International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Snowball-Miner: Integration of Deep Learning for Extraction of Cyber Threat Intelligence from Dark Web

[1]Abir Dutta, [2]Bharat Bhushan, [3]Shri Kant
[1]PhD. Scholar
Deptt. of Computer Sc. and Engg,
Sharda University, Greater. Noida
e-mail: abir_wbsetcl@yahoo.com
[2]Deptt. of Computer Sc. and Engg,
Sharda University, Greater. Noida
e-mail: bharat.bhushan@sharda.ac.in
[3]Center of Cyber Security and Cryptology
Deptt. of Computer Sc. and Engg,
Sharda University, Greater. Noida
e-mail: shri.kant@sharda.ac.in

Abstract— In Cyber threat intelligence is a crucial component in defending against cybersecurity threats. Cyber security dark web, security Blogs, Hackers' community, news forums, Open-Source Intelligence (OSINT) are known as the harbor of illicit activities and serve as a breeding ground for cybercriminals. Extracting actionable intelligence from the dark web is challenging due to its anonymous and encrypted nature. State-of-art work proposed machine learning and deep learning approach to aggregate the dark web for cyber threat intelligence from data present in the dark web. This paper proposes, a novel approach utilizing Snowball-Miner for cyber threat intelligence discovery from the dark web. The model is trained on a diverse dataset consisting of dark web forums, hidden .onion based marketplaces and other underground platforms using Snowball-crawler. However, we have employed hybrid convolutional model CNN-LSTM and CNN-GRU adopting doc2vec word embedding to classify into four domains viz Energy Sector, Finance, Illicit Activities and illegal Services. From our experiment it emerged that, CNN-LSTM outperforms as 96.37% for classification of domain specific threat documents. Furthermore, after data preparation we implemented NLP technique and extracted the domain specific Indicator of Compromise (IoCs) using RegEx parser and Subject, Object and Verb (SOV) semantics dependency analysis. Finally, we have integrated IoCs and Threat keywords with respective domains to generate domain specific threat intelligence which enhance the quality of the domain specific CTI based on R-dimension (Relevance).

Keywords- Power Sector; Cyber Threat Intelligence; Natural Language Processing; Web crawler; Machine Learning; Naïve Bayes

I. INTRODUCTION

The sophistication of cyber-attack methods has increased recently. The majority of attacks in the past were meant to cause personal trouble, but now, organized [1]. Attacks with the intention of committing money laundering are rising. Additionally, in the past, the majority of cyber-attacks goals were indifferent, but today, a certain target is repeatedly assaulted with a defined goal, have become more common. The rapidly expanding number of novel varieties of malware cannot be detected by traditional methods of malware detection that depend on signatures [2]. Due to these recent events, it is challenging to entirely stop all attacks, even when defense measures are adopted against cyber-attacks. Even if 99.9% of businesses with a sizable network were successful in thwarting assaults, the remaining 0.10% of failures would be catastrophic for the entire business. In earlier days, attackers are used to target the Information Technology Infrastructure and the associated data center of the enterprises by means of malware attack, ransomware for stealing data, financial benefits and defacement of the organization [3]. But in recent days hackers are stealthier for the intension of a potential disrupt or destroy of critical physical infrastructures [4]. Using modern security tools and artefacts, malicious intender is more insidious in the network solution of the Critical Information Infrastructure (CII) like, power production units, HV sub-station, Fiber optics and MPLS-VPN communication, VHF and UHF Transmission Lines, distributed networks etc. [5-6]. Cyber criminals targeted the backdoor of the legitimate software and breach the applications to find out the vulnerabilities, which have been used to find out the web facing components of the connected networks to exploit Industrial Control Systems software and SCADA system [7].

