Exploring the Dark Web for Cyber Threat Intelligence

Shrithi Batthini*, Vaibhavi Honagekar*, Ishika Joshi*, and Prof.Charumathi K.S.**

student*,professor**

Department of Information Technology, Pillai College Of Engineering,

Navi Mumbai,Maharashtra, India

Abstract— Cyber attack strategies have become more prominent in recent years, making it more difficult to stop an assault, even if some form of countermeasure is used. Even with protection systems in place, it is challenging to totally avoid all cyber intrusions. We are, in a sense, in a defense-only posture. A strategy for exploiting this threat will be provided, which will combine social network analysis with cognitive computing to filter hostile behavior and forecast their vulnerabilities. The major goal is to harvest crucial postings from dark web forums, including posts from malicious hackers, and disseminate information like virus trading and hacking strategies. Hence ,We are proposing an approach to extract forums which include important information or intelligence and identify traits of each forum using AI and ML Techniques. Information gathering from the dark-web communities is quite a challenging task, data cannot address real-world cybersecurity problems considering the importance of information gathering. The dataset containing forum posts from the darknet were extracted from different websites using different web crawlers .Then the retrieved data is cleaned , normalized, parsed, processed and classified and labeled according to their severity. After this phase different ML algorithms were applied on labeled data . Finally the model is tested with the inputs and used for predicting the cyber threats for future. The main focus of this research will assist us in identifying future developing risks in cyberspace and high propensity measures to counter such malicious tasks. With recent technological advancements, recent research has revealed that dark-web indications may be connected with event data and leveraged to forecast intrusions, indicating perhaps a more threat-focused cybercrime is on the horizon.

Keywords—Intelligence, Darkweb, Forums, Machine Learning

1. Introduction

The majority of cyber attacks today are orchestrated attacks targeted at financial fraud, and they are becoming more common. To get out of this scenario, you need to be able to forecast cyber attacks and adopt proper remedies ahead of time. Current methods for dealing with cybercrime are reactive, which means that cyber specialists respond only after a breach has occurred. Because the hackers who carry out these attacks frequently mask their activity and goals, cyber-threats are difficult to detect and forecast. They may, however, continue to discuss vulnerabilities and offer tradecraft on how to compromise them in publicly accessible forums.We're demonstrating a way for leveraging this concern that combines social network analysis, machine learning, and cognitive computing to filter hostile actions and foresee their vulnerabilities. This analysis concentrates on the representation with several sorts of threat and attack information. The identification of a threat actor, such as a person, an organization, or a nation state, is the most important aspect of the model since it determines any methodology, tactics, strategies, and procedures that were already expected to be employed for various attacks.

Dark web

The dark web is made up of content that isn't indexed by google search and it can only be accessible with specific software or authorization. The darknet is a portion of the internet that is only viewable to certain browsers either through specific network setups. The most popular and accessible forums for cybercriminals to share hacker resources are hacker forums. These discussion boards are used by hackers to post messages in threads about hacking tools, strategies, and harmful source code.

Dark Web Forums

On the dark web, anonymity takes precedence. This is where unauthorized hacking services can be sold. All of these adverts are false. Furthermore, some members take advantage of their anonymity to defraud others. As a result, you can't be certain that such services are reliable 100 percent of the time. A dark web forum is an environment where users can openly discuss topics such as drug smuggling, child exploitation, hacking, data dumps, racist as well as extremist content, and more. Cybercriminals use underground forums to discuss a variety of subjects ranging from operational security to server functionality. Users can choose from a variety of membership options, along with VIP, Premium, or Moderator. Because active forum users are frequently sought by criminal justice or intel agencies, monitoring them provides valuable information that can help law enforcement or security agencies catch criminals in the act.

2. Literature Survey

A. Exploring the Dark Web for Cyber Threat Intelligence using Machine Learning:

The major goal was to use machine learning to provide more precise threat detection from the darknet using the doc2vec tool as an "active defense" against cyber threats. The goal of this study is to use machine learning and doc2vec, a natural language processing approach, to extract postings from dark web forums that contain vital information in order to conduct preventive countermeasures. We also determine the characteristics of several dark web forums based on this conclusion. On the dark web, there are a plethora of forums. In these forums, malicious intruders share content data such as malware dealing and hacking strategies. We hope to extract essential posts from these forum postings using doc2vec and machine learning. In this study, we set ”posts related to malware offers” as critical posts [1].

