ document) \\ /\ (total\ no.of\ terms\ in\ the\ document) \qquad (1)$$

Here, T is refer to a particular Term from the corpus F stands for frequency; TF (t) is the formula for counting the number of occurrences of a term.

5.8. *Word Associations Extraction:*

Once we understand some of the most common phrases in our corpus, we can go further into it by looking at different relationships between them. While this may not be the ide-al technique to investigate the substance of each text or the corpus as a whole, it may give some helpful information for future research. Term Association based on the standard correlation formula.

$$r = \frac{Covariance(x,y)}{StandardDeviation(x)xStandardDeviation(y)} \qquad (2)$$

$$mk(t1, t2, \ldots.., tk) = SentenceFrequency(t1, t2, \ldots., tk)X\prod Tf(tj) \qquad (3)$$

Where j = 1, 2,……..k.

These words are related to other words at a set threshold value because they are the most commonly recurring terms in various tales or documents. As discussed in the data description, we have the top frequent terms from the corpus of the particular category of the dark web website. Using Word Association mining, we find the terms associated with the frequent terms with the higher value of association. Below listed Tables are the result set of each category.

It is an Approach to organize, understand and summaries an extensive collection of textual information. Extracting the veiled concepts, protruding features, and latent variables from data that depend on the application context, we use Latent Dirichlet Allocation (LDA) [11] to analyze a sizeable heterogeneous text corpus from the dataset. A topic model is a probabilistic model that connects texts and words using variables that reflect the suitable subjects identified in the text. In this perspective, a document may be considered a collection of themes represented by probability distributions used to produce the words in a document based on these themes. LDA [8] generative model that says documents have multiple subjects. The topic is a distribution to a fixed vocabulary. All homogeneous typesetting documents contribute to similar topics, but each document shows these topics in a different relationship.

5.9. *Topic Modelling:*

It is an Approach to organize, understand and summaries an extensive collection of textual information. Extracting the veiled concepts, protruding features, and latent variables from data that depend on the application context, we use Latent Dirichlet Allocation (LDA) [11] to analyze a sizeable heterogeneous text corpus from the dataset. A topic model is a probabilistic model that connects texts and words using variables that reflect the suitable subjects identified in the text. In this perspective, a document may be considered a collection of themes represented by probability distributions used to produce the words in a document based on these themes. LDA [8] generative model that says documents have multiple subjects. The topic is a distribution to a fixed vocabulary. All homogeneous typesetting documents contribute to similar topics, but each document shows these topics in a different relationship.

6. Result Analysis

6.1. *Frequent Term Extraction*

Frequent term Extraction is the process of Data Exploration; frequent terms in the same para-graph found using the document term matrix, ref [3.1] of Methodology to understand the process. Ref. Table [2, 3, 4, 5, 6, 7, and 8] was the frequent terms of each category of the website mentioned.

6.2. *Word Associations Extraction*

Frequent Term Extraction VI.A, gives us the top five frequent in each mentioned categories of the dark web website. Using Word Association Mining, we find the terms associated with the frequent terms with the higher value of association. Below the Listed Tables are the result set of each category.

Frequent Terms	Term Associated with the Frequent Term				
	Association Value				
Child	hardcore	assorted	collections	http	groups
	0.81	0.79	0.79	0.78	0.78
Free	xxx	expect	ass	russian	tits
	0.97	0.95	0.94	0.94	0.94
Girls	young	load	teen	sex	clips
	0.83	0.71	0.7	0.7	0.7
Porn	porno	smotret	gay	russian	collapsed
	0.99	0.97	0.96	0.96	0.96
Video	porno	Online	Sex	smotret	tubes
	0.91	0.9	0.9	0.9	0.9

Table 2. Association Result of frequent terms in Adult Category.

The Sentiment impact of the Adult Category is Negative. In addition, many words are from Anger, Disgust, and Fear Category as shown in VI.D. Association and Topics also show evidence of child porn and rape.

