Darknet And Deepnet Mining For Proactive Cyber Security Threat Intelligence

Sasikala M[1], Abhinaya S[2], Sharmila R[3],Anitha M[4]
[1, 2, 3, 4]Dept of Computer Science and Engineering
[1, 2, 3, 4] Adhi College of Engineering and Technology, Kanchipuram– 631605.

Abstract*- In this project, we present an operational system for cyber threat intelligence gathering from various social platforms on the Internet particularly sites on the Darknet and Deepnet. We focus our attention to collecting information from hacker forum discussions and marketplaces offering products and services focusing on malicious hacking. We have developed an operational system for obtaining information from these sites for the purposes of identifying emerging cyber threats. Currently, this system collects on average 305 high-quality cyber threat warnings each week. These threat warnings include information on newly developed malware and exploits that have not yet been deployed in a cyber-attack. This provides a significant service to cyber defenders. The system is significantly augmented through the use of various data mining and machine learning techniques. With the use of machine learning models, we are able to recall 92% of products in marketplaces and 80% of discussions on forums relating to malicious hacking with high precision. We perform preliminary analysis on the data collected, demonstrating its application to aid a security expert for better threat analysis.*

Keywords*-* cybersecurity, cyber threat intelligence, dark web, deep learning, Grad-CAM,CNN multimodal classification, onion services.

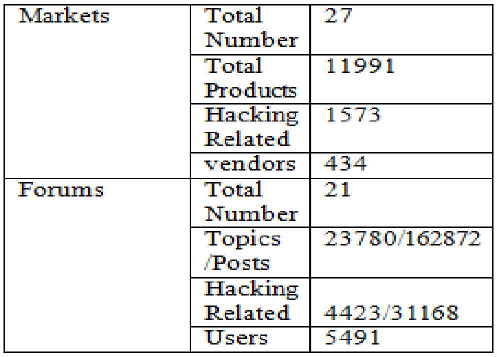
I. INTRODUCTION

GENERAL

Pre-reconnaissance cyber threat intelligence refers to information gathered before a malicious party interacts with the defended computer system. An example demonstrating the importance of cyber threat intelligence is shown in Table 1.

A Microsoft Windows vulnerability was identified in Feb. 2015. The release of the vulnerability was essentially Microsoft warning its customers of a security flaw. Note that at this time, there was no publicly known method to leverage this flaw in a cyber-attack (i.e. an available exploit). However, about a month later an exploit was found to be on sale in darknet market. It was not until July when FireEye, a major cybersecurity firm, identified that the Dyre Banking Trojan designed to steal credit cards exploited this vulnerability - the first time an exploit was reported. This vignette demonstrates how threat warnings gathered from the darknet can provide valuable information for security professionals. The average global exposure of the Dyre Banking Trojan was 57.3% along with another banking malware Dridex1. It means that nearly 6 out of 10 organizations in the world were affected, and this is a significantly high number on a global level. In this project, we examine how such intelligence can be gathered and analyzed from various social platforms on the Internet particularly sites on the darknet and deepnet. In doing so, we encounter several problems that we addressed with various data mining techniques. Our current system is operational and actively collecting approximately 305 cyber threats each week. Table 1 shows the current database statistics. It shows the total data collected and the data related to malicious hacking. The vendor and user statistics cited only consider those individuals associated in the discussion or sale of malicious hacking related material, as identified by the system. The data is collected from two sources on the darknet/deepnet: markets and forums.

