IDENTIFICATION AND CLASSIFICATION OF ILLEGAL CONTENT ON TOR DARK WEB

Thesis submitted in fulfilment of the requirement for the degree of

Doctor of Philosophy

IN

INFORMATION TECHNOLOGY

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
LUCKNOW
प्रज्ञा शील करूणा
ESTABLISHED 1996

Submitted by

Mohd Faizan

Supervised by

Prof. Raees Ahmad Khan

Submitted to

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

JULY-2021

ABSTRACT

The dark web is a layer of the Internet that lies beneath the publicly accessible layer of the surface web. The dark web sometimes also referred to as the darknet and can be accessed only through specialized software. The part of the dark web employing the onion routing technique is accessible through the Tor browser and constitutes the Tor dark web. The websites on the Tor network are referred to as the hidden services. Several studies in the literature have reported the presence of illegal and controversial content on the hidden services including but not limited to illegal drugs and firearm trafficking, child abuse content, violence and extremism.

The present work in this thesis attempts to provide solutions to the law enforcement agencies to tackle illegal content on the Tor hidden services. These solutions leverage the machine learning techniques, natural language processing and graph theory. The techniques utilize the textual content of the Tor hidden services for identification and classification of the illegal content. Among all the dark web networks, it has been found that the Tor network is forefront in hosting and promoting the illicit content, therefore, in this thesis we have exclusively focused our work on the Tor dark web. The proposed solutions can be broadly divided into three components: Hidden Services Classification, Content Based Identification and Link Based Identification.

The first component is a machine learning classification model employing a two-step dimensionality reduction scheme for identification and classification of the illegal content on the Tor hidden services. The two-step Dimensionality Reduction (DR) scheme uses Mutual Information to select representative features in the first step followed by the transformation of the selected features by the Linear Discriminant

Analysis in the second step. The effectiveness of the two-step DR scheme is evaluated on the three classifiers namely: Logistic Regression, Naïve Bayes and the Support Vector Machines. The proposed model requires a related dataset for the training and validation purpose. Therefore, a labeled dataset called dark web text dataset has been created that contains 4102 instances categorized into 31 classes including both licit and illicit content. The samples in the dataset are collected with the help of a customized Python crawler.

The proposed model with the two-step DR scheme is implemented on the dark web text dataset and evaluated with the standard metrics. The Logistic Regression classifier produces the best classification performance in terms of the *f-score* when supplied with the feature set obtained using the two-step DR scheme. Moreover, the two-step DR techniques produced better results as compared to other dimensionality reduction schemes. The proposed model with the two-step DR scheme also outperformed the baseline approach for classification of the illegal content on the Tor network.

The second component is based on the textual content of the hidden services for identifying key hidden services involved in drugs and firearm trafficking. In case of drugs trafficking, a ranking methodology is proposed to rank the Tor hidden services based on the severity of the drugs available on it. Drugs Name Recognition (DNR) is used to extract the name of drugs and their street names from the product listings on the hidden services. A metric is proposed that calculate the harm score of a hidden service based on the toxicity of available drugs identified using DNR. The hidden services are then ranked in order of their computed harm score. The top ranked hidden services are the ones that deal in potentially harmful drugs. The effectiveness of the ranking methodology is evaluated against the ground truth data specifically created from the

dataset. The proposed ranking methodology produces good result in terms of the Kendall's tau and the Rank Biased Overlap metrics. The law enforcement agencies could prioritize their efforts on the hidden services that sit at the top positions in the ranked list.

The next part of the component focuses on the illegal firearm trafficking on the Tor network. The literature review of the studies on the firearm trafficking on the dark web has identified that the handguns and rifles are among the popular and preferred items. Moreover, it has been found in several instances of violence in the past in major cities of the world that handguns and rifles were used to spread terror. Therefore, we have proposed an ensemble machine learning model for identification of handguns and rifle listings on the hidden services. The ensemble consists of the Naïve Bayes and Random Forest as the base classifier that were combined using stacking technique. The ensemble model was provided an engineered feature set consisting of Part-of-Speech (PoS) tagged features and their N-grams. The PoS tagged feature set with unigrams and bigrams produce better result than the complete feature set. Also, the ensemble model outperforms the individual classifiers in terms of Precision, Recall and *f-score*.

In our next contribution, we uncover the topological properties of the Tor network that motivate us to develop the third and final component of our proposed solutions. A web graph of the Tor network is generated where a node represents the hidden service and an edge represent a hyperlink between the two hidden services. The majority of the nodes in the Tor web graph have in-degree and out-degree below ten. Also, the graph is weakly connected with only few connected pairs of nodes. Moreover, the Tor network possesses a bow-tie structure similar to that found in the surface web though smaller in size. The Tor network also

exhibit the small-world and scale-free characteristics as found in the surface web.

Finally, in our last component, we propose a hyperlink based algorithm for identifying prominent nodes in the Tor web graph. The link analysis algorithm is based on a modified PageRank algorithm for identifying the influential nodes in the Tor network. The influential nature of a node depends on two factors: number of hyperlinks originating from a node towards the surface web and the central location of the node in the graph which is obtained from the graph centrality metrics. The combined influence score is embedded into the PageRank algorithm to compute the overall influence of the node in the network. The nodes are then ranked using the overall influence score. The accuracy of the ranking algorithm is evaluated using graph robustness metrics.