ISSN No:-2456-2165

Classification of Illicit Venture on Dark Web:- A Survey

Zalak Kansagra[1] , Kiran Ahire[2] , Pravin Mishra[3] , Ankur Sarkar [4]
[1]Assistant Prof., Dept. of Computer Science and Technology
Parul University

Abstract:- Dark Web is that part of the internet which cannot be indexed by normal search engines and require special browser to access them i.e TOR. The dark web corpus has huge amount of illicit ventures which are illegal in our country as per the Indian penal code This leads to growth of illegal activities on the web. To collect and categorize the web pages that has illicit content is very time consuming and difficult as well. Hence we propose a method to effectively classify, visualize illicit ventures on the dark web. We select laws and regulations related to each type of illegal activities and trained the classifiers. From various categories of drugs, gamblers, weapons, child pornography and counterfeit credit cards, the corresponding law which prohibits them is selected for training the classifier. Then classifier algorithms like Naives Bayes classifier classify the illicit content on the web pages. The whole purpose of this project is to help the cyber bodies to be versatile and keep record of the illegal contents on web.

Keywords:- *Dark Web, Categorization, Illegal Websites, Indian Penal Code.*

I. INTRODUCTION

To most users, the Internet appears as an endless virtual world of global e-commerce and information. Goods can be purchased and delivered to your doorstep in 2 days (or less), information can be accessed at the swipe of a finger, and dreams come true. But this is only the surface web. The Surface Web also called the Visible Web, Indexed Web, Indexable Web is the portion of the World Wide Web that is readily available to the general public and searchable with standard web search engines. The Surface Web only consists 10 percent of the information that is on the internet. But the Internet has a dangerous, colossal secret: the Dark Web. Today's number one source for stolen information, illegal paraphernalia, .Unlike the Surface Web and the majority of Deep Web content, the Dark Web cannot be accessed through regular web browsers such as Internet Explorer, Google Chrome, or Mozilla Firefox. The most common tool for accessing the Dark Web remains a browser called Tor, which was created by the military to protect oversea communications. Tor was eventually released to the public in 2004, leading to a of developers creating the Tor Project, the method most use to access Dark Web today. While not every site on the Dark Web supports illegal activity, the vast majority of black market activity occurs on the Dark Web today.

II. OBJECTIVES AND SCOPE OF THE STUDY

The objective of our paper is to research in the field of machine learning so as to find an efficient way to categorize the illicit ventures on dark web. The amount of illicit ventures in web are increasing day by day and it has been found that dark web is one of the root source of it. Due to high anonymity of the dark web it is easier for people to do illegal stuffs and don't have to worry about getting caught.

As the dark web is hard to access, research on this field is limited. So we are here to design a efficient mod which will help the Indian Cybercrime department to control these activities and take action in timely manner. What we plan to do is to design a machine learning algorithm which will classify the various illicit ventures under the Indian penal code and use then categorize various type of illicit content present in it. This will help the cyber bodies to monitor these activities in an efficient way. In future we can add new illegal things as we plan to make our model dynamic and versatile.

III. BACKGROUND KNOWLEDGE

Each classification pipeline is comprised of three stages. First –text preprocessing, then features extraction, and finally, classification. We used two famous text representation techniques across three different supervised classifiers resulting in six different classification pipelines, and we examined every pipeline to figure out the best combination with the best parameters that can achieve high performance. Text pre-processing initially, we eliminated all the HTML tags, and when we detected an image tag, we preserved the image name and removed the extension. Furthermore, we filtered the training set for the non-English samples using Langdect11 Python library and stemmed the text using Porter library from NLTK package. After pre-processing the text, we used two famous text representation techniques; (A) Bag-of-Words(BOW) is a well known model for text representation, that extracts the features from the text corpus by counting the word frequency, (B) Term Frequency Inverse Document Frequency Model (TFIDFM) (Aizawa, 2003) is a statistical model that assigns weight for the vocabularies where it emphasizes the words that occur frequently in a given document. For each features representation method, we examined three different supervised machine learning algorithms.