B. Cyber Threat Discovery from Dark Web**:**

The major goal was to extract cybercrime intelligence from Darknet hacker forum postings and predict attack type using a Random Forest classifier, which has a higher accuracy rate. Manually monitoring and evaluating this data is difficult due to the large amount and unstructured nature of forum posts as well as other darknet data. On a darknet forum postings dataset, this article mixes advanced statistics and predictive analytics with machine learning to uncover useful cyber risk intelligence [2].

C. Dark-Web Cyber Threat Intelligence: From Data to Intelligence to Prediction:

The key objective was to investigate the evolution of cybersecurity threats, as well as how information must be retrieved, processed, and potentially used for forecast purposes in order to reduce their impact. As the title of this book implies, there has been a shift with how the deep web can be used to influence cyber threat intelligence. Simply said, the sensitive data must be gathered, processed, and maybe used for prediction, all of which are difficult tasks [3].

D. Predicting Cyber-Events by Leveraging Hacker Sentiment:

The project proposes a way for preventing cyber-events on both the surface and dark web that have the potential of being used as signals for attack predictions. We looked at over 400 thousand posts from over 100 hacking communities on either the surface and deep web from January 2016 to January 2018. Some forums have much higher predictive power than others, according to our findings. Sentiment-based models that use specific forums can be used to supplement state-of-the-art time-series models for predicting cyber-attacks weeks in advance [4].

2.1 Summary of Related Work

The study of various machine learning algorithms and artificial intelligence techniques are studied. Furthermore the fundamentals, objectives and scope of the project is presented. Literature survey has been done in review and survey papers which has led to the identification of various research gaps. The proposed system architecture has been presented along with the hardware and software specifications. Applications of this project are mentioned in the report.

The summary of methods used in literature is given in Table 1.

Table 1 Summary of literature survey

Literature	Methods used	Accuracy
Masashi Kadoguchi et al. 2019[1]	Sixgill,Doc2vec , ML techniques	90%
Azene Zenebe et al. 2019 [2]	ML techniques, NLP,Offline Explorer	78%
Paulo Shakarian et al. 2018[3]	ML techniques, Web crawlers, parsers	-
Ashok Deb et al. 2018[4]	Sentimental Analysis	37%

3. Proposed Work

We are proposing an approach to extract forums which include important information or intelligence and identify traits of each forum using methodologies.The dataset is extracted using different web crawlers from different websites. Dataset is extracted in the form of a csv file from the website azuresec.net. Data was wholly in the form of Russian language which was translated into english using different translators. Then the retrieved data is cleaned, normalized, parsed, processed, classified and labeled. To extract data into different categories based on the level of threat, Machine Learning Techniques were used. Final step is to test the model with the inputs and predict the cyber threats for the future. This will help us know the future emerging threats in cyberspace and take appropriate measures to avoid such malicious activities. We aim to extract critical data. The use of information from hacker communities such as the dark web has great promise in leading to a more threat-focused cybersecurity.

3.1 System Architecture

The system architecture is given in Figure 1. Each block is described in this Section.

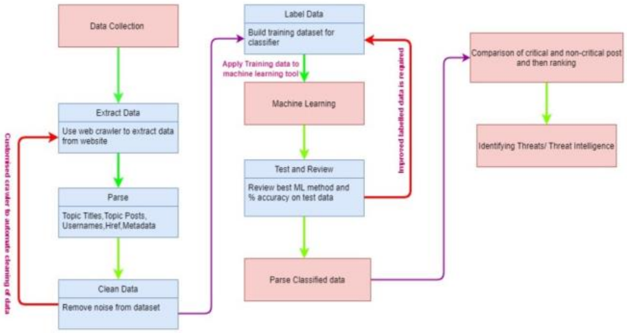
Data Collection
Extract Data
Use web crawler to extract data from website
Parse
Topic Titles,Topic Posts, Usernames,Href,Metadata
Clean Data
Remove noise from dataset
Customised crawler to automate cleaning of data
Label Data
Build training dataset for classifier
Apply Training data to machine learning tool
Machine Learning
Test and Review
Review best ML method and % accuracy on test data
Improved labelled data is required
Parse Classified data
Comparison of critical and non-critical post and then ranking
Identifying Threats/ Threat Intelligence

Fig. 1 Proposed system architecture

A. Data collection:The first part is to collect the data.Tor browser and Virtual Private Network is used to collect data from dark web.Tor, acronym for The Onion Router, is a free and open-source programme that allows users to communicate anonymously. It hides a user's location and activity from anyone doing network surveillance or network configuration by routing Internet traffic.Tor makes it increasingly challenging to track down a user's online activities. Tor's purpose is to safeguard its users' personal privacy, as well as their liberty and capacity to communicate in confidence, by preventing their Internet activity from being monitored.A virtual private network (VPN) connects a private network to a public network, allowing users to transmit and receive data as if their computers were directly linked to the private network. Increased functionality, security, and control of the private network are some of the advantages of using a VPN. It is often used for distant employees and gives access to services that are not available on the public network. A VPN is established by using dedicated circuits or tunneling technologies to build a virtual point-to-point connection across an existing network. A VPN that can be accessed through the public Internet can give some of the advantages of a wide-area network.