Hence, for Energy sector usage of AI and machine learning can be a game changer for detecting and preventing ever-growing cyber security threats. Also, these technologies can provide Energy sectors many energy-specific solutions. Customer Engagement/ Energy conservation – With the use of AI and machine learning in energy sector, energy companies can provide consumers their need base customized information through data analytics [8]. Simultaneously, AI and ML can automatically detect illegal tapping of energy from the grid and intentional misrepresentation of energy data or energy usage and flag them for energy companies and enables them to reduce energy waste, protect their assets, and reduce expenditure. Smart Grids and Grid Security – With the usage of smart grids, energy data can be collected from very granular level, which helps to develop energy efficient project plans [9].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Also, this strategy allows energy companies to monitor energy flow and real-time energy consumption. This leads to controlled energy consumption through automated demand response systems resulting in energy savings [10]. AI and machine learning can be used to improve a complex and vulnerable system of energy grids by preventing cyber-attacks by using data analytics to identify patterns in energy data and by providing a quick response to it.

In order to handle this issue, it is necessary to anticipate cyber-attacks and take the required precautions in advance, and using intelligence is crucial to making this feasible. We anticipate that much cognitive ability, exists in cyberspace assaults on the dark web or in particular forums [11-12]. It is anticipated that employing threat intelligence will enable early attack detection and active defense. However, various research efforts are being carried out to improve this process. This study uses the natural language processing method doc2vec to more effectively extract threat intelligence. From forum discussions on the dark web, we extract the material that requires aggressive action (hence referred to as "critical posts"). knowledge on virus trading and hacking methods on the dark web, as we will cover later [13]. Our goal is to effectively extract these crucial postings from the dark web forums using doc2vec and machine learning. Depending on what the subjects are, many things are discussed in critical postings. We use forum posts on encouragement of cyber-attacks as examples. In this study, we concentrate on gathering forum posts linked to malware offers to show the efficacy of our suggested strategy. The goal of this project is to develop a technique for automatically extracting significant postings in order to increase the effectiveness of security analysis that uses intelligence [14]. The movement to actively defend against cyber-attacks by using intelligence is currently most prevalent in the industrial world, and some of these efforts have produced impressive outcomes. However, the Dark Web contains more than just posts that can be exploited as intelligence [15]. Role of Threat Intelligence in Energy Sector is depicted in Figure 1. Given that timeliness is a requirement for intelligence, the security analyst's intelligence will be timely and effective if it can successfully extract a key post from the black web, which is composed of both helpful and useless content.

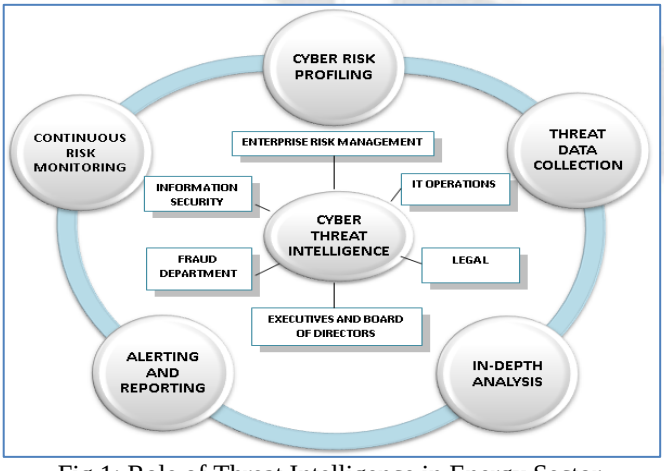

Fig 1: Role of Threat Intelligence in Energy Sector

Following research queries are attended in our study: **RQ1:** What are the substantial sources of dark web, that contain threat intelligence along with their domain? **RQ2:** How snowball sampling approach is effective to search new Drakweb links that contains threat related documents? **RQ3:** Which Machine learning and Deep learning classifiers are best fit for classifying the domain wise threat documents that are collected from dark web? RQ4: How to build a domain-oriented threat intelligence ontology?

