Stage I report
on

LABELLING HIDDEN SERVICES WITH IMAGE TAGGING

Submitted
in partial fulfillment of
the requirement of the degree of

M.Tech in Software Engineering

by

Akansha Sudhirkumar Singh
(202191015)

Under the Guidance of
Dr. S. G. Bhirud

Department of Computer Engineering and Information Technology
VEERMATA JIJABAI TECHNOLOGICAL INSTITUTE
(An Autonomous Institute Affiliated to Mumbai University)
(Central Technological Institute, Maharashtra State)
Matunga, MUMBAI - 400019
A.Y. 2021-2022

STATEMENT OF CANDIDATE

I state that work embodied in this Project entitled **"Labelling Hidden Services with Image Tagging"** form my own contribution of work under the guidance of **Dr. S. G. Bhirud** at the Department of Computer Engineering, Veermata Jijabai Technological Institute, Mumbai. The report reflects the work done during the period of candidature but may include related preliminary material provided that it has not contributed to an award of previous degree. No part of this work has been used by us for the requirement of another degree except where explicitly stated in the body of the text and the attached statement.

Akansha Sudhirkumar Singh
Roll No:- 202191015
Date:
Place: VJTI, Mumbai

CERTIFICATE

This is to certify that Akansha Sudhirkumar Singh, a student of M.Tech in Software Engineering, has completed the Stage 1 report entitled, “***Labelling Hidden Services with Image Tagging***” to our satisfaction.

Dr. S. G. Bhirud	Dr. M. R. Shirole
Project Supervisor	***Head, Department of CE and IT***

Place: VJTI, Mumbai
Date:

APPROVAL SHEET

The report "***Labelling Hidden Services with Image Tagging***" submitted by Akansha Sudhirkumar Singh[202191015], is found to be satisfactory and is approved for the Degree of M.Tech in Software Engineering.

Dr. S. G. Bhirud
Project Supervisor

Examiner

Examiner

Examiner

Place: VJTI, Mumbai
Date:

ACKNOWLEDGEMENT

I would like to thank all those people whose support and cooperation has been an invaluable asset during the course of this Project.

I would also like to thank our Guide **Dr. S. G. Bhirud** for guiding me throughout this project and giving it the present shape. With his encouraging words and reminders, he empowered me to improve in personal as well as academic life. It would have been impossible to complete the project without his support, valuable suggestions, criticism, encouragement and guidance.

I convey my gratitude also to **Dr. M. R. Shirole**, Head of Department for his motivation and providing various facilities, which helped us greatly in the whole process of this project.

I am also grateful to all other teaching and non-teaching staff members of the Computer Engineering and Information Technology Department for directly or indirectly helping us for the completion of project and the resources provided.

Akansha Sudhirkumar Singh

ABSTRACT

Dark Web forms a large portion of the internet and provides anonymity to its users. This anonymity provides privacy to users but is also taken advantage of by the cyber criminals or terrorists to fulfil their illicit motives. Various studies points out to how dark web is used for illegal activities. Any service which is hosted on the Dark Web is a hidden service. A hidden service is a combination of text and graphical data. Most of the techniques available focus on hidden service text data analysis. However graphical data forms a large portion of Dark Web which cannot be ignored. Most of the methods or techniques are directed towards analysis of Dark Web graphical data focused on a particular domain like Child Sex Abuse (CSA). This research makes use of a dataset which consists of a number of hidden services containing both text and graphical data. The image data present in this dataset will be labelled into unique classes or categories that can be formed using available dataset. The model will be trained using train data and validated using validation data. Model will be subjected to identify the category of any test image. Once identified, text data respective to that image will also be analyzed to confirm if the text and image data are correlated. This will help in identifying the type of hidden service effectively.

Keywords: Dark Web, Image Analysis, Hidden Service, Image Categorization

Contents

STATEMENT OF CANDIDATE i

CERTIFICATE ii

APPROVAL SHEET iii

ACKNOWLEDGEMENT iv

ABSTRACT v

Contents vi

List of Figures vii

Abbreviations viii

1. **INTRODUCTION** 1

1.1. Overview 1
1.2. Motivation 3
1.3. Problem Statement 3
1.4. Project Scope 3
1.5. Future Scope 3

2. **LITERATURE REVIEW** 4
2.1. Literature Survey 4
2.2. Comparative Research Table 12
2.3. Literature Gap 18

3. **PROPOSED APPROACH** 19
3.1. Dataset Overview 19
3.2. Overview of Proposed Approach 20
3.3. Data Cleaning 21

4. **CONCLUSION** 22

Bibliography 23

List of Figures

1.1 Identification of Objects in Images as Weapons 2
2.1 Classes in TOIC Dataset 4
2.2 Overview of Proposed Method for TOR Image Content Categorization 5
2.3 Overview of the Proposed SAKF Method 6
2.4 The main features of DUSI dataset 7
2.5 Overview of Perceptual Hashing Process 8
3.1 A Hidden Service in the Dataset 19
3.2 An Overview of Proposed Approach 20
3.3 Data Cleaning Approach 21

Abbreviations

TOR	The Onion Router
CSA	Child Sex Abuse
TOIC	TOR Image Categories
CREIC	Compass Radius Estimation for Image Classification
SIFT	Scale-invariant feature transform
YOLO	You Only Look Once
BOW	Bag of Words
BOVW	Bag of Visual Words
SAKF	Semantic Attention Keypoint Filtering
DUSI	Darknet Usage Service Images
HSV	Hue Saturation Value
NSFW	Not Safe For Work
SSD	Single Shot Detection
CNN	Convolutional Neural Networks
SAGE	Scientific Advisory Group for Emergencies
IEEE	Institute of Electrical and Electronics Engineers
DUTA	Darknet Usage Text Addresses
TF-IDF	Term Frequency — Inverse Document Frequency
JSON	JavaScript Object Notation
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets

Chapter 1

INTRODUCTION

1.1 Overview

1.1.1 What are Hidden Services?

A hidden service is a site you visit or a service you use that uses Tor technology to stay secure and, if the owner wishes, anonymous. “Hidden services” are also known as “onion services”.[2]

1.1.2 Why and Who uses it?

Hidden services are used to carry out activities that are otherwise illegal and unethical on the surface Web by cyber criminals or terrorists. These activities include distribution of child pornography, access and sale of illegal drugs, and the sale of weapons.

1.1.3 What is TOR?

One of the famous networks into the Dark Web is The Onion Router (TOR), and its content can be consulted through the TOR Browser or through the Surface Web thanks to projects like TOR2WEB.[1]

1.1.4 What is Image Tagging?

Image tagging is the process of labelling images with keywords to make them more searchable.[5]

1.1.5 What is the significance of Image Tagging?

A company benefits from image tagging because of how quickly they organize images and how accessible their images become to users[4]. When a user searches their website, they can find what they need based on simple keywords.The time saved on key wording is immense.[4]

1.1.6 Image Tagging in Dark Web

Identifying the object present in dark web images for any hidden service is crucial. This can help in understanding the kind of hidden service one is dealing with.

