Categorization of Illegal Activities on Dark Web using Classification

Hrushikesh Thorat[1], Shubham Thakur[2], Amit Yadav[3]

[1-3]*Department of Computer Engineering PHCET, Rasayani*

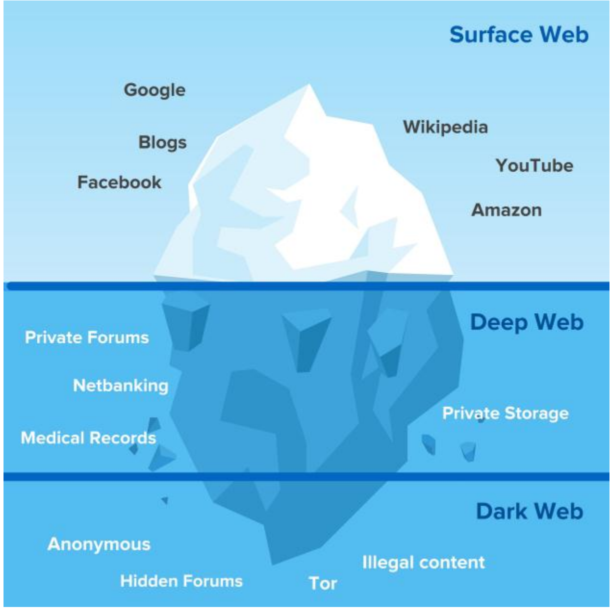
---***---

Abstract— The Dark Web is a part of WWW, which is only accessible by means of special software, which allows users and website owners to remain anonymous. The dark web corpus contains activities that are illegal under the Indian Penal Code and IT Amendment Act 2008. This leads to growth of illegal activities on the web. The collection and labelling of illegal dark web content web pages is difficult and time consuming. The method proposed in this project can effectively classify, visualize illegal activities on the dark web. We creatively select laws and regulations related to each type of illegal activities and trained the classifiers. From the categories of drugs, gamblers, weapons, child pornography and counterfeit credit cards, we picked corresponding legal documents from the Indian Penal Code (IPC) for supervised training. Then classifier algorithms like Naive Bayes classifier classify the illegal content on the web pages. This will help Indian Cyber Crime Department to monitor potential illegal activities and their corresponding websites in a timely manner. This classification defines a new way of categorizing illegal activities on the dark web.

Keywords: Dark Web, Categorization, Illegal Activities, Visualization

I. INTRODUCTION

Dark Web, also known as 'Dark Net' is the online content consisting of web pages and forums that are encrypted with least possibility of tracing the exact location of servers hosting the online content. Dark Web content also cannot be indexed by typical search engines such as Google, Bing, Yahoo, etc. On Dark Web, the identity of the user and the website owner, server locations remain anonymous. Today, various open source browsers such as Tor, Freenet, I2P are available for free to access Dark Web content. According to Tor Metrics Project [1], there are more than 200K registered .onion addresses as of May 2020, and on an average, 2 million users use Tor on daily basis.



[2]

Fig. 1. Divisions of World Wide Web

The Tor network was developed by United States Naval Research Laboratory in 1990's for military purposes as the onion routing principle gave anonymity to transmit confidential data with encryption. Later, in 2004, the original 'The Onion Routing Project' was made free to public under a free and open source license by the name, 'Tor Project'.

Every computer has a specific address called Internet Protocol (IP) which is provided by local Internet Service Provider (ISP). An IP address is also associated with a Domain name and both IP address and domain name is routed through ISP's servers. To identify a user's location, one can easily trace the IP address for the user's device. Tor, on the other hand, uses onion routers which bounce a connection through a wide network of relays all over the world. This gives anonymity to users and the web page they are accessing. Now, most of illegal activities such as buying and selling of drugs, weapons, confidential data, sensitive records are fortifying in the Tor network because of its anonymous nature.