Table 1: current database status



Markets	Total Number	27
	Total Products	11991
	Hacking Related	1573
	vendors	434
Forums	Total Number	21
	Topics /Posts	23780/162872
	Hacking Related	4423/31168
	Users	5491

II. RELATED WORKS

DARKNET AND DEEPNET SITES

Widely used for underground communication, “The Onion Router” (Tor) is free software dedicated to protect the privacy of its users by obs\curing traffic analysis as a form of network surveillance. The network traffic in Tor is guided through a number of volunteer operated servers(also called “nodes”). Each node of the network encrypts the information it blindly passes on neither registering where the traffic came from nor where it is headed, disallowing any tracking. Effectively, this allows not only for anonymized browsing (the IP-address revealed will only be that of the last node), but also for circumvention of censorship2. Here, we will use “darknet” to denote the anonymous communication provided by crypto-networks like “Tor”, which stands in contrast to “deepnet” which commonly refers to websites hosted on the open portion of the Internet , but not indexed by search engines.

MACHINE LEARNING APPROACHES

In this work, we leverage a combination of supervised and semi-supervised methods. Supervised methods include the well-known classification techniques of Naive Bayes (NB),random forest (RF), support vector machine (SVM) and logistic regression (LOG-REG).However,supervisedtechniquesrequiredlabeleddata,andthisisexpensiveandoftenrequires expert knowledge. Semi-supervised approaches work with limited labeled data by leveraging information from unlabeled data. We discuss popular semi-supervised approaches used in this work. We perform a grid search to find optimal parameters for the learning techniques

TOR NETWORK

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

III. 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 (TFIDF). 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.

Wang et.al. [1] present probabilistic model for relevant web page selection. Crawler employed the TF-IDF algorithm to quote the feature of page content and Bayes classifier to evaluate page rank.

Whereas, Saleh et al. [2] present a probabilistic domain distiller for a focused crawler. Domain distiller combines SVM, naïve Bayes, and genetic algorithm (GA) and present optimized instance of probabilistic classifier, i.e., optimized Naïve Bayes (ONB) classifier. Where initially, GA optimized the marginal distance of different class labels for vector space, and SVM alienates the outlier's keywords. NB scrutinized the ambiguous nature of the outlier domain keyword and finally proposed a disambiguation model for classification.

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- of Illegal Forum Activities inside the Dark Net” . 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.

Zhao et al. [7] presents deep web interface for harvesting Smart Crawler. Smart Crawler use search engine for extra cting center page and prioritize highly relevant URLs. Simultaneously, Smart Crawler employed adaptive link-ranking algorithm for extracting most relevant links.

Whereas, Xiaojun Liu et al. [8] presents Sina Weibo based web crawler for extractin Chinese citizen sentiment about green building. This approach applied text mining, ontology, and keywords search for dictionary-based sentiment crawler. Apart from search domain focused crawler also being employed for vulnerability and security analysis, Kim and Pant [9]-focused crawler for Detection of Malicious Web Page with the help of machine learning technique and audience Demography. Malicious web page focused crawler employed Naive Bayes, SVM, Logistic Regression with Statistical Analysis of web content.

Khalil et al. [10] present a multithreaded duplicate content detection web crawler as RCrawler that useful for web crawling, scraping, and link analysis. RCrawler extract URL, Page Content, and depth level feature and employed similarity hash function algorithm for parallel web crawling and scraping.

Janis Dalins et al.[11] presents focused crawler for extracting and blocking dark web for child pornography. Focused crawler for dark web used labeled and page content as suspicious text feature.

Pedro Ivo et al. [12] present a focus crawler for analyzing business threats and opportunities analysis by integrating pattern recognition, ontology, and weak signal monitoring. Anchor text for Business threats is extracted by using Part of Speech tagging and SVO Typology.

Rong Wang et al. [13] present a focused crawler for blacklisting malicious web page. The focused crawler identifies malicious URLs by applying decision tree-based machine learning techniques over correlation and DOM tree-based feature.

Whereas Harry T Yani et al.[14] present distributed Focused crawler for tracking cyber attack, which is a multi-thread web crawler that use the optimal number of threads to maximize the download speed.

Fayyad et al. [15] have presented an analysis on KDD process, Knowledge discovery in databases (KDD) has been characterized as the non-trivial procedure of distinguishing substantial, novel, possibly valuable, and at last justifiable information from the data.