Our contribution to this research is to extract threat intelligence in the form of IoCs and threat related keywords from various sources available in dark web and integrate this intelligence to the specific domains which are as follows-

- Snowball-Crawler which employs a breadth-search approach to scrap the dark web pages having the extension.onion to gather the threat related documents.
- Collected threat documents are classified into four specific domains such as Energy Sector, Finance, Illicit Activities and illegal Services. In order to achieve this, we have designed machine learning models like NB, LR and XGB. Moreover, we have implemented deep learning hybrid convolutional model like CNN-GRU and CNN-LSTM to classify the domains into aforementioned four domains. Types. At the same time, we have tested the accuracy of all the model for predicting the categorization.
- Furthermore, implementing NLP we have prepared our dataset and from these threat documents we have extracted the IoCs and threat keywords using RegEx parser and Subject-Object-Verb (SVO) combination approach respectively.
- Thereafter, we have integrated the extracted threat intelligence to the specific domain to generate domain-oriented CTI and we focused to generate Energy sector specific CTI. Finally, we have implemented RT dimension (Relevance and Timeliness) for evaluation of quality of extracted CTI

Rest of the paper is organized into six segments such as: in section II we will discusses about the state-of-the-art of cyber threat intelligence in various domain. Whereas, Section III describe the overview of the framework which facilitates the extraction of the IoCs from multiple data sources. Section IV illustrates the methodology of our experiment using Natural language processing and Machine Learning approach. Section V gives us the glimpse of the outcome of our experiment along with concerned discussion and lastly, Section VI specify the conclusion of our research work.

II. Related Work

Research on the dark net market places is carried out in different study. Such market places are the rich sources of cyber threat intelligence. Numerous research has been carried out in the domain of darknet resources. In this section, the most recent advances in ML and DL techniques, as well as other work relevant to frameworks for generation of domain wise cyber threat intelligence, will be covered.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

The state-of-art study was conducted by Jun Zhao et al. [16] which was focused on security resources of surface web. Initially in order to determine the intended domain of CTIs, the TIMiner for CTI mining framework, an effective domain recognizer based on convolutional neural network, is constructed. A method for extracting indicators of compromise (IOC) utilizing word embedding and structural dependencies is provided, and a categorized CTI is created by integrating the retrieved IOC and its domain tag. Huixia Zhang et al. [17], presented the EX-Action framework, which uses NLP to locate threat events and a multimodal approach to learning to identify activities. The knowledge integrity of the extracted action obtained by EXAction is also evaluated using a metric and CTI reports that had complicated phrase structures reflects that EX-Action performs superior to other cutting-edge action extraction techniques regarding performance evaluation metrices. To aid security researchers in their equipment prioritization and risk mitigation attempts, Sagar Samtani et al. [18] suggested a device vulnerability severity meter (DVSM) that includes the exploitable release timestamp and vulnerability intensity. Two CTI instance such as openly available systems from the most prominent hospitals and global Supervisory Control and Data Acquisition (SCADA) systems—were used by the author to further illustrate the real-world importance of the EVA-DSSM and DVSM. Harm Griffioen et al. [19] assess the quality of multiple open-source cyber threat intelligence feeds over a period of times. According to their investigation, the majority of indicators are operational for a minimum tenure prior to listing. Additionally, they have discovered that many lists include prejudices towards particular nations and demonstrate how banning IP addresses on a list can result in a significant amount of collateral harm.

A multimodal classification strategy employing comprehensible deep learning that categorizes onion services depending on the visual and textual content of each site was proposed by Harsha Moraliyage et al. [20]. The main components of this technique that categories and contextualize the representative aspects of an onion service are a neural network based on convolution employing Gradient-weighted Class Activation integration and a pre-trained word embedding. Kate Connolly et al. [21], recommended open-access release of an analysis of data on cybersecurity-related postings on dark web markets for many types of dangerous digital commodities available on online marketplaces, together with their pricing, steadfast merchants, reviews, and other rudimentary details about their storefronts, where author searched for webpages that offered illegally obtained digital items for several years while compiling information on the commodities that were offered using adaptable Selenium WebDriver to explore the web sites and gather data due to the marketplaces' diverse and sophisticated levels of protection. Varsha Varghese et al., [22] proposed an architecture using crawling and scraping, together with an anonymous intelligence toolkit, to search and gather information from the dark web forums. The acquired data is then fed into a cutting-edge NLP model that uses NER technology to derive useful threat intelligence. According to the testing findings, the model could more effectively and accurately identify organizations, HackerIds, software, tools along with other things in the conversations of dark web forums. Apurv Singh Gautam et al. [23] proposed a model that uses deep learning and machine learning techniques to automatically categorized hacker forum information into predetermined groups and create dynamic illustrations that allow CTI professionals to explore gathered data for diligent and appropriate CTI. According to the study's findings, the deep learning model RNN GRU produces the best classification outcomes of all the models.