Illegal activities and services being a major part of online activities that are practiced on Dark Web when categorized, would be very helpful for Indian Cyber Crime Department and various Research Organization to keep track of types of activities that evade the Indian Law and its Rules and Regulations.

II. LITERATURE REVIEW

Several researches have been made in the field of website classification. The classification includes extraction of text content from web pages, classifying the text by weighing techniques like Term Frequency-Inverse Document Frequency (TF-IDF). And various classification algorithms such as Naïve Bayes and SVM are used to build a classifier. In case of Dark Web, very few researches have been done so far in the field of classification of illegal activities due to its anonymous nature. Most of the researches make use of large datasets collected from Dark Web to train their classifiers which is lengthy and time consuming. While re- searches using a different training data only shows the classification of activities that are considered illegal under US Legislative.

Siyu He, Yongzhong He and Mingzhe Li, in ”Classification of Illegal Activities on Dark Web” [3] proposed a classification method that uses 'Federal Code of United States of America' as training data to their model which gave them accuracy of 0.935

Al Nabki, M. W., Fidalgo, E., Alegre, E., and de Paz, I., in ”Classifying illegal activities on TOR network based on web textual contents” [4] classified certain categories of activities in Dark Web by creating using DUTA (Darknet Usage Text Address) which has to manually label the extracted web content Hussein Alnabulsi1, Rafiqul Islam, in ”Identifi-cation of Illegal Forum Activities inside the Dark Net” [5]. They made use of posts from selective Dark Web forum URLs and trained their model to classify those posts into different activities upon testing on new set of URLs.

III. SYSTEM ARCHITECTURE

The architecture is divided in three main parts– Collecting relevant laws and regulation (Train- ing Data), Extracting Dark Web content (Testing Data), and Classification of illegal activities. The system Architecture is shown below. The system

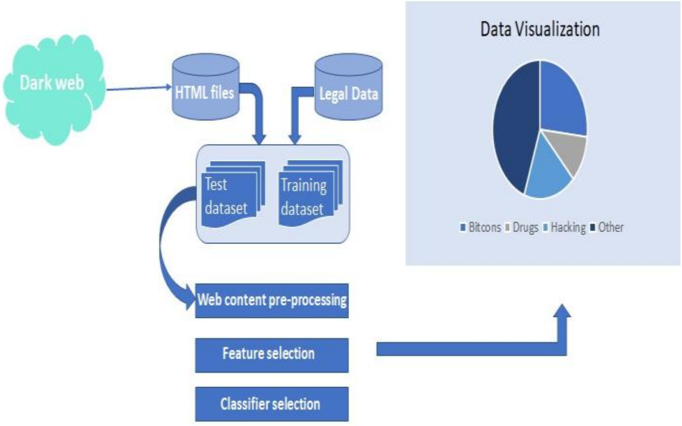


Fig. 2. System Architecture

consists of 2 different datasets - Legal and Illegal data for Training and Testing purpose, respectively. The training data is the data we will select from legal documents while Testing data is the one we will extract from the Dark Web forums. We will use training data to train our model and our classifier will test it of Test data we extracted.

A pre-processing is need to be done on both the datasets before we use them. We then use, Feature extraction techniques on training data using TF- IDF feature extraction to build a Vector Space Model. After applying classification algorithm on Test dataset i.e. collected Dark Web Forum pages, we get to know the meaningful insights of different categories of illegal activities carried on Dark web considered criminal as per IPC (Indian Penal Code). We use different Visualization techniques mainly, pie chart to display our result which makes easier to analyze our research. We will also com- pare the accuracy and choose the best algorithm for classification of illegal activities on Dark Web forums.

IV. METHODOLOGY

A. Selecting Legal Documents

In order to get better results, we decided to select few Laws, Legal amendments and Judiciary case studies which are related to illegal activities that can be done online in Dark Web corpus. We choose, few Sections of IPC [6], IT Amendment Act 2000, 2008 [7] and Judiciary case studies in Cyber crime as the part of our training dataset. We removed the subject chapter names, irrelevant information and labelled each extracted text ac- cording to the crime involved.