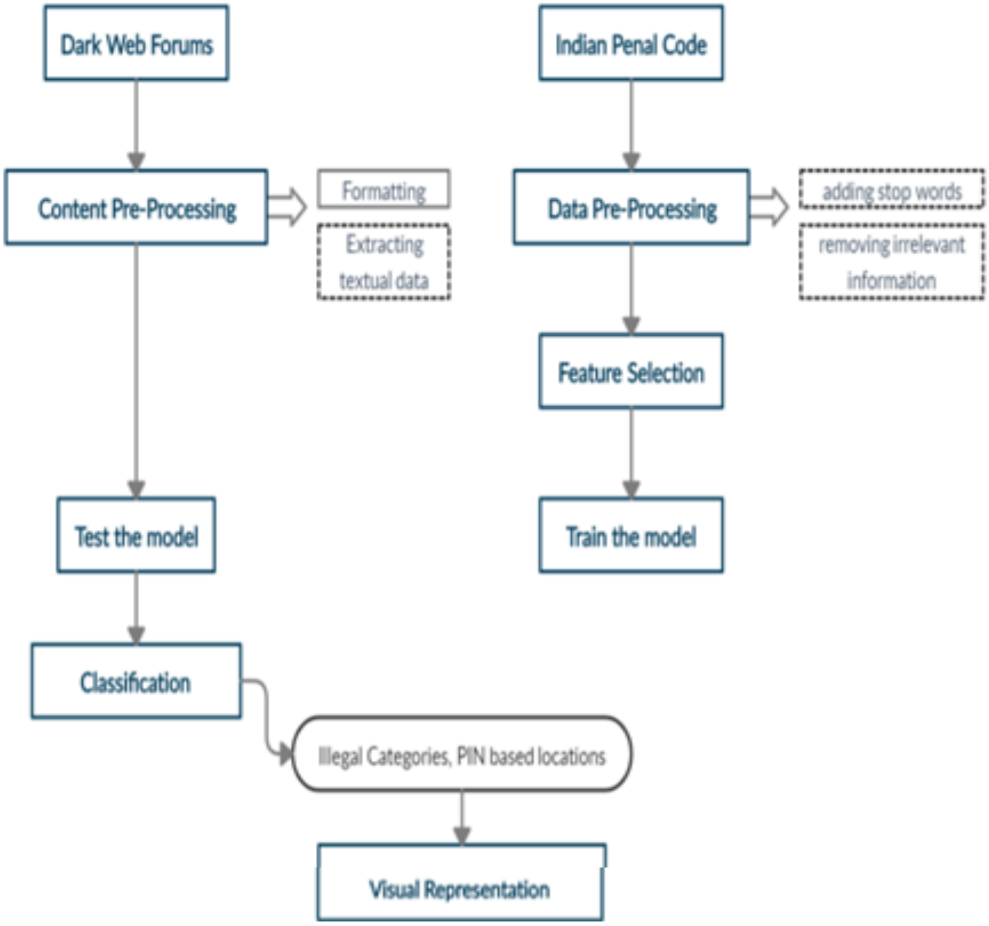

Fig. 1

We are going to classify illegal activities on dark web as following:

1. First we will design a crawler which will help us to extract the .onion forums from dark web
2. Then we will Select legal documents as per The Indian penal code which we provide to our classifier as training data
3. Then we test the classifier by providing the dark web forums as test data sets
4. We will use various machine learning algorithm to find which algorithm gives us the best result .

IV. SURVEY OF THE LITERATURE

This section comprises some of the literature's on various classification techniques that has been used to classify the various illicit content's on the dark web, which was developed by various researchers and we have used them.

The brief knowledge about the deep web how it works what we see on surface web and what's going on behind it. All the activities which have some private content related to bank account details, health and etc. are managed by deep web. According to research about ninety-five percent of the content on deep web is accessible without any subscription On deep web there is wealth of information which can be leveraged by different sectors of the society such as: Researchers, Finance & Markets. In terms of size the deep web is about 400 -550 times larger than the surface web. Deep web content is believed to be about 500 times bigger than normal search content, and it mostly goes unnoticed by regular search engines.

As Dark web is hidden from the common web, user dark web by nature is anonymous, most attempts has that uses the dark web are far from being safe if they were to just access it blindly. Many precautions need to be taken before one can even thinking of playing around on the dark web. Many security agencies such as the CIA & FBI [8] do surf the dark web to monitor the activity there but still due to the fact that the most attempt are failed. Here are several ways to access the dark web, one of which includes the use of a Tor browser, Freenet and I2P, among which the most popular method is by using Tor. Tor was initially known as the onion router because of the dark websites having the dot onion domain.(.onion).