Apte [16] described Data Mining, Data Mining is a process by which precise and beforehand obscure information can be released from a large amount of data in a form that can be accepted, influenced, and employed for enhancing decision-making procedures.

Q. Luo [17] Data mining or knowledge discovery is the way toward dissecting data from different perceptions and briefing it into valuable information that can be utilized to recognize risks relevant to a particular project.

Bennouas et al. [18] proposed a random Web crawl model. That deals with the hyperlink structure, whose vertices are the pages and whose edges are the hyper textual links. Its model the simpler web graph crawling process instead of the page writing procedure.

IV. SYSTEM STUDY

EXISTING SYSTEM

- Classification, and consequently filtering of text content plays an important role. The rapid growth of internet recourses creates the necessity for accurate classification, which is supposed to help determine their ever-changing nature.
- In the case of hidden web-resources, classification helps to identify the degree of threat, and therefore to apply blocking measures in a timely manner.
- This Existing research looks at the possibility of onion resource classification by assigning categories to the text blocks contained there in.
- It has been developed an algorithm to collect the elements of a training sample, which allows us to extract all the relative URLs belonging to the absolute URL.
- After processing the obtained relative URLs, the program receives blocks of text content which are sorted later by the number of characters and written to a text file in the format: “category;source;text”.
- Passing text blocks that contain less than a certain number of characters helps to avoid getting elements into a training sample which in most cases do not carry any “semantic load”. The upper limit of the number of characters in a string is similarly defined.

DISADVANTAGES

- In the KNN method, the smallest values of distances were selected from the input text to the learning sample elements. In this case, the amount of information about the block of text becomes less than expected.
- It is possible to achieve a lot of class categorization by introducing a standard deviation among the admissible values and classify using the distances belonging to the given interval.

PROPOSED SYSTEM

The image and text modalities contain the main features required for the detection and
classification of an onion service.

- A Convolutional Neural Network with Gradient-weighted Class Activation Mapping (Grad-CAM) and a pre-trained word embedding with Bahdanau additive attention are the core capabilities of this approach that classify and contextualize the representative features of an onion service.
- The proposed multimodal classification approach based on explainable deep learning consists of two pathways for the two modalities of images and text. It is illustrated in Fig.3.3. Each pathway has a learning phase followed by an explain ability phase and then merged for the final phase of multilabel classification. Unavailability of enough training data is a common problem in certain application domains including onion service classification.

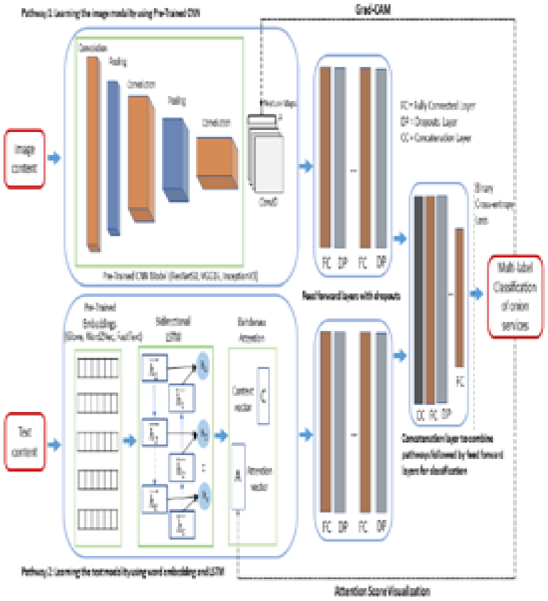

Pathway 1: Learning the image modality using Pre-Trained CNN
Grad-CAM
Convolution
Pooling
Convolution
Pooling
Convolution
Feature Maps
Image content
FC = Fully Connected Layer
DP = Dropout Layer
CC = Concatenation Layer
FC DP FC DP
Pre-Trained CNN Model (ResNet50, VGG16, InceptionV3)
Feed forward layers with dropouts
Binary Cross-entropy Loss
Multi-label Classification of onion services
Pre-Trained Embeddings (Glove, Word2Vec, FastText)
Bidirectional LSTM
Bahdanau Attention
Context vector
Attention vector
Text content
CC FC DP
FC
Concatenation layer to combine pathways followed by feed forward layers for classification
Pathway 2: Learning the text modality using word embedding and LSTM
Attention Score Visualization