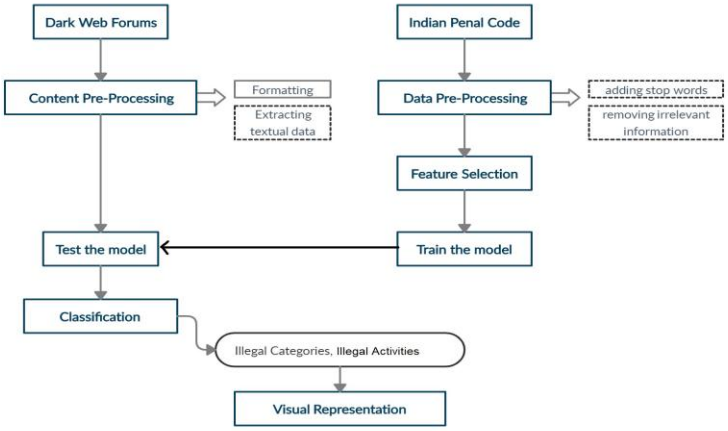


Fig. 3. Process

B. Extraction of Dark Web Forums

Illegal data (Testing Data) in our project refers to the data extracted from Dark Web. This data is nothing but the web content of few websites we choose where we believe that illegal activities are handled. Generally, large amount of web pages are downloaded and few pages are used as training data while the rest of the corpus is tested accordingly. But this is a time-consuming process, and requires manual labeling of large number of web pages. In our project, we choose a unique way to extract the dark web data by making our crawler to crawl only the web pages with .onion extension. To start with which website to crawl first, we selected few known Dark Web forums where it seems to conduct illegal activities such as selling of Drugs, Weapons, Child Pornography, etc. Following are the forums we choose for our project from DeepWebsiteLinks[8]: http://underdj5ziov3ic7.onion/thread/drugs/pg http://oxwugzccvk3dk6tj.onion/index.html http://xmh57jrzrnw6insl.onion/ http://parazite.nn.fi/roguesci/

C. Classification of Illegal Activities

Since, to make classification as accurate as pos-sible, we divided it into 3 parts: Pre-processing, Feature extraction and Classifier selection

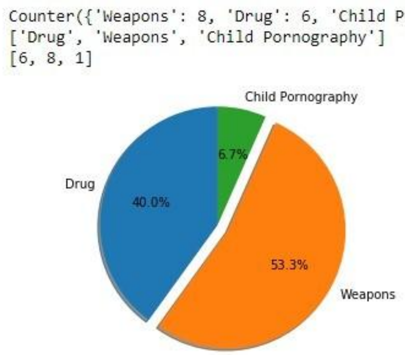
Web Content Pre-Processing: Pre-processing is done on extracted website data from the Dark Web forums by crawling selected .onion links. We involved formatting of extracted HTML content by removing irrelevant content such as images attached, HTML tags, and outbound and 'noindex' Links.

Feature Selection: To build a Vector Space Model, we used TF-IDF (Term Frequency - Inverse Document Frequency) weighing technique. Reason to choose TF-IDF was it gives importance of each words rather choosing words based upon their number of occurrences throughout the dataset.

Classifier Selection: Now that of Vector Space Model is ready, we have to choose a Classifier algorithm to test it on our preprocessed testing data. We selected Naïve Bayes, Support Vector Machine (SVM) and Random Forest (RM) algo-rithms for classification.

V. RESULTS

This section will give us the results we achieved after applying different classification algorithms. First, with the dark web forums pages we extracted following are the types of illegal activities we were able to classify as we labelled them according to the selected legal documents. Looking at the results (Fig. 4), we found that 53.3 percent of the forum pages that we selected are involved trading of weapons, illegally. Whereas, 40 percent of pages were involved in Drug market and 6.7 percent of which were involved in Child Pornography. This tells us, the Dark Web corpus is mostly use for



```
Counter({'Weapons': 8, 'Drug': 6, 'Child Pornography': 1})
['Drug', 'Weapons', 'Child Pornography']
[6, 8, 1]
```

Fig. 4. Classification Results

Following three illicit campaigns - Weapons, Drugs and Child Pornography.