Frequent Terms	Term Associated with the Frequent Term Association Value				
Bitcoin	addr	beginers	bitcoindouble	bitcoingen	bitcoingenerators
	0.94	0.94	0.94	0.94	0.94
eur	aaron	abdelilah	Afeez	agnes	aguilar
	1	1	1	1	1
btc	hack	faucet	view	free	maker
	0.9	0.8	0.8	0.87	0.76
ethereum	ace	adc	afcee	alinir	arbitrage
	0.97	0.97	0.97	0.97	0.97
Invest	buy	hacks	top	market	adder
	0.94	0.94	0.92	0.92	0.92

Table 3. Association Result of frequent terms in Crypto Category.

The Sentiment impact of the Crypto Category is Positive. In addition, many words are from the Trust Category as shown in VI.D. In conclusion, we can say that Crypto services websites use words from trust sentiments to add trust to their users. Association and Topics also show the evidence of Bitcoin Hacking, Doubling, and Faucet.

Frequent Terms	Term Associated with the Frequent Term Association Value				
buy	online	save	Canada	read	real
	0.94	0.94	0.94	0.94	0.94
card	ATM	information	country	addres	first
	0.89	0.88	0.81	0.79	0.79
credit	ATM	trusted	information	thank	version
	0.94	0.88	0.8	0.75	0.75
cvv	aus	min	master	Account number	ages
	0.98	0.98	0.97	0.97	0.97
dumps	track	good	shop	pin	sell
	1	0.98	0.97	0.96	0.96

Table 4. Association Result of frequent terms in Counterfeit Category.

The Sentiment impact of the Counterfeit Category is Positive. In addition, many words are from the Trust Category as shown in VI .D. In conclusion, we can say that Counterfeit services websites use words from trust sentiments to add feelings of trust to their users. Association and Topics also show the evidence of Fake Passports, Fake Bank Notes, Card shops, counterfeit iPhones, and sell purchase of Cards Details (Like CVV, credit card, ATM services.

Frequent Terms	Term Associated with the Frequent Term				
	Association Value				
cocaine	peruvian	crack	trusted	via	ships
	0.78	0.78	0.76	0.75	0.71
pure	fish	flake	scale	address	across
	0.69	0.69	0.67	0.66	0.64
cannabis	cbd	edibles	seeds	recommended	home
	0.73	0.73	0.71	0.65	0.6
drugs	agaric	amani	antipsychotics	balls	beans
	0.83	0.83	0.83	0.83	0.83
marijuana	thc	hash	bad	bad	acopulco
	0.95	0.94	0.93	0.91	0.91

Table 5. Association Result of frequent terms in Drug Category.

The Sentiment impact of the Drug Category is Positive. In addition, many words are from the Trust Category as shown in VI.D. In conclusion, we can say that Drug services websites use words from trust sentiments to add trust to their users. The transactions and statements showed in the drug category proved the purchase and selling evidence of cocaine, cannabis, and marijuana. We also get some of the pharmaceutical-related chemical drugs in common topics. Association Results also show the imports of Drugs from Peru.

Frequent Terms	Term Associated with the Frequent Term				
	Association Value				
cards	visa	cloned	debit	union	western
	0.87	0.85	0.82	0.81	0.81
buy	cash	online	steal	make	Caught
	0.58	0.55	0.55	0.53	0.53
onion	http	wiki	cannabisuk	tor	Hidden
	0.81	0.5	0.49	0.48	0.46
money	cash	make	making	steal	Rich
	0.77	0.76	0.73	0.71	0.68
paypal	accounts	mastercard	debit	western	Hacked
	0.87	0.86	0.83	0.81	0.78

Table 6. Association Result of frequent terms in Market Category.

The Sentiment impact of the Market Category is Positive. In addition, many words are from the Trust Category as shown in VI.D. In conclusion, we can say that Market services websites use words from trust sentiments to add feelings of trust to their users. Association and Topics also show the evidence of card cloning, cash stealing, onion related PayPal/western union services, and sell purchase of miscellaneous items.