Figure 1 .Proposed System

- The need for a large amount of training data is one of the major drawbacks in deep learning applications. But the transfer learning has emerged as a solution to this problem which uses a knowledge transferring approach where it uses knowledge gathered from one domain to solve a problem in another domain.
- The proposed modular AI component which integrates the multimodal deep learning architecture can efficiently categorizes the onion services in the Tor communication network to provide enhanced threat intelligence capabilities.
- Multimodal classification learns to classify inputs from multiple modalities, instead of a single modality.
- This has been effectively used in many domains, such as image captioning, sentence matching in images, and speech recognition .

V. SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

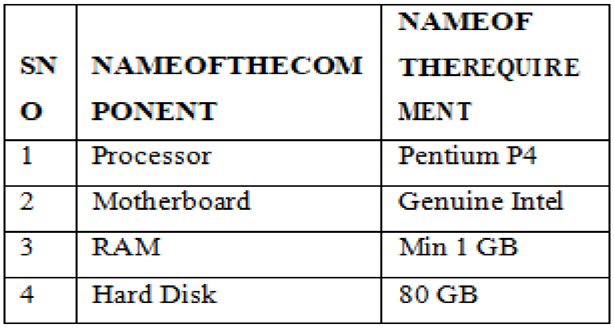
SOCIAL FEASIBILITY

*The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by thesystem, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

VI. SYSTEM SPECIFICATION

HARDWARE SPECIFICATION

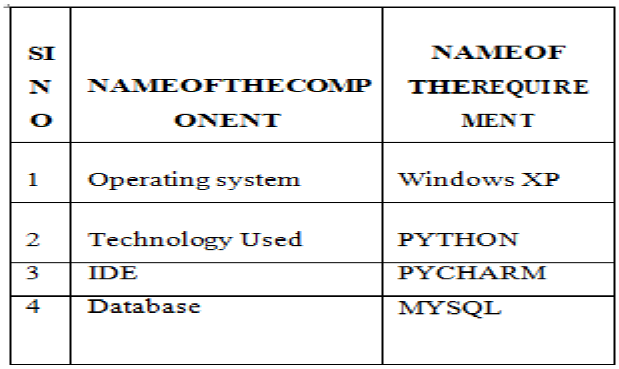
Table 2:Hardware specification



SN O	NAMEOFTHECOM PONENT	NAMEOF THEREQUIRE MENT
1	Processor	Pentium P4
2	Motherboard	Genuine Intel
3	RAM	Min 1 GB
4	Hard Disk	80 GB

SOFTWARE SPECIFICATION

Table 3:Software Specification



SI N O	NAMEOFTHECOMP ONENT	NAMEOF THEREQUIRE MENT
1	Operating system	Windows XP
2	Technology Used	PYTHON
3	IDE	PYCHARM
4	Database	MYSQL

SOFTWARE DESCRIPTION

Python is a very popular general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python is dynamically-typed and garbage-collected programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License Python is a MUST for students and working professionals to become a great Software Engineer specially when they are working in Web Development Domain. I will list down some of the key advantages of learning Python:

Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

Python is Interactive − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

CHARACTERITISTICS OF PYTHON

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

APPLICATIONS OF PYTHON

- Easy-to-learn − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- Easy-to-read − Python code is more clearly defined and visible to the eyes.
- Easy-to-maintain − Python's source code is fairly easy-to-maintain.
- A broad standard library − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.