B.Preprocessing of data: The extracted data is then parsed. In the proposed method, we performed word tokenization, cleaning, word normalization, stemming, and stop-words processing as preprocessing steps. Word tokenization is a process of separating individual words. Cleaning refers to removing unnecessary characters such as numbers and parentheses in the text. Word normalization unifies upper/lower cases of each word.

The process of shortening a word to its word stem, which mainly refers to prefixes and suffixes or to the base of words known as a lemma, is known as stemming. Natural language understanding (NLU) and natural language processing (NLP) both benefit from stemming (NLP).NLP is interaction between humans and systems.NLP allows computers to comprehend natural language in the same way that humans do. Natural language processing employs artificial intelligence to accept real-world data, interpret it, and make sense of it in a way that a computer can comprehend, whether the language is spoken or written.

C.Labelling of data :Here, we receive the preprocessed data. This data is labeled according to severity of threats (such as medium threat ,high threat,low threat,no threat) and machine learning tools are used to this preprocessed data. Afterwards, this labeled data is tested and reviewed using best machine learning methods and algorithms. For the test and review of data, improved labeled data is required.

D.Machine Learning: First, the trained dataset is given to the machine learning model as the database server for authentication. The preprocessed data acts as input for the machine learning model. Naive Bayes and support vector machine algorithms are used for classification of data. Natural Language Processing is used to train the model about all the functionalities. It aims to teach machines how we humans utilize natural languages to communicate with one another.Most of the NLP techniques use various supervised and unsupervised machine learning algorithms for extracting valuable insights from the human language. A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points.The Support Vector Machine, or SVM, is a popular Supervised Learning tool for solving classification and regression problems. However, it is mostly used in Machine Learning to solve classification problems.The objective of the SVM method is to determine the best line or deciding boundary for classifying n-dimensional space into categories so that subsequent data points may be easily placed in the appropriate category. The ideal choice boundary is known as a hyperplane. SVM is used to choose the extreme vectors that help build the hyperplane.Support vectors are the extreme instances, and the method is called a Support Vector Machine.

E.Output Block: After preprocessing and labeling of data and successfully applying the algorithms the output is displayed with threat analysis.

3 Requirement Analysis

The implementation detail is given in this section.

3.1 Software

Naive Bayes and Support Vector Machine algorithm of the machine learning techniques is used to classify and identify the threats from the dataset. We are using python as our programming language and jupyter notebook as python IDE. Jupyter Notebook is a free, open-source web tool that lets you create and share documents with live code, equations, visualizations, and narrative prose. Data cleansing and processing, mathematical modeling, statistical modeling, data visualization, machine learning, and many more applications are all possible. Python is a high-level, general-purpose programming language that is interpreted. Python's design philosophy places a strong emphasis on programming languages, as seen by the frequent usage of indentation.Since Python uses minimal code, it becomes easier for the programmer to debug the error and at the same time, reduce the risk of language complexity and issues.

3.2 Hardware

Desktop/Laptop is used to display the final output report through a web application.

3.3 Dataset and Parameters

We used a dataset available from azuresec.net website which was in russian language and translated into english and also used datasets from kaggle and mendeley sites.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Prof. K.S. Charumathi For the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks to our Head of the Department Dr. Satishkumar Varma and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

REFERENCES

[1] Exploring the Dark Web for Cyber Threat Intelligence using Machine Learning, 2019, Graduate School of Information Security Yokohama, Japan.

[2] Cyber Threat Discovery from Dark Web, 2019, University of Maryland College Park, Bowie State University, Bowie, Maryland, Farmingdale State College, Farmingdale, NY.

[3] Dark-Web Cyber Threat Intelligence: From Data to Intelligence to Prediction, 2018 School of Computing, Informatics, and Decision Support Engineering, Arizona State University, USA.

[4] Predicting Cyber-Events by Leveraging Hacker Sentiment,2018, Information Sciences Institute, University of Southern California, Marina del Rey, USA.

[5] Framework for More Effective Dark Web Marketplace Investigations, 2018, Department of Business and Management, LUISS Guido Carli University, Viale Pola, Rome, Italy.