Frequent Terms	Term Associated with the Frequent Term				
	Association Value				
onion	cockmail	gum	torch	clockwise	mailman
	0.93	0.93	0.88	0.85	0.84
tor	native	install	styleguide	websit	offline
	0.58	0.55	0.55	0.54	0.54
http	carts	wiki	darktor	deepboard	deeplink
	0.98	0.98	0.98	0.98	0.98
log	daily	iso	armel	armhf	busteramd
	1	1	1	1	1
les	pour	une	des	dans	faire
	0.98	0.97	0.94	0.93	0.93

Table 7. Association Result of frequent terms in Service Category.

The Sentiment impact of the Service Category is Positive. In addition, many words are from the Trust Category as shown in VI.D. In conclusion, drawn, service services websites use words from trust sentiments to add feelings of trust into their users. Association and Topics also show the evidence of onion tor network-related services, hacking Service like (WhatsApp Hack).

Frequent Terms	Term Associated with the Frequent Term				
	Association Value				
gun	firearm	luty	individual	anti	never
	0.92	0.91	0.91	0.9	0.9
glock	package	abbasy	accessories	additio	advantage
	0.99	0.99	0.99	0.99	0.99
ammo	link	abbasy	accessories	addition	advantage
	0.89	0.82	0.82	0.82	0.82
magazine	gen	auto	full	point	version
	0.91	0.9	0.89	0.79	0.79
firearms	design	defence	known	great	com
	0.95	0.94	0.89	0.87	0.86

Table 8. Association Result of frequent terms in Weapon Category.

The Sentiment impact of the Weapons Category is Positive. In addition, many words are from the Trust Category as shown in VI.D. In conclusion, we can say that Weapons services websites use words from trust sentiments to add feelings of trust to their users. Fear Category also has a good percentage of sentiments score, which is evident as well. Association and Topics also show the Selling Purchasing evidence of a gun, Glock, ammo, magazine, and firearms.

6.3. *Topic Analysis*

In this part, we look at the most important terms used by Latent Dirichlet Allocation (LDA) to learn about different topics. In our dataset, we cover a wide range of subjects. Below the Tables is the Output Set of Topic Analysis.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
der und die hannover video	video april photo posts post	porn porno video free sex	anonymous jpg reply src read	porn porno video sex free

Table 9. Topic Analysis Adult Category.

Adult Category Analysis: Posting Adult Content on the website, Free Adult Content, Anonymity, and Free Content. There are also sections for Unidentified, as the term shows no relevance to the category. In conclusion, most of the topics inclines toward free Adult content.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
btc bitcoin add cart explorer	les can ethereum est des	bitcoin double bitcoins btc ethereum	usd eur united zip states	usd btc casino money games

Table 10. Topic Analysis Crypto Category.

Crypto Category Analysis: Crypto Category show that most of the Crypto Service is used in the transaction such as sale and purchase of bitcoin cryptocurrency and other related Tokens one of the topics is doubling of bitcoin Scam and gambling. In conclusion, most of the topics incline toward bitcoin-related services and scams.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
fake passport buy banknotes counterfeit	dumps pin sell card shop	eur grams cards cloned btc	cvv iphone fullz apple per	name card number credit birth

Table 11. Topic Analysis Counterfeit Category.

Counterfeit Category Analysis: The topics in Counterfeit Category show that most of the services are related to counterfeiting currency and other personal financial details such as Credit Card details, sale of counterfeit electronic goods such as iPhones. In conclusion, most of the topics incline toward transaction fiat currency and dumped financial details.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
heroin cocaine online marijuana buy	view quick sale pcs wishlist	cannabis cocaine seeds drugs pure	kamagra cocaine gbp drugs gram	cocaine marijuana kush hash shrooms

Table 12. Topic Analysis Drugs Category.

Drug Category Analysis: The topics in Drug Category show that most of the services are related to drug selling. In a conclusion, most of the topics incline toward various types of drugs and e-commerce services of drugs.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cards card buy transfer payment	onion http bitcoin tor credit	buy money order btc sale	reply service bank contact hack	web dark darknet april scam

Table 13. Topic Analysis Market Category.