In order to find useful cyber threat intelligence, Azene Zenebe et al. [24] suggested both descriptive and predicative analytics utilizing machine learning on the forum posts dataset from darknet. We employed the WEKA machine learning technology and IBM Watson Analytics. The data was analyzed using Watson Analytics to reveal patterns and connections. In order to categories the types of vulnerabilities that hackers are targeting from the form posts, WEKA provided machine learning models. The findings demonstrated the prevalence of Cryptor, password crackers, RATs (Remote Administration Tools), buffer overflow exploit tools, and keylogger system exploit tools in the darknet as well as the regular attendance of significant writers in forums. Machine learning also aids in the development of exploit type classifiers. Compared to the Random Forest classifier, accuracy was greater. A DT-BERT-BiLSTM-CRF based model on the dictionary template was suggested by Xuren Wang et al. [25]. which is a prior training approach effectively utilizing the corpus of context-dependent a semantic data and reduces inconsistency in the threat intelligence object detection process. The accuracy of NER in the threat intelligence area has been enhanced by creating a lexical template of threat intelligence entities. In this research author used remote supervision techniques to create the association mining set of data and the NR-RL-PCNN-ATT model, incorporates reinforcement learning and the attention mechanism into conventional neural networks to reduce the variability in the interpretation data. Ghaith Husari et al. [26], developed a text mining approach that combines improved strategies of information retrieval and Natural Language Processing in order to retrieve threat measures based on a semantic connection. They also proposed a risk measures an ontological framework which is adequate for comprehending the specifications and implications of malicious behaviour. Every potential risk activity is mapped to the kill chain phases, proper methods and tactics and translates any threat sharing standards like STIX 2.1. Kanti Singh Sangher et al. [27], uses the Agora dataset obtained from the DarkNet market store, to employ deep learning topologies based on categorization techniques utilizing the pre-trained word-encoding visualizations to detect illegal actions connected to online crime on Dark Web forums. Author annotated the data using a painstakingly planned human labelling method, accounting for every characteristic that implied the context of the data. To analyses Dark Web forums and detect cybercrimes by law enforcement organizations, the author presented a BERT-based categorization model that can open up possibilities for the development of advanced systems as per the needs.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

III. Overview of the Architecture

In our experiment we have adopted Snowball-Miner which comprises of Six phases such as-

- *Threat data feed gathering:* Dark web is a significant sources of threat related data and in order to collect relevant threat data we have to adopt additional security measures. As the dark web comprises of lethal entities such as infected script, malware, advance persistent threat, ransomware and many more. To safely collection of threat document we have designed snowball-crawler which is a python-based script and employ in a virtual environment in collaboration of VPN and tor browser [28]. We have monitored around 62 distributed onion sites to collect our intended threat documents. To achieve this, we have designed 48 nos of individual TI crawlers which monitor such pages and collect the relevant threat related resources.
- *Data Preparation and Natural Language Processing:* Threat Posts collected from different. onion pages contain irrelevant data, advertisement and other information which makes the dataset noisier [29]. To obtain higher degree of accuracy in respect of the desired outcome, data sanitization is a crucial aspect. We have implemented few data cleansing techniques such as stop-words removal, duplicate removal, stemming, normalization and tokenization.
- *Word Embedding:* In order to build any ML and DL model threat documents are needed to be converted into spatial vector. We have implemented doc2vec approach to convert the entire threat document into latent vector space [30]. Doc2Vec, as opposed to conventional bag-of-words simulations, may record the semantic content of whole articles or document.
- *Domain Classification Model:* To mapped the CTI with respective domain, we have to identify the domain of the specific threat document. In our research, we have employed two hybrid model of CNN-LSTM and CNN-GRU to classify the domain of the CTI based data feed. Our model will categorize the CTIs into four domains viz Energy Sector, Finance, Illicit Activities and illegal Services. Orchestration of our model is depicted in Figure 2.
- *Threat Intel Extraction:* Various IOCs are embedded inside of the threat document. Additionally, these threat feeds are comprising of IOCs and CTI keywords. In order to extract the IoCs we have implemented RegEx parser for validating Malicious IP address, domains, executables, hashes and spam emails. However, a separate approach has been adopted to extract the CTI keyword [31-32]. For this purpose, we have adopted Subject, Object and Verb (SOV) semantics dependency analysis. By virtue of such approach, we have designed an ontology of threat action candidate keys and employing similarity index we have filtrate threat keywords and IoCs that are analogy to the threat candidate keys.
- *Integration of IoCs to generate Domain specific CTIs which enhance the quality of CTI:* Integration of IoCs with respective domain will enhance the productivity of the Threat intelligence that may be integrated in security measures. CTI extracted from the datafeed contains voluminous IoCs and threat description. However, in real enterprise scenario majority of the TI are not relevant or the precautious measures regarding the vulnerability has already been taken care of. In this context, indexing the CTI is a precursor for building a qualitative and resilient security solution. This component ranked the extracted CTI based on their relevancy from the client purview.