PYTHON –FUNCTIONS

A function is a block of organized, reusable code that is used to perform a single, related action. Functions provide better modularity for your application and a high degree of code reusing.As you already know, Python gives you many built-in functions like print(), etc. but you can also create your own functions. These functions are called user-defined functions.

SYNTAX

```
def functionname( parameters ):
                    "function_docstring"
function_suite
                      return [expression]
```

By default, parameters have a positional behavior and you need to inform them in the same order that they were defined.

MACHINE LEARNING WITH PYTHON

Machine Learning (ML) is basically that field of computer science with the help of which computer systems can provide sense to data in much the same way as human beings do. In simple words, ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method. The key focus of ML is to allow computer systems to learn from experience without being explicitly programmed or human intervention.

NEED FOR MACHINE LEARNING

Human beings, at this moment, are the most intelligent and advanced species on earth because theycan think, evaluate and solve complex problems. On the other side, AI is still in its initial stage and haven't surpassed human intelligence in many aspects. Then the question is that what is the need to make machine learn? The most suitable reason for doing this is, "to make decisions, based on data, with efficiency and scale".

Lately, organizations are investing heavily in newer technologies like Artificial Intelligence, Machine Learning and Deep Learning to get the key information from data to perform several real-world tasks and solve problems. We can call it data-driven decisions taken by machines, particularly to automate the process. These data-driven decisions can be used, instead of using programming logic, in the problems that cannot be programmed inherently. The fact is that we can't do without human intelligence, but other aspect is that we all need to solve real-world problems with efficiency at a huge scale. That is why the need for machine learning arises.

VII. SYSTEM DESIGN

SYSTEM ARCHITECTURE

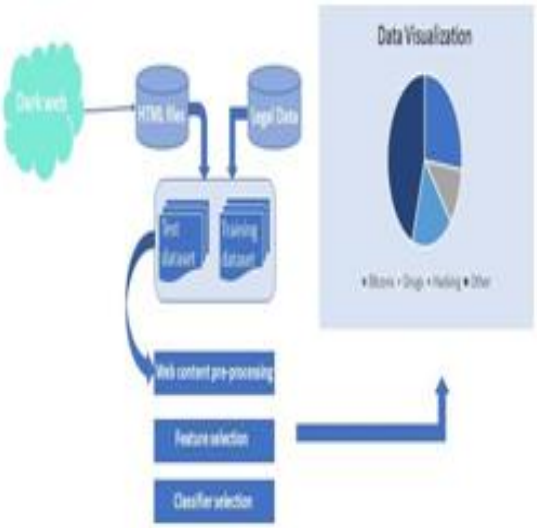


Figure 2 .System Architecture

The architecture is divided in three main parts– Collecting relevant laws and regulation (Training Data), Extracting Dark Web content (Testing Data), and Classification of illegal activities. The system Architecture is shown. The system 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. Afetr 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 compare the accuracy and choose the best algorithm for classification of illegal activities on DarkWeb forums.

ACTIVITY DIAGRAM

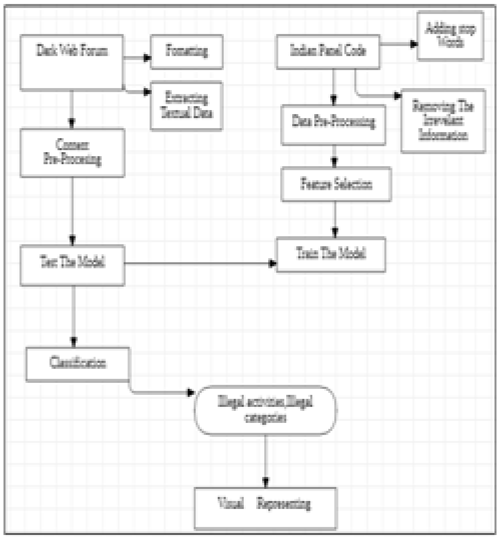


Figure 3.Activity Diagram

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 webpages with .onion extension.