A. Accuracy

We chose three different classifier algorithms - Naive Bayes, SVM and Random Forest classifica-tion Algorithms and check for the accuracy of the results. We received the Accuracy score of 88.889

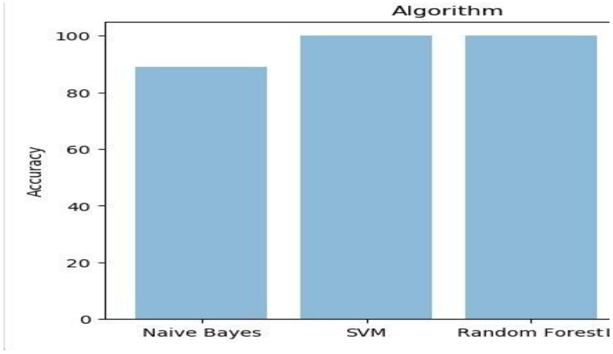


Fig. 5. Classification Results

from Naïve Bayes Algorithm with recall and F1 score of 0.875 and 0.8831168831 respectively. While both SVM and Random Forest showed same accuracy of 100 with recall and F2 score of 1.0 and 1.0 respectively.

VI. CONCLUSION

This paper successfully shows the classification of Illegal activities that are being conducted on different Dark Web Forums by selecting legal data from few Indian Criminal Codes such as IPC (Indian Penal Code) and IT Amendment Act 2008. We effectively train the model with training dataset by first removing all the irrelevant information and then using TF-IDF (Term Frequency – In- verse Document Frequency) as a feature extraction model to create Vector Space Model. This makes it easy to add a greater number of categories to classify. We used Spyder crawler to crawl illegal Dark Web corpus using Tor network. We gave certain forum links to the crawler to kickstart the crawling and collect the textual data. Later, we used SVM (Support Vector Machine) and Naïve Bayes algorithm for classification on which we achieve a performance with good accuracy. There hasn't been any classification technique to classify illegal activities on Dark Web based on Indian Criminal Code. Our method combining with legal documents such as IPC gives us classification of activities considered illegal under Indian Legisla-tive.

REFERENCES

[1] Tor Metric Project, https://metrics.torproject.org/hidserv- dir-onions-seen.html, https://metrics.torproject.org/userstats- relay-country.html

[2] Iceberg Free License Image by VvStudio, https://www.freepik.com/

[3] Siyu He, Yongzhong He, Mingzhe Li (2019), "Classifica-tion of Illegal activities on Dark web". Published on ACM Proceedings of the 2019 2ndInternational Conference on Information Science and Systems.

[4] Al Nabki, M. W., Fidalgo, E., Alegre, E., dePaz, I. (2017). "Classifying illegal activities onTOR network based on web textual contents". In Proceedings of the 15th Conference of the Euro-pean Chapter of the Association for Computation-alLinguistics.

[5] Hussein Alnabulsi1, Rafiqul Islam, School ofComputing and Mathematics, Charles Sturt Uni-versity, Albury (2018), "Iden-tification of Illegal Forum Activities inside the Dark Net" 2018 In-ternational Conference on Machine Learning andData Engineering (iCMLDE)

[6] Indian Penal Code (Amendment) Act, 1921 (16 of 1921), sec.4. "INDIAN PENAL CODE (IPC)"

[7] The Information Technology ACT, 2008, Ministry of Law, Justice and Company Affairs (Legislative Department) New Delhi, the 9th June 2000/Jyaistha 19, 1922 (Saka), https://meity.gov.in/content/information-technology-act.

[8] Deep Website Links website - https://www.deepwebsiteslinks.com/deep-web-forums-links/