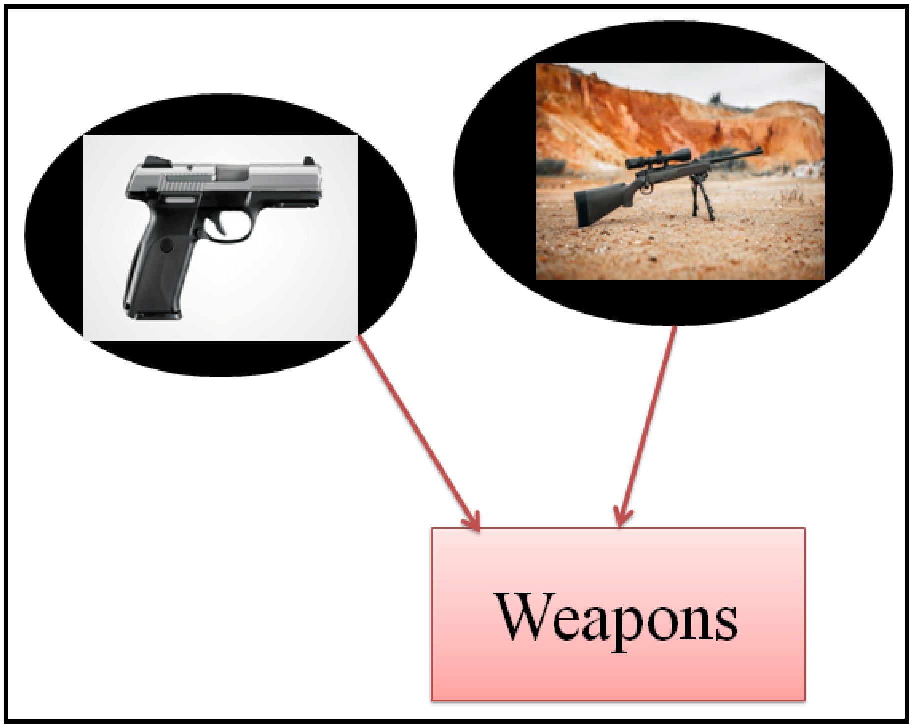


Figure 1.1: Identification of Objects in Images as Weapons

For example in Figure 1.1 shows how identifying the object in images as weapons helps one understand that this hidden service is dealing with weapons.

1.2 Motivation

Approximately 75% of dark marketplace listings include image data, indicating the importance of considering image content for investigative analysis[13]. However, visiting thousands of domains to look for visual information containing illegal acts manually requires a considerable amount of time and resources. Thus a system which can automatically label images in various hidden services is highly desirable. Existing researches focus on specific domain such as identifying Vendors or Child Sex Abuse Material. Even if some researches considers content having more than one class (example: weapons, drugs, counterfeit money, etc.), focus is on some specific classes leaving other classes out of consideration. Thus the motivation of this project is to contribute to the field of dark web image analysis and confirm the result of analysis with textual analysis of hidden services using the available data-set.

1.3 Problem Statement

For this project, dataset that is available is of hidden services containing around 48,700 files and 532 folders consisting of both text and image data. A model that can provide highest possible accuracy for existing dataset in identifying the objects in the dataset will be identified and implemented, it will help researchers in future who will work on dataset containing similar classes. The dataset that will be used in this research will be made publicly available for research community to work on it. Image labeling will be confirmed with the textual data of the hidden services.

1.4 Project Scope

- Hidden service labeling by image tagging for generic keywords.
- Confirming co-relation between text data labeling and image tagging

1.5 Future Scope

Predicting type of hidden service based on correlation between text and image tagging.
For example Bitcoin, Drugs, Weapons, etc.

Chapter 2

LITERATURE REVIEW

2.1 Literature Survey

Fidalgo et al[1] presents TOIC (TOR Image Categories) dataset with five different illegal classes. These images in dataset is classified using Bag of Visual Words model with fusion of dense SIFT and Edge-SIFT features that can create an efficient model to detect and categorise illegal content. Edge-SIFT descriptors with fixed radius was proposed by Xie et al[18]. Fidalgo et al[20] introduced concept of Ideal Radius Selection which performs better. Fidalgo et al introduced method of Compass Radius Estimation for Image Classification (CREIC) which estimates the optimum radius value of the Compass Operator for a given dataset to extract the most relevant descriptors from the edge images.Using fusion of these features and SIFT descriptors, Fidalgo et al. obtained higher accuracy rates than Edge-SIFT descriptors. First Fidalgo[1] tested CREIC method on a well-known dataset for classification, Butterflies [13] and then on TOIC. The method yields an accuracy of 86.62% in Butterflies and 92.49% in TOIC.

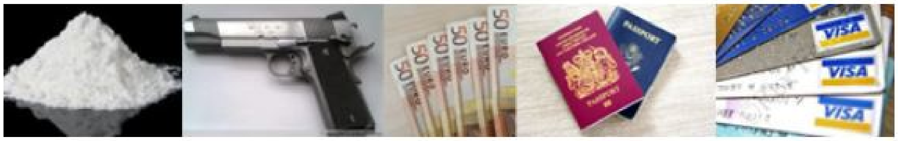


Figure 2.1: Classes in TOIC Dataset

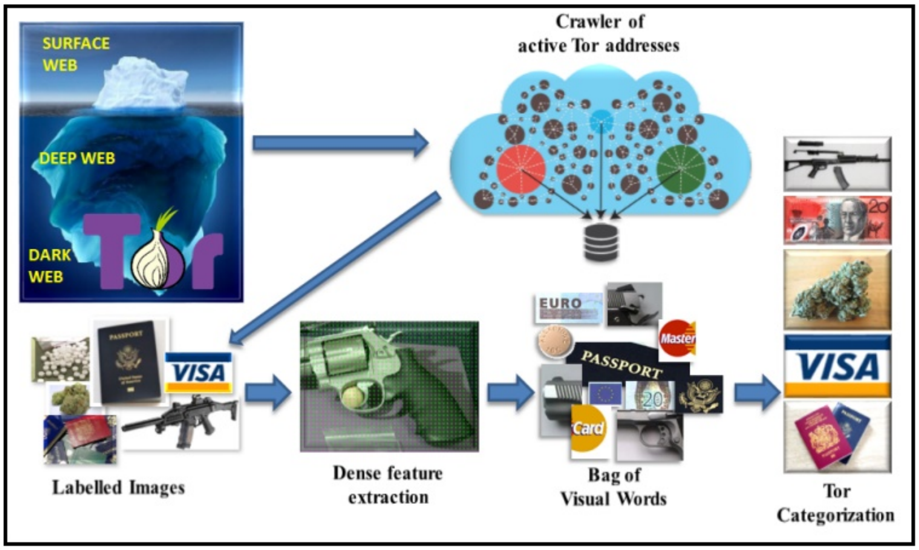


Figure 2.2: Overview of Proposed Method for TOIC Categorization [1]

Vaibhav Pandit.,[9] presents an approach to caption black and white images using transfer learning by implementing Inception v3, a model which is developed by Google. The dataset consists of 8000 photos with upto five captions for each photo. This method yields an accuracy of 45.77% on the validation set.