Market Category Analysis: The topics in Market Category show that most of the services are related to carding, crypto, hacking, tor, and darknet. In conclusion, each topic concludes to a different kind of category and related services.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
privacy http tor forum anonymous	les des pour une sur	tor die und der use	log cronjob daily que hacker	whatsapp tor hack debian build

Table 14. Topic Analysis Service Category.

Service Category Analysis: The topics in Service Category show that most of the services are related to hacking and tor. Topic 2 and Topic 3 contain words from the French language. In conclusion, each topic concludes various hacking services.

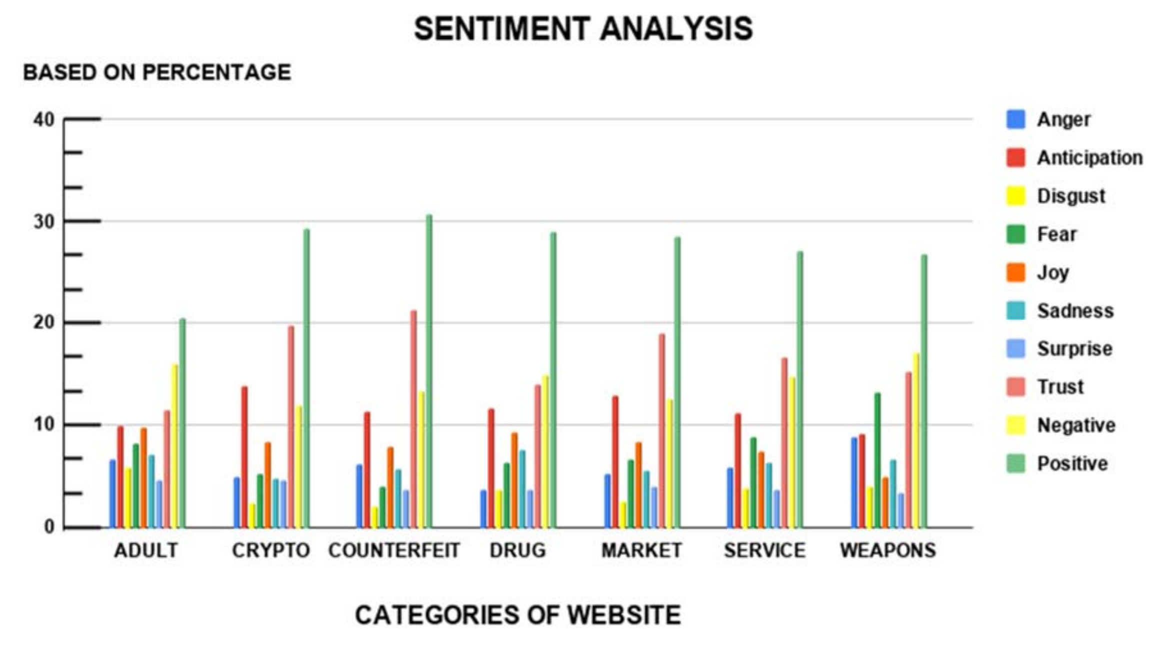
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
iphone gold silver gray space	printable glock frame receiver data	read specs rounds ammo magazine	gun firearms homemade firearm control	amount btc photo barrel weight

Table 15. Topic Analysis Weapons Category.

Weapons Category Analysis The topics in Weapons Category show that most of the services are related to the sale/purchase of weapons, homemade weapons, firearms attachments. In conclusion, each topic concludes transactions of weapons.

6.4. *Sentiment Analysis*

Sentiment Analysis performed on the text of each category of the website so that it gets easy to analyse. We have created a subset from the dataset of each category. Each subset further processed for the text cleaning and, ultimately, the sentiment analysis. Further, we processed the result set of each category into a percentage and gave a better visualization of the Sentiments.



6.5. *Case Study*

During this study, the authors noticed some intriguing patterns in the text data, and we investigated further on the patterns we were able to analyse in some of the cases studied.