USE CASE DIAGRAM

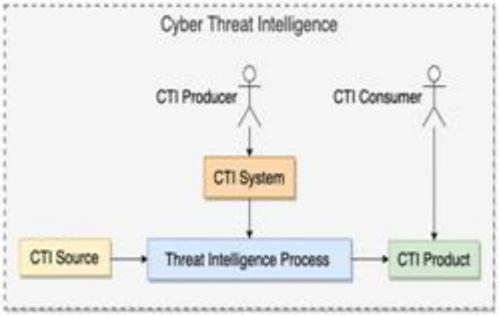


Figure 4. Use Case Diagram

MODULES

1. Web crawler module
2. Parser module
3. Classifier module
4. Forums module

WEB CRAWLER MODULE

The crawler is a program designed to traverse the website and retrieve HTML documents. Topic based crawlers have been used for focused crawling where only web pages of interest are retrieved.

More recently, focused crawling was employed to collect forum discussions from dark net. We have designed separate crawlers for different platforms (markets/forums) identify by experts due to the structural difference and access control measures for each platform

PARSER MODULE

- We implemented a parser to extract specific information from marketplaces (regarding sale of malware/exploits) and hacker forums (discussion regarding services and threats).
- This well-structured information is stored in a relational database. We maintain two databases, one for marketplaces and the other for forums. Like the crawler, each platform has its own parser.
- The parser also communicates with the crawler from time to time for collection of temporal data. The parser communicates a list of relevant web pages to the crawler, which are re-crawled to get time-varying data.

FORUMS MODULE

Forums are user-oriented platforms that have the sole purpose of enabling communication. It provides the opportunity for the emergence of a community of like-minded individuals - regardless of their geophysical location. Administrators set up Darknet forums with communication safety for their members in mind. While structure and organization of Darknethosted forums might be very similar to more familiar web-forums, the topics and concerns of the users vary distinctly. Forums addressing malicious hackers feature discussions on programming, hacking, and cyber-security. Threads are dedicated to security concerns like privacy and online safety - topics which plug back into and determine the structures and usage of the platforms.

TESTING

Testing is the process of detecting errors. Testing plays a critical role in assuring quality and ensuring the reliability of software. The results of testing are used later on during maintenance also.

TESTING OBJECTIVES

The main objective of testing is to uncover a host of errors, systematically and with minimum effort and time. Testing is a process of executing a program with the intent of finding an error. A good test case is one that has a high probability of finding error, if it exists The tests are inadequate to detect possibly present errors The software more or less confirms to the quality and reliable standards.

TESTING LEVELS

System testing is stage of implementation which is aimed at ensuring that the system works accurately and efficient before live operation commences. Testing is vital the success of the system. System testing makes a logical assumption that if all the parts of the system are correct, the goal will be successfully achieved.

1. UNIT TESTING
2. INTEGRATION TESTING
3. ACCEPTANCE TESTING

VIII. CONCLUSION

In this project, we implement a system for intelligence gathering related to malicious hacking.

Our system is currently operational. We are in the process of transitioning this system to a commercial partner. We consider social platforms on dark net and deep net for data collection.

We address various design challenges to develop a focused crawler using data mining and machine learning techniques. Here we develop a novel multimodal classification approach that utilizes explainable deep learning to classify onion services based on the image and text content of each site. This approach consists of two learning pathways,

1. image Modality based on CNN with explain ability using Grad-CAM
2. text modality based on pre-trained word embedding and BiLSTMs with Bahdanauadditive attention mechanism for explain ability.

We evaluated this approach on a state-of the-art onion services dataset curated by CIRCL, containing more than8000 instances of onion services across 51 categories. The results of this experiment confirm the value and effectiveness of the proposed multimodal classification approach in enabling proactive CTI decision-making with interpretable and explain able out comes to monitor and detect the cybersecurity threats originating from dark web onion services.