Dr. S. V. Viraktamath.,[10] provides insights on some related works of paper-currency recognition and has explained the benefits and disadvantages of various currency recognition systems. Most of the algorithms analysed used images of original currencies taken from camera to create datasets and compared these values with the test images to differentiate between original and counterfeit.

Risab Biswas.,[11] illustrates the importance of drug identification and drug discovery in world healthcare sector and how this solution shows accurate results in identifying the drugs correctly, given the molecular structures.

Tufail Sajjad Shah Hashmi.,[12] presents a comparative analysis between YOLOV3 and YOLOV4 for weapons detection. A total of around 9000 images containing gun and pistol was used for this research. The dataset was divided into train and test data and evaluated using both YOLOV3 and YOLOV4 models. YOLOV4 outperforms YOLOV3 in this comparison.

Enrique Alegre.,[7] introduces Semantic Attention Keypoint Filtering, which combines saliency maps with Bag of Visual Words (BOVW). This strategy filters non-significant features from the object of interests at the pixel level. This paper addresses the problem of separating foreground objects with background objects to remove unnecessary details from the subjected images. Eduardo Fidalgo[8] evaluated SAKF on a custom Tor image dataset against CNN features: MobileNet v1 and Resnet50, and BoVW using dense SIFT descriptors, achieving a result of 87.98% accuracy that is outperforming all other approaches.

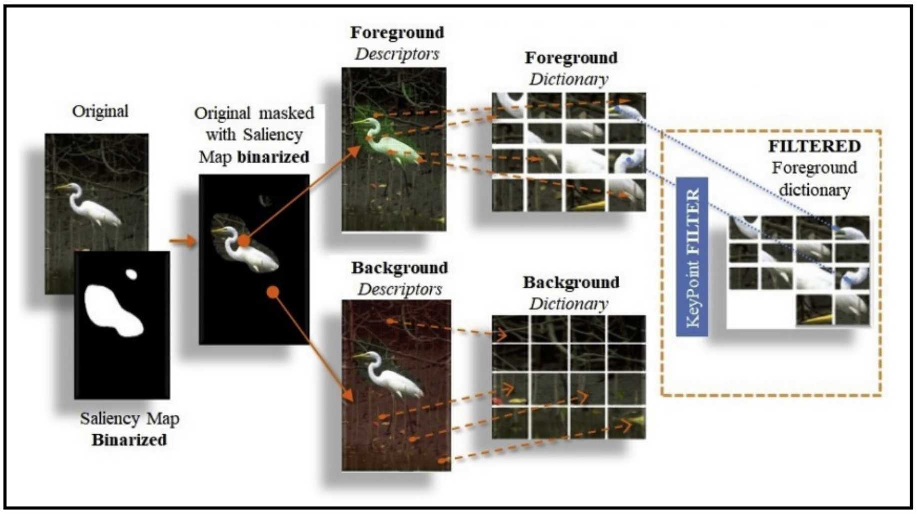


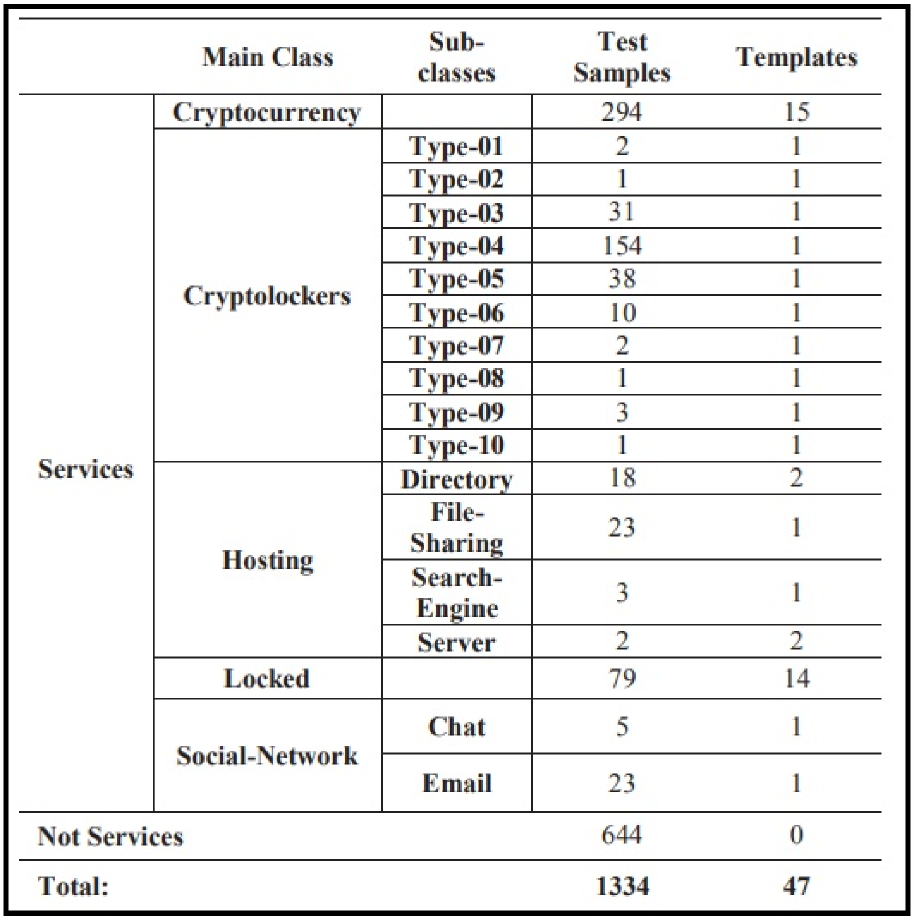
Figure 2.3: Overview of the Proposed SAKF Method[7]

Joao Marques[22] thousands of onion addresses can be collected using memory extraction from servers belonging to the Distributed Hash Table using cheap resources in a small amount of time. This research can be used as a stepping stone to extract intelligence that can be used to to secure legitimate content and monitor or block illegitimate content as anonymity is not only used for providing privacy to legitimate users but also a shield to dark vendors dealing with illegitimate content on the dark web.