6.5.1. Drugs Import and Exports

Drugs are one of the commonly used services over the dark web. According to the Association Table of Drug Category, we observed an interesting pattern of words among the keyword Cocaine that is whenever the cocaine keyword has used the keywords, i.e., Peruvian, ships are repetitively used which implies that the drugs imported to other countries via or from the country Peru. "Under Peruvian law, drug use is not punishable by imprisonment, which is appropriate, because problematic drug users deserve treatment, not imprisonment" [23]; possession of up to 2 grams of cocaine or up to 5 grams of coca paste is legal for personal use in Peru per Article 299 of the Peruvian Penal Code. [24] This statement also supports this case study, as the law of drugs in Peru is not that harsh. In conclusion, we can say that the drugs imported to Countries like Mexico, the USA, etc. There might be a chance also these website administrators are from Peru.

6.5.2. Crypto Gambling and Scam

Dark Web where anonymity is guaranteed to make it more secure regarding the virtual transaction currencies used, such as bitcoin. These Crypto Currencies have been used for illegal purposes like money laundering, weapons sale and purchases, Drug Mafia, terrorist activity, hit for hire, and sex trafficking. We observed in the crypto category websites that provide core services of crypto like wallet services transferring of bitcoin Gambling games, doubling coin. In gambling/Casino games on the dark web, they all are frauds. They have implemented the zero-sum game algorithm sometimes to win but lose most of the time. In addition, we observed that there is no such thing that bitcoin doubled by transferring to someone else so far in the crypto field. Even staking of bitcoin cannot ensure doubling. The moral of the story is that there is no such thing to double/triple a Bitcoin. If we send our Bitcoin, we will never get the return from the double Bitcoin sites. Be aware of these kinds of frauds on the dark web.

7. Conclusion

Dark Web and Digital Investigation are some of the top emerging fields among the researcher. This study inclined towards understanding the overhaul sentiments and digging out some exclusive patterns result in each category mentioned earlier of the dark web. This study aims to understand the environment of dark web websites and perform sentiment analysis of text under various categories. Along with sentiment, we also looked for some interesting patterns in the text data. The dark web's complex structure necessitates advanced techniques for accessing and navigating the material and data stored in dark web databases. Combining clustering techniques and pattern mining may give a more efficient and refined data access method on the deep web. This study aims to understand the environment of dark web websites and perform sentiment analysis of text under various categories. Along with sentiment, we also looked for some interesting patterns in the text data.

References

[1] https://metrics.torproject.org/
[2] Chowdhary, Ankur. Autonomous Security Analysis and Pretesting (ASAP) *DEFCON Red Team Village***2020**
[3] Sawant, Poonam. (2012). A Review Of Web Mining Research. International Journal of Management, IT and Engineering. 2. 153-166.
[4] Xu, Zhengqiao & Zhao, Dewei. (2012). Research on mobile learning system based on Web mining. ICICIP 2012 - 2012 3rd International Conference on Intelligent Control and Information Processing. 565-568. 10.1109/ICICIP.2012.6391484.
[5] Kumar, G. & Gosul, Manohar. (2011). Web Mining Research and Future Directions. Communications in Computer and Information Science. 196. 10.1007/978-3-642-22540-6_47.
[6] Selamat, Ali & Alghamdi, Hanan. (2012). Topic Detections in Arabic Dark Websites Using Improved Vector Space Model. 10.1109/DMO.2012.6329790.
[7] Mäntylä, Mika & Graziotin, Daniel & Kuutila, Miikka. (2016). The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers. Computer Science Review. 27. 10.1016/j.cosrev.2017.10.002.
[8] Blei, David & Ng, Andrew & Jordan, Michael & Lafferty, John. (2003). Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03 Latent Dirichlet Allocation.
[9] Park, Andrew & Beck, Brian & Fletche, Darrick & Lam, Patrick & Tsang, Herbert. (2016). Temporal Analysis of Radical Dark Web Forum Users. 10.1109/ASONAM.2016.7752341.
[10] Al-Rowaily, Khalid & Abulaish, Muhammad & Smieee, & Haldar, Nur & Alrubaian, Majed. (2015). BiSAL-A Bilingual Sentiment Analysis Lexicon to Analyze Dark Web Forums for Cyber Security.
[11] Zimbra, David & Chen, Hsiu-chin. (2012). Scalable Sentiment Classification Across Multiple Dark web Forums. Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics (ISI 2012). 78-83. 10.1109/ISI.2012.6284095.
[12] Chen, Hsiu-chin. (2008). Sentiment and affect analysis of Dark Web forums: Measuring radicalization on the internet. 104 - 109. 10.1109/ISI.2008.4565038.
[13] Zhou, Yilu & Qin, Jialun & Lai, Guanpi & Reid, Edna & Chen, Hsiu-chin. (2006). Exploring the Dark Side of the Web: Collection and Analysis of U.S. Extremist Online Forums. 621-626. 10.1007/11760146_67.