Detection we presented and made publicly available TOIC (TOR Image Categories), a dataset which comprises five categories representing part of the illegal content found in the Darknet TOR related with Money, Weapons, Drugs, Personal Identification and Credit Card counterfeit. We applied the ideal Radius Selection and Compass Radius Estimation for Image Classification (CREIC) methods proposed by Fidalgo et al. [10] to both TOIC and Butterflies

dataset. The results obtained demonstrate how the selection of relevant edge features can boost the performance of a basic Bag of Visual Words (BoVW) coded with dense SIFT and Edge-SIFT features. The promising results obtained on TOIC dataset make CREIC an interesting framework to be used by law enforcement agents in their fight against criminal activity in the network by supporting them in the categorization of the TOR content.

A method is proposed that can effectively classify illegal activities on the dark web. Instead of relying on the massive dark web training set, laws and regulations related to each type of illegal activities are carefully selected to train the machine learning classifiers and achieve a good classification performance. The results show that combined with TF-IDF feature extraction and Naive Bayes classifier, an accuracy of 0.935 is achieved in the experimental environment. This approach allows researchers and the network law enforcement to check whether their dark web corpus contains such illegal activities based on the relevant laws of the illegal categories they care about in order to detect and monitor potential illegal websites in a timely manner.

The study on a set of topologically dedicated hosts discovered from malicious Web infrastructures. Those hosts are found to play central roles in the Dark Web, serving over 70% of the nearly 4 million malicious redirection paths collected in the research and rarely being connected to by any legitimate hosts. Leveraging this unique feature, a topology based technique is developed that detects these hosts without even knowing exactly what they do. This approach utilizes the Page Rank algorithm to capture those with high status on the dark side of the Web but very much unknown on the bright side, and brings to the light thousands of dedicated hosts missed by the state-of-the-art malware scanner. Taking a close look at these findings, we learn that many of those hosts are actually TDSes, which play a key role in traffic exchange in malicious activities. Further study on such services reveals their unusually long life span(65 days), compared with the exploit services studied before, and the way they are used to monetize traffic, even after their domains have been taken down. It's also observed that 61.66% of the parked TDSes use ad-networks and ad exchanges such as Double Click and Bid System. While 56% go through tracker networks used for targeted advertising, 3.94% of the parked paths monetizing traffic directly through the Zero Click model.

Categorizing illegal activities of Tor HS by using two text representation methods, TFIDF and BOW, combined with three classifiers, SVM, LR, and NB .Their aim is to build an image classifier to work in parallel with the text classification. The high accuracy that they have obtained in this work might represent an opportunity to insert their research into a tool that supports the authorities in monitoring the Darknet.

V. CONCLUSIONS

As the project is still on the development phase , the only conclusion we can draw is that the given system will categorize various illegal activities on dark web which will help the cybercrime agencies to take rapid actions . In future we may have new categories of illegal content ,so we try to make our system versatile so as to accommodate future changes

VI. FUTURE WORK

In case of illegal activities detection or any other model the higher the accuracy better the model. So in future we should try to increase the accuracy of the model. In some activities the dataset is not proper or contain less number of text which effects the accuracy. So in future we should try and add more text and increase our dataset.

REFERENCES

[1]. DavidMatilla,V´ıctorGonz´alez-Castro, Laura Fern´andez-Robles,EduardFidalgo, MhdWesam Al-Nabki - IEEE Transactions on Information Forensics and Security, 2018
[2]. Robert Koch - 11th International Conference on Cyber Conflict, 2018
[3]. James Cardon –Department of Business and Management ,LUISS Guido Carli University,Viale Pola 12,00198 Roma,Italy,2018.
[4]. L Rubel Biswas, Eduardo Fidalgo and Enrique Alegre-Department of Electrical, Systems and Automation , Universidad de Leon, Spain,2017.
[5]. RZhou Li, SumayahAlrwais,YinglianXie, Fang Yu, XioaFeng Wang – IEEE Symposium on Security and privacy,2013.
[6]. E.Fidalgo, E.Alegre,V.Gonzalez-Castro,andL.FernAndez-Robles-Illegal activity categorization in Dark net based on image classification using CREIC method,2017.
[7]. M.Dahiya,IJCSE- International Journal of computer Science & Engineering, volume-5/ Issue-5,13 May 2017
[8]. Barber S., Boyen X., Shi E., Uzun E. (2012) Bitter to Better —Jizhou Huang, Ming Zhou, Dan Yang, IJCAI - International Joint Conferences onArtificial Intelligence, Volume – 07, 2006.
[9]. JOOSasaArsovski , Adrian David Cheok , Muniru Idris , MohdRadzee Bin Abdul Raffur, Imagineering Institute Article, February 2017.
[10]. Arbër S. Beshiri, ArsimSusuri - Dark Web and Its Impact in Online Anonymity and Privacy: A Critical Analysis and Review, 2019.