Xiangwen Wang.,[21] presents an approach to link multiple accounts of the same darknet vendors through photo analytics. Xiangwen analysed 3 large markets Agora, Evolution, and SilkRoad2, which are now closed, using deep neural networks. However, in this research neither image metadata nor text-based data was considered for classification, which might have improved the model's accuracy.

Susan Jeziorowski.,[13] examines image metadata and explore several image hashing techniques. Their study reveals that approximately 75% of dark marketplace listings include image data, indicating the importance of considering image content for investigative analysis. Also it was found that 2% of image data considered had metadata associated with them and 50% of images hashes were observed to be repeated among marketplace listings which tells about the frequency of image reuse among dark vendors. Thus this research reveals the effectiveness of image hash analysis for identifying similar images between dark marketplaces.

Rubel Biswas.,[16] presents a custom dataset named DUSI (Darknet Usage Service Images) evaluation using both Perceptual Hashing and Bag of Visual Words (BoVW). The dataset is divided into services and not services images.



	Main Class	Sub-classes	Test Samples	Templates
Services	Cryptocurrency		294	15
	Cryptolockers	Type-01	2	1
		Type-02	1	1
		Type-03	31	1
		Type-04	154	1
		Type-05	38	1
		Type-06	10	1
		Type-07	2	1
		Type-08	1	1
		Type-09	3	1
		Type-10	1	1
	Hosting	Directory	18	2
		File-Sharing	23	1
		Search-Engine	3	1
		Server	2	2
	Locked		79	14
	Social-Network	Chat	5	1
		Email	23	1
Not Services			644	0
Total:			1334	47

Figure 2.4: The main features of DUSI dataset[16]

In perceptual Hashing, hash code of each image of the services image is calculated and stored. New image is subjected to hashing and then the hash code is compared with the hash code that is stored. If the Hamming Distance between both the hash code is less than the threshold value(10) then the new image is a service image otherwise it is not. Perceptual Hashing do not require training of the model thus the time investment is less than BOVW. Also Perceptual Hashing provided higher accuracy than BOVW for the given dataset.

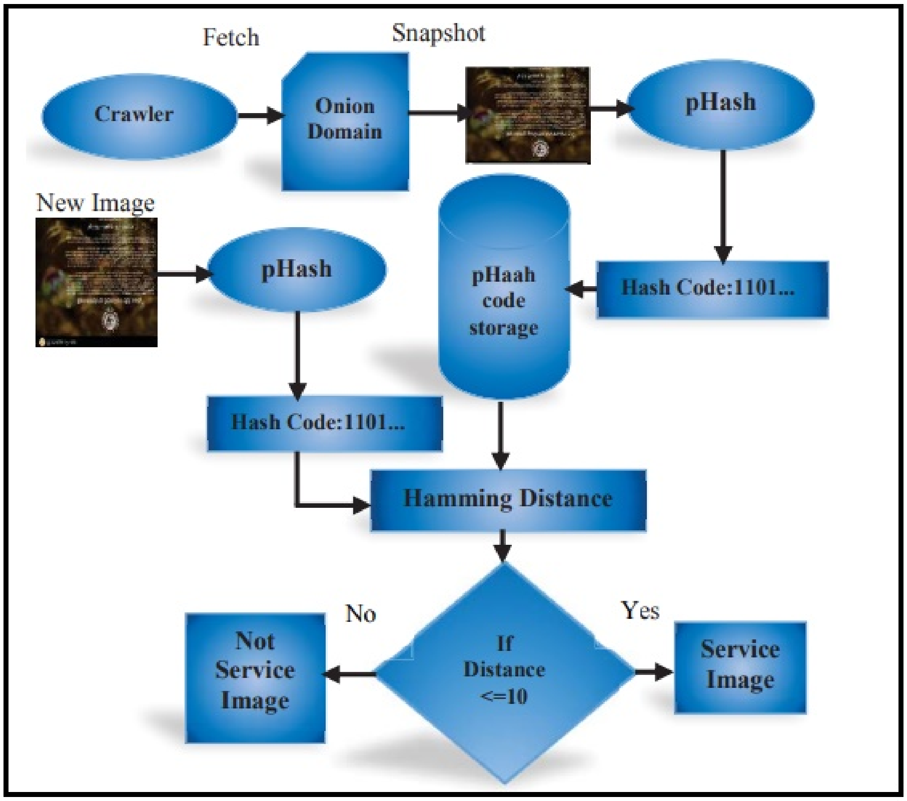


Figure 2.5: Overview of Perceptual Hashing Process[16]

Saiba Nazah.,[3] provides Dark Web threat analysis and detection using a systematic approach following steps as define research questions, develop search strategy, screening and selecting study, data extraction, synthesize and analyze data, report review. Around 65 papers were analysed from various sources. The list of the crime threats were identified from the papers. Many crime detection studies have been done to locate the crimes or criminals in the dark web. The detection techniques and law enforcement methods applied and initiated for this purpose were discussed.

Abhishek Gangwar.,[24] presents a critical review of automatic pornography and Child Sex Abuse (CSA) detection techniques in images and videos. Two publicly available pornographic databases and a real world CSA dataset provided by Spanish Police Forces. Five methods evaluated were Skin Detection by color, Nudity Detection by skin color, HSVColor-SIFT, ShallowCNN, Open NSFW. This research observed that the methods consisting of multiple features performed better than those using simple features like skin color or single image descriptor. Deep learning based methods were observed to outperform all other methods.

Shrey Srivastava.,[25] compares 3 major image processing algorithms: Single Shot Detection (SSD), Faster Region based Convolutional Neural Networks (Faster R-CNN), and You Only Look Once (YOLO) to find the fastest and most efficient of three. Out of the three Object Detection Convolutional Neural Networks that were analysed, Yolo-v3 shows the best overall performance. This comparison is done on the open-source COCO dataset by Microsoft, to ensure a homogeneous baseline. However the choice of algorithms and the result is largely dependent on the use case.

Pankaj Kumar.,[26] introduces a deep learning convolutional neural network (CNN) model to identify Anthracnose disease of mango, which is one of the common diseases in mango plants and can be detected from mango leaves. The images of leaves are captured from the Mango farms in small village in Kolhapur city Maharashtra state and Khanapur taluka in Karnataka state, India. The dataset has been classified in four classes named as Mango Anthracnose, Mango healthy, other diseased and other healthy. The dataset is split into train and validation sets. Model is trained on train dataset and validated using validation dataset. Any real time image of leaf can be subjected to classification through this model.

Wisam A. Qader.,[17] presents overview of Bag of Visual Words (BOW), its importance, its working, applications and challenges. In bag of words (BOW), the number of each word is counted that appears in a document, the frequency of each word is used to know the keywords of the document, and a frequency histogram is made from it.