[14] Chen, Hsiu-chin & Chung, Wingyan & Qin, Jialun & Reid, Edna & Sageman, Marc & Weimann, Gabriel. (2008). Uncovering the Dark Web: A case study of Jjihad on the Web. JASIST. 59. 1347-1359. 10.1002/asi.20838.
[15] Qin, Zengchang & Lawry, Jonathan. (2008). Fuzzy Label Semantics for Data Mining. Studies in Fuzziness and Soft Computing. 218. 10.1007/978-3-540-73185-6_11.
[16] Qin, Jialun & Zhou, Yilu & Chen, Hsiu-chin. (2011). A multi-region empirical study on the internet presence of global extremist organizations. Information Systems Frontiers. 13. 75-88. 10.1007/s10796-010-9277-6.
[17] Al Nabki, Wesam & Fidalgo, Eduardo & Alegre, Enrique & Paz, Ivan. (2017). Classifying Illegal Activities on Tor Network Based on Web Textual Contents. 35-43. 10.18653/v1/E17-1004.
[18] Park, Andrew & Beck, Brian & Fletche, Darrick & Lam, Patrick & Tsang, Herbert. (2016). Temporal Analysis of Radical Dark Web Forum Users. 10.1109/ASONAM.2016.7752341.
[19] Zhang, Xuan & Chow, K.P. (2020). A Framework for Dark Web Threat Intelligence Analysis. 10.4018/978-1-7998-2466-4.ch017.
[20] L'Huillier, Gaston & Alvarez, Hector & Ríos, Sebastián & Aguilera, Felipe. (2011). Topic-Based Social Network Analysis for Virtual Communities of Interests in the Dark Web ABSTRACT. SIGKDD Explor. Newsl.. 12. 66-73. 10.1145/1964897.1964917.
[21] Ríos, Sebastián & Muñoz, Ricardo. (2012). Dark Web portal overlapping community detection based on topic models. 10.1145/2331791.2331793.
[22] Gregor Heinrich (2004) Parameter estimation for text analysis http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.216.695
[23] https://www.tni.org/es/node/16661
[24] http://druglawreform.info/en/country-information/latin-america/peru/item/207-peru

Authors

Ch A S Murty, obtained a Master of Science, University of Hyderabad and Master of Technology degree in JNT University, Hyderabad, India. He is currently pursuing Ph.D. in Computer Science from National Forensic Sciences University, Gandhinagar, India. His research areas of interest are Dark web, AI/ML Techniques, Auditing and Assessment environment vulnerability assessment and penetration testing of infrastructure, web in IT/ ICT

Dr Parag H Rughani obtained his PhD in computer science from Saurashtra University. He is currently working as an associate professor in digital forensics at the Institute of Forensic Science, Gujarat Forensic Sciences University, Gandhinagar. He has more than 14 years of teaching experience and has published more than 15 research papers in reputed international journals. His areas of expertise include machine learning, digital forensics, memory forensics, malware analysis and IoT security and forensics.