REFERENCES

[1] J.H.Li,‘‘Cyber security meets artificial intelligence: A survey,’’ Frontiers Inf. Technol. Electron.Eng., vol. 19, no. 12, pp. 1462–1474, Dec. 2018

[2] Chetry and U. Sharma, ‘‘Dark web activity on Tor—Investigation challenges and retrieval of memory artifacts,’’ in Proc. Int. Conf. Innov. Comput. Commun., 2021, pp. 953–964.

[3] P.Burda, C.Boot, and L.Allodi,‘‘Characterizingthe redundancyDarkWeb.Onionservices,’’inProc.14thInt. Conf.Availability,Rel.Secur.,Aug. 2019, pp. 1–10, doi: 10.1145/3339252.3339273.

[4] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, ‘‘Accessing the deep web,’’ Commun. ACM,vol. 50, no. 5, pp. 94–101, May 2007, doi: 10.1145/1230819.1241670.

[5] S. Ghosh, A. Das, P. Porras, V. Yegneswaran, and A. Gehani, ‘‘Automated categorization of onion sites for analyzing the darkweb ecosystem,’’ in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2017, pp. 1793–1802, doi: 10.1145/3097983.3098193

[6] J. Saleem, R. Islam, and M. A. Kabir, ‘‘The anonymity of the dark web: A survey,’’ IEEE Access, v ol. 10, pp. 33628–33660, 2022, doi: 10.1109/access.2022.3161547.

[7] M. Ebrahimi, M. Surdeanu, S. Samtani, and H. Chen, ‘‘Detecting cyber threats in non-english dark net markets: A cross-lingual transfer learning approach,’’ in Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI), Nov. 2018, pp. 85–90, doi: 10.1109/ISI.2018.8587404

[8] M. Mirea, V. Wang, and J. Jung, ‘‘The not so dark side of the darknet: A qualitative study,’’ Secur. J., vol. 32, no. 2, pp. 102–118, Jun. 2019, doi: 10.1057/s41284-018-0150-5.

[9] J.Bergmanand O.B.Popov,‘‘The digital detective’s discourse—A tool set for forensically sound collaborative dark web content annotation and collection,’’ J. Digit. Forensics, Secur. Law, vol. 17, no. 1, p. 5, 2022, doi: 10.15394/jdfsl.2022.1740.

[10] C. Wagner, Dulaunoy, G. Wagener, and A. Iklody, ‘‘MISP: The design and implementation of a collaborative threat intelligence sharing platform,’’ in Proc. ACM Workshop Inf. Sharing Collaborative Secur., Oct. 2016, pp. 49–56, doi: 10.1145/2994539.2994542.

[11] L. Dandurand and O. S. Serrano, ‘‘owards improved cyber security information sharing: Requirements for a cyber security data exchange and collaboration infrastructure (CDXI),’’ in Proc. 5th Int. Conf. Cyber Confl. (CyCon), Jun. 2013, pp. 14–29, 2013.

[12] U. Noor, Z. Rashid, and A. Rauf, ‘‘A survey of automatic deep web classification techniques,’’ Int. J. Comput. Appl., vol. 19, no. 6, pp. 43–50, Apr. 2011, doi: 10.5120/2362-3099.

[13] OnionServices.TorMetrics.Accessed:Nov.21,2021.[Online].Available:https://metrics.torproj-t.org/hidserv-dir-onions-seen.html

[14] J.Ngiam,A.Khosla,M.Kim,J.Nam,H.Lee,andA.Y.Ng,‘‘Multimodal deep learning,’’ in Proc. 28th Int. Conf. Mach. Learn. (ICML), 2011, pp. 689–696.

[15] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, ‘‘Explain images with multimodal recurrent neural networks,’’ ArXiv, vol. abs/1410.1090, 2014.