Shubhdeep Kaur and Sukhchandan Randhawa[29] and Abhineet Gupta[15] explains how dark web is misused and provides an overview of dark web, ways to access dark web, types of criminal activities on dark web, types of attacks using dark web, impact of dark web on cyber security and role played by law enforcement agencies in dealing with it. Thus providing reader with insights on the dark side of dark web so that the reader can take preventive measures while accessing dark web.

Arber S. Beshiri.,[28] discussed and provided results on the indirect number of users in Kosovo and in the world in 2018. Thus providing insights about the influence of dark web in different spheres of society. Not only dark web has influence on the world but global events also have impact on dark web.

Influence of pandemic is discussed by Abdul Razaque.,[23] showing how dark websites related to PPE has impact of COVID-19. It was also observed that the provider, vendor and user of dark web had increased post COVID-19 and thus there is increase in criminal activities whether it is sale of illegal goods or accessing illicit materials.

Akshaya Udgave.,[34] provides the analysis on the use of text mining techniques to help readers keep track of recent developments in the field of design science. Text mining techniques discussed are Information Extraction, Information Retrieval, Categorization, Clustering, Summarization.

Said A. Salloum.,[32] collected and analyzed three hundred different articles from Springer, Wiley, Science Direct, SAGE, IEEE, and Cambridge using text mining techniques. Main tasks for the analysis of text in this study were text clustering (k-means), association rule, word cloud, and word frequency. The articles were analyzed for topic of mobile education for medical domain.

Albert Weichselbraun.,[30] introduces Inscriptis which provides a library that can help converting HTML page contents to plain text. Albert also points out how Inscriptis excels when it comes to interpreting complex HTML constructs such as nested tables when alternatives fail to interpret them. Inscriptis also supports annotation rules which can be used for analysing the HTML structure.

David Mathew Thomas.,[27] used a methodology to gather data from required service and analyse it for the requirements of customer. Social network site Reddit was crawled using the web crawler scrapy and the data was analysed to find out the number of times topics were searched.

Mhd Wesam Al Nabki.,[14] created Darknet Usage Text Addresses (DUTA) dataset that is extracted from the Tor hidden service Darknet. Dataset is divided into twenty-six classes for this research. Wesam has categorized illegal activities of Tor hidden service by using two text representation methods, Term Frequency Inverse Document Frequency and Bag of Words combined with three classifiers, Support Vector Machine, Logistic Regression, and Naive Bayes with high accuracy. The combination of TF-IDF text representation with the Logistic Regression classifier achieved highest accuracy.

Oleksandr Matveiev.,[31] investigated two text categorization approaches K-Nearest Neighbour and the Support Vector Machine algorithms on a JSON dataset containing 40000 entries. The performance of both the algorithms were evaluated on how accurate and how quickly they classified that shoes category based on the brand. The results can further be improved by adding more testing data.

Bassel Alkhatib.,[33] introduced a three step approach to put a dark website under investigation. The approach consisted of three parts. The first part consisted of The Dark Crawler Darky which scans whole website and extracts data from it. The second part included The Cleaner which performs prepossessing on the extracted data. The third part included The Dark Miner which applied Association Rules and Clustering to illustrate all the gained results. The design of crawler and the usage of data mining techniques may differ from one website to another.

2.2 Comparative Research Table

Sr No.	Research	Strengths	Limitations
1	Illegal Activity Categorisation in DarkNet Based on Image Classification Using CREIC Method	Efficient classification of five categories of dark web images	Some specific category of images was addressed. Proprietary software Matlab is used.
2	Evolution of Dark Web Threat Analysis and Detection: A Systematic Approach	Systematic Literature Review (SLR) is provided to identify threats in Dark Web for the researchers and specialists in the Cyber security field.	Methods to analyse image contents of Dark Web to identify the hidden service is not covered.
3	Classifying suspicious content in TOR darknet through Semantic Attention Keypoint Filtering	An efficient method that focuses only on the object of interest is proposed.	The images extracted from TOR do not always show object of interest in ideal situation for the method proposed.
4	Classifying Suspicious Content in Tor Darknet	SAKF is being evaluated on a custom dataset against CNN features: MobileNet v1 and Resnet50, and BoVW using dense SIFT descriptors. It is outperforming other methods.	SAKF can be compared with some additional methods apart from the ones mentioned to check its performance.
5	DeepCap: A Deep Learning Model to Caption Black and White Images	In this paper researcher focuses on a method of captioning of black and white images using transfer learning unlike the existing models which are for coloured images.	Accuracy obtained by the model is around 45%. Accuracy of the model is required to be increased probably by training the model on a bigger dataset.

6	Review on Detection of Fake Currency using Image processing Techniques	The research concluded that K-means algorithm and SVM algorithm provides 97% accuracy in detection of fake currency.	The currency image is taken from only one side or either front or back, which can be further improved by taking the images from different angles.
7	Drug Discovery and Drug Identification using AI	The method introduced can reduce the entire drug discovery process of clinical trials to a very small time of 3-4 months (which generally takes 10-12 years).	Research is focused on drug identification using molecular structure. This will not be helpful in detection of drugs through images.
8	Application of Deep Learning for Weapons Detection in Surveillance Videos	Comparative analysis of YOLOV3 and YOLOV4 for weapons detection. YOLOV4 outperforms YOLOV3.	Weapons used in dataset is only gun and pistol.
9	Towards Image-Based Dark Vendor Profiling	This paper focuses on the effectiveness of using image hashing to identify similar images between dark marketplaces for vendor identification.	This research leaves behind a large set of data which are not related to vendors on the dark web.
10	Classifying Illegal Activities on Tor Network Based on Web Textual Contents	The combination of TFIDF with the Logistic Regression classifier achieved highest accuracy for text dataset Darknet Usage Text Addresses (DUTA) created from hidden services and labelled for research.	Graphical data analysis is not considered leaving out the large portion of hidden services data.

11	The Dark Web as a Phenomenon: A Review and Research Agenda	A literature review was conducted into the roles the dark web plays in modern digital society, its enablement of cybercrime and its relationship with law enforcement.	Methods to analyze dark web data are not covered in the research.
12	Recognition of Service Domains on TOR Dark Net using Perceptual Hashing and Image Classification Techniques	The research demonstrates that perceptual hashing performs better than Bag of Visual Words on DUSI Image dataset.	There was no mention of text data correlation with images. Dataset used is not available publicly.
13	An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges	This study is useful in terms of introducing the BoW method to the new researchers and providing a good background with associated related works.	BOW may face some challenges in which the image will be difficult to be detected or fully unrecognized such as viewpoint variation, illumination, deformation.
14	Spatial Pooling of Heterogeneous Features for Image Classification	A novel framework fusing complementary descriptors for image classification is introduced to provide high accuracy.	Model works well for some datasets but not suitable for others.
15	Semi-local Affine Parts for Object Recognition	The model is focused on identifying image features having a characteristic appearance and elliptical shape and performs well for butterflies' dataset.	Model may not perform well for other type of datasets.

16	Compass radius estimation for improved image classification using Edge-SIFT	The research suggested how different radii of compass operator of original image can have impact of classification accuracy using BOVW	The better radius selection method such as saliency map can be explored.
17	You Are Your Photographs: Detecting Multiple Identities of Vendors in the Darknet Marketplaces	The research demonstrated how photo analytics can be used to identify multiple accounts of same vendor.	This research leaves behind a large set of data which are not related to vendor accounts on the dark web.
18	Tor: Hidden Service Intelligence Extraction	The research shows that it is possible to gather thousands of onion addresses through memory extraction of servers belonging to the Distributed Hash Table in a small amount of time.	The research shows the necessity for improvement in this area to protect the anonymity of the users.
19	Influence of COVID-19 Epidemic on Dark Web Contents	The research identified how the Dark Web has been influenced by recent global events, such as the COVID-19 epidemic.	The investigation experienced drawbacks, such as covering a relatively small portion of the Dark Net.
20	Pornography and Child Sexual Abuse Detection in Image and Video: A Comparative Evaluation	This research observed that the methods consisting of multiple features performed better for automatic pornography and Child Sex Abuse (CSA) detection. Deep learning based methods were observed to outperform all other methods.	Since there is no storage of pornography and Child Sex Abuse (CSA) contents for research. Larger datasets can be used for testing for better and accurate results.

21	Comparative analysis of deep learning image detection algorithms	The research concludes Yolo-v3 shows the best overall performance against SSD, Faster R-CNN for open-source COCO dataset.	This comparison is done on the open-source COCO dataset, to ensure a homogeneous baseline but the choice of algorithms is largely dependent on the use case.
22	Classification of Mango Leaves Infected by Fungal Disease Anthracnose Using Deep Learning	The research will help farmers save their plants by early detection of Anthracnose disease of mango.	The method was focused on a particular disease of the plant. The efficiency of this model on other kinds of diseases is not explored.
23	Data Analysis by Web Scraping using Python	The research shows crawling of a social network site Reddit and the number of times terms were searched on it thus this research can be helpful if the goal is to analyse a site.	The research misses out on analysis of other data apart from the searched terms on site.
24	Dark Web and Its Impact in Online Anonymity and Privacy: A Critical Analysis and Review	The research gives the number of anonymous users in Kosovo and worldwide and provides results about the influence of the Dark Web in different spheres of society.	The accuracy of the counts provided in this paper cannot be verified as anonymity is not verifiable.
25	Dark Web: A Web of Crimes	This research makes aware the reader criminal activities and incidents which take place over the Dark Web and take preventive measures.	The research does not include hidden services analysis.

26	Inscriptis - A Python-based HTML to text conversion library optimized for knowledge extraction from the Web	Inscriptis excels when it comes to interpreting complex HTML constructs such as nested tables when alternatives fail to interpret them and supports annotation rules.	Incriptis working on multiple pages was not discussed in this research.
27	Towards Classifying HTML-embedded Product Data Based On Machine Learning Approach	The research evaluated two popular algorithms for text classification that are KNN and SVM on a JSON dataset providing knowledge about text analysis techniques.	The research gave satisfactory results but not the best. Increasing the testing data can help get better results.
28	Using Text Mining Techniques for Extracting Information from Research Articles	The research provided insights on various text mining techniques and these techniques are used to analyse 300 articles from six databases for the topic of mobile education for the medical domain.	The research was focused on a particular topic rather than exploring more topics which can lead to more interesting patterns.
29	Mining the Dark Web: A Novel Approach for Placing a Dark Website under Investigation	The research introduced an efficient three step method to investigate any website consisting of crawling, cleaning and mining.	The design of crawler and the usage of data mining techniques may differ from one website to another.
30	Text Mining and Text Analytics of Research articles	The research discusses text mining techniques such as Information Extraction (IE), Information Retrieval (IR), Categorization, Clustering, and Summarization.	The text mining techniques are not discussed in detail and are limited.

2.3 Literature Gap

1. Most of the methods or techniques are directed towards analysis of Dark Web graphical data focused on a particular domain like Child Sex Abuse (CSA).
2. A large portion of dark web image data related to other categories is left out.
3. Even if some researches considers having content from more than one class as seen in TOR Image Categories(TOIC) and Darknet Usage Service Images(DUSI), focus is on some specific classes leaving other classes out of consideration.
4. Dataset used these researches are not public.
5. The analysis of text data related to respective Image data to confirm the correlation. This correlation between text and Image data can help in identifying the type of hidden service accurately.

Chapter 3

PROPOSED APPROACH

3.1 Dataset Overview

Dataset that is to be used for the research consists of set of hidden services. There are around 48,700 files in 532 Hidden Services Each hidden service consists of number of HTML, CSS, JavaScript, Image and other files. Each hidden service in the dataset looks as shown in figure 3.1

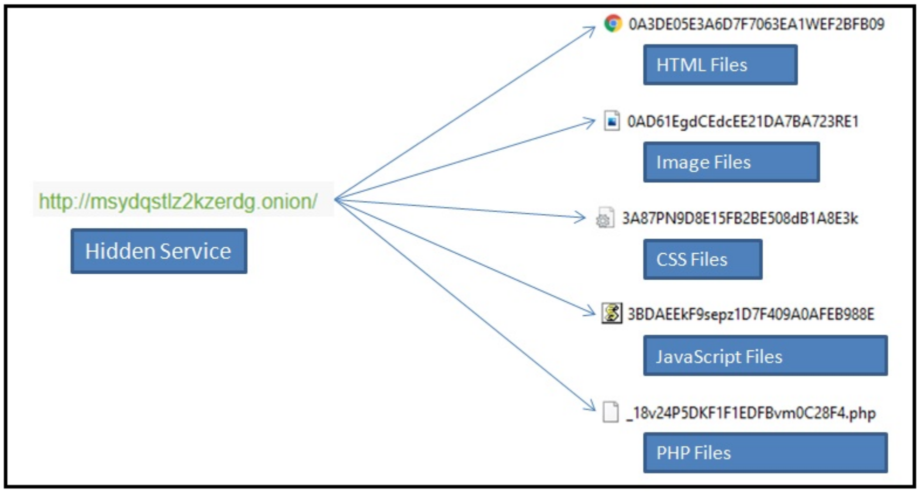

http://msydqstlz2kzerdg.onion/
Hidden Service
0A3DE05E3A6D7F7063EA1WEF2BFB09
HTML Files
0AD61EgdCEdcEE21DA7BA723RE1
Image Files
3A87PN9D8E15FB2BE508dB1A8E3k
CSS Files
3BDAEEkF9sepz1D7F409A0AFEB988E
JavaScript Files
_18v24P5DKF1F1EDFBvm0C28F4.php
PHP Files

Figure 3.1: A Hidden Service in the Dataset

3.2 Overview of Proposed Approach

The proposed approach is shown in figure 3.2.

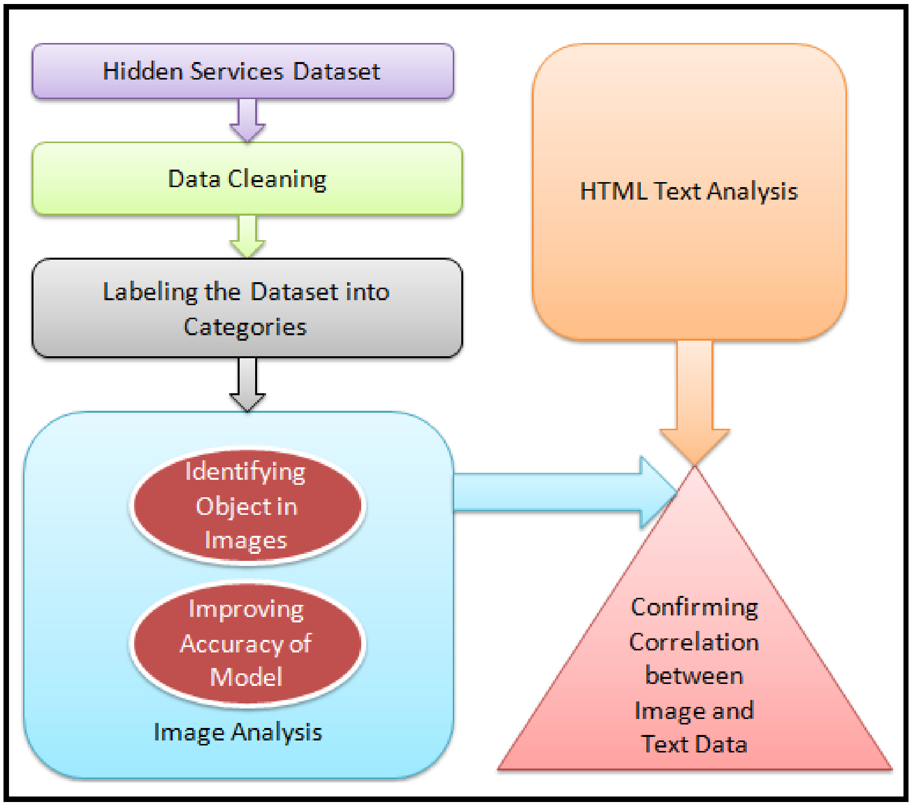

Hidden Services Dataset
Data Cleaning
Labeling the Dataset into Categories
Identifying Object in Images
Improving Accuracy of Model
Image Analysis
HTML Text Analysis
Confirming Correlation between Image and Text Data

Figure 3.2: An Overview of Proposed Approach

1. The dataset will be analysed and cleaned manually.
2. The classes or categories and may be subcategories will be identified and image data will be labelled accordingly.
3. The dataset will be divided into train and test data.
4. The model will be trained on the train data and validated using test data.
5. This trained model can be used to classify any image from the dataset.
6. The text data associate with image mentioned in step 5 will be analysed.
7. Correlation between the Image data and text data will be confirmed on the basis of results obtained in step 5 and step 6.

3.3 Data Cleaning

Figure 3.3 shows approach of data cleaning

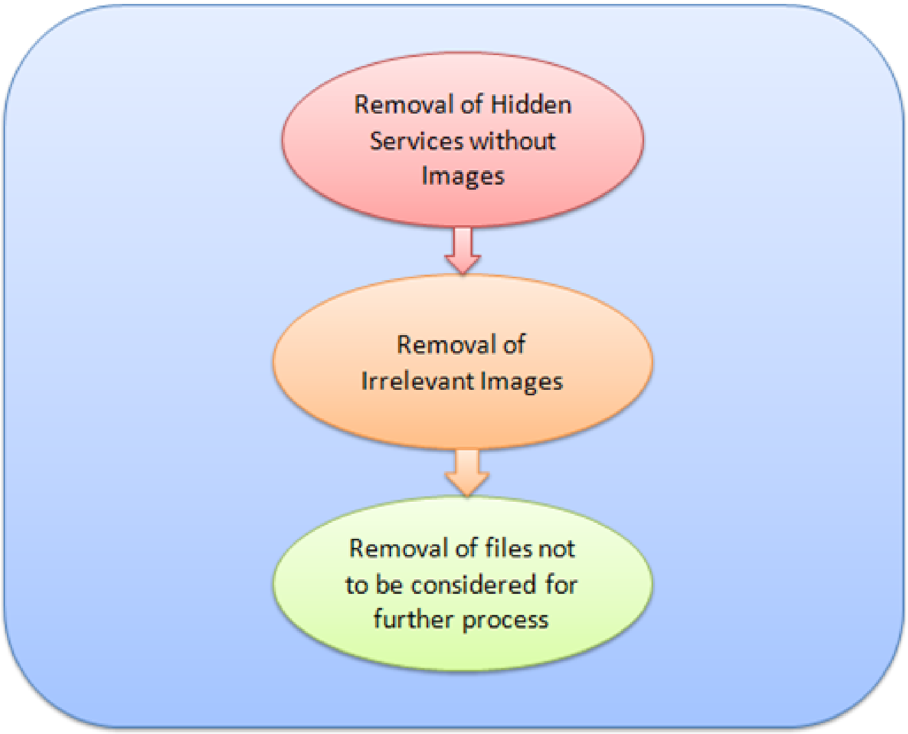

Removal of Hidden Services without Images
Removal of Irrelevant Images
Removal of files not to be considered for further process

Figure 3.3: Data Cleaning Approach

1. The hidden services without image data will be separated.
2. The images which are irrelevant like blank images will be eliminated.
3. The files which will not be considered for further analysis like cascading style sheets will be eliminated.

Chapter 4

CONCLUSION

This research will focus on addressing the gaps of existing researches identified. For this project, dataset that is available is of hidden services containing around 48,700 files and 532 folders consisting of both text and image data. Identifying the model that can provide highest possible accuracy for existing dataset in identifying the objects in the dataset and implementing it will help researchers in future who will work on dataset containing similar classes. The dataset that will be used in this research will be made publicly available for research community to work on it. Image labeling will be confirmed with the textual data of the hidden services.

Bibliography

1. Eduardo Fidalgo, Enrique Alegre, Víctor González-Castro , Laura Fernández-Robles ; "Illegal Activity Categorisation in DarkNet Based on Image Classification Using CREIC Method"; Conference: International Workshop on Soft Computing Models in Industrial and Environmental Applications Computational Intelligence in Security for Information Systems Conference International Conference on EUropean Transnational Education, 2018

2. Ailanthus; https://blog.torproject.org/nine-questions-about-hidden-services/, 2015

3. Saiba Nazah, Shamsulhuda, Jemal Abawajy, Mohammad Mehedi Hassan; "Evolution of Dark Web Threat Analysis and Detection: A Systematic Approach", IEEE Access, 2020

4. Casey Schmidt; https://www.canto.com/blog/image-tagging/, 2019

5. Ryne Knudson; https://brandfolder.com/blog/image-tagging-software, 2021

6. Bethea Davida; https://towardsdatascience.com/bag-of-visual-words-in-a-nutshell-9ceea97ce0fb, 2018

7. Eduardo Fidalgo , Enrique Alegre , Víctor González-Castro , Laura Fernández-Robles ; "Classifying suspicious content in tor darknet through Semantic Attention Keypoint Filtering" , Digital Investigation Journal, 2019

8. Roberto A. Vasco-Carofilis , Eduardo Fidalgo, Francisco Janez-Martino, Pablo Blanco-Medina; "Classifying Suspicious Content in Tor Darknet" , JNIC 2020 Conference.

9. Vaibhav Pandit, Rishabh Gulati, Chaitanya Singla, Dr. Sandeep Kr Singh; "DeepCap: A Deep Learning Model to Caption Black and White Images", 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020

10. Dr. S. V. Viraktamath, Kshama Tallur, Rohan Bhadavankar, Vidya; "Review on Detection of Fake Currency using Image processing Techniques", Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)

11. Risab Biswas, Avirup Basu, Abhishek Nandy, Arkaprova Deb, Kazi Haque, Debashree Chanda; "Drug Discovery and Drug Identification using AI" , Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN 2020)

12. Tufail Sajjad Shah Hashmi, Nazeef Ul Haq, Muhammad Moazam Fraz, Muhammad Shahzad; “Application of Deep Learning for Weapons Detection in Surveillance Videos”, International Conference on Digital Futures and Transformative Technologies, 2021

13. Susan Jeziorowski; Muhammad Ismail; Ambareen Siraj; “Towards Image-Based Dark Vendor Profiling”, IWSPA ’20, New Orleans, LA, USA, March 18, 2020

14. Mhd Wesam Al Nabki, Eduardo Fidalgo, Enrique Alegre, and Ivan de Paz; “Classifying Illegal Activities on Tor Network Based on Web Textual Contents”, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017

15. Abhineet Gupta; “The Dark Web as a Phenomenon: A Review and Research Agenda”, The University of Melbourne, 2018

16. Rubel Biswas, Eduardo Fidalgo, Enrique Alegre; “Recognition of Service Domains on TOR Dark Net using Perceptual Hashing and Image Classification Techniques”, 8th International Conference on Imaging for Crime Detection and Prevention, ICDP-2017, Madrid 13-15 Dec. 2017

17. Wisam A. Qader, Musa M.Ameen, Bilal I. Ahmed; “An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges”, Fifth International Engineering Conference on Developments in Civil & Computer Engineering Applications 2019

18. Lingxi Xie, Qi Tian, Meng Wang, and Bo Zhang; “Spatial Pooling of Heterogeneous Features for Image Classification”, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 5, MAY 2014

19. Svetlana Lazebnik, Cordelia Schmid, Jean Ponce. “Semi-local Affine Parts for Object Recognition”. British Machine Vision Conference (BMVC ’04), Kingston, United Kingdom. pp.779–788. ffinria-00548542f, 2010

20. E. Fidalgo, E. Alegre , V. González-Castro , L. Fernández-Robles; “Compass radius estimation for improved image classification using Edge-SIFT”, Neurocomputing Journal, 2016

21. Xiangwen Wang, Peng Peng, Chun Wang, Gang Wang; “You Are Your Photographs: Detecting Multiple Identities of Vendors in the Darknet Marketplaces” , ASIACCS’18, Incheon, Republic of Korea, June 4–8, 2018

22. Joao Marques ; “Tor: Hidden Service Intelligence Extraction”, University of Amsterdam, 2018

23. Razaque, A.; Valiyev, B.; Alotaibi, B.; Alotaibi, M.; Amanzholova, S.; Alotaibi, A. “Influence of COVID-19 Epidemic on Dark Web Contents”. Electronics 2021, 10, 2744. https://doi.org/10.3390/ electronics10222744 , 2021

24. Abhishek Gangwar, E. Fidalgo, E. Alegre, V. González –Castro; “Pornography and Child Sexual Abuse Detection in Image and Video: A Comparative Evaluation”, 8th International Conference on Imaging for Crime Detection and Prevention, ICDP-2017, Madrid 13-15 Dec. 2017

25. Shrey Srivastava , Amit Vishvas Divekar, Chandu Anilkumar, Ishika Naik, Ved Kulkarni and V. Pattabiraman; “Comparative analysis of deep learning image detection algorithms”, Journal of Big Data, 2021

26. Pankaj Kumar , Sunidhi Ashtekar , Jayakrishna S. S , Bharath K P , Vanathi P. T , Rajesh Kumar M; “Classification of Mango Leaves Infected by Fungal Disease Anthracnose Using Deep Learning”, Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)

27. David Mathew Thomas, Sandeep Mathur; “Data Analysis by Web Scraping using Python”, Proceedings of the Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019]

28. Arber S. Beshiri , Arsim Susuri; “Dark Web and Its Impact in Online Anonymity and Privacy: A Critical Analysis and Review”, Journal of Computer and Communications, 2019

29. Shubhdeep Kaur, Sukhchandan Randhawa; “Dark Web: A Web of Crimes”, Springer Science+Business Media, LLC, part of Springer Nature 2020

30. Albert Weichselbraun; “Inscriptis - A Python-based HTML to text conversion library optimized for knowledge extraction from the Web”, Journal of Open Source Software, 6(66), 3557. https://doi.org/10.21105/joss.03557 , 2021

31. Oleksandr Matveiev, Anastasiia Zubenko, Dmitry Yevtushenko and Olga Cherednichenko; “Towards Classifying HTML-embedded Product Data Based On Machine Learning Approach”, National Technical University “Kharkiv Polytechnic Institute”, Kirpicheva st. 2, Kharkiv, 61002, Ukraine, 2021

32. Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and Khaled Shaalan; “Using Text Mining Techniques for Extracting Information from Research Articles”, Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence 740, 2018

33. Bassel Alkhatib, Randa S. Basheer; “Mining the Dark Web: A Novel Approach for Placing a Dark Website under Investigation”, I.J. Modern Education and Computer Science, 2019

34. Akshaya Udgave, Prasanna Kulkarni; “Text Mining and Text Analytics of Research articles”, - Palarch’s Journal Of Archaeology Of Egypt/Egyptology 17(6). ISSN 1567-214x, 2020