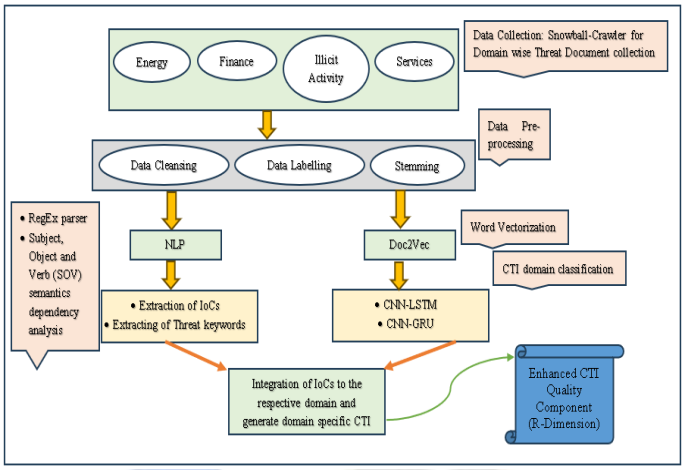

Fig 2: Outline of the Proposed Model

IV. Methodology

In this part, we will discuss our suggested approach to clustering the dark web for cyber threat information using machine learning and deep learning. According to the proposed method, it is necessary to develop separate models for each kind of intelligence to be retrieved. Whenever a model is developed, it is necessary to accumulate posts corresponding to it and identify the gathered posts, and the costs of developing the model are considerable. In the proposed technique of this investigation, first characteristics are extracted using doc2vec from the collected posts, as in our earlier work. Then, deep clustering, which is a machine learning methodology that incorporates auto-encoding and clustering, is performed on the collected features. In this method, the posts are categorized into each cluster by kind without labelling the gathered posts. The goal is to construct a model that designates numerous sorts of posts as essential posts and divides them into distinct clusters. In this study, we want to check if our technique can classify posts on the dark web appropriately. In the research of this study, we set forum postings connected to malware offerings as important posts, as in the proposed manner of our earlier work. The following covers each stage of the proposed method.

A. Data Collection:

Identify reliable sources is one of the most crucial aspects for data collection. In this stage we have to first identify Dark Web sources [33] which are well known for cyber threat intelligence resources. These could include forums, marketplaces, chat rooms, or other platforms frequented by cybercriminals [34]. Develop a web crawler capable of accessing and retrieving relevant data from the Dark Web sources. Additionally, we have to ensure the crawler adheres to ethical guidelines and legal requirements for scraping dark web. In this context we have designed a Snowball-crawler, which is

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

an automated framework and able to collect security related data feed from darknet portals. The. onion subdomains are frequently followed by an arbitrary array of characters and numbers in the URL addresses of the websites on the dark web. These domains are not accessible using conventional browsers and must be resolved using the TOR browser [35, 36].

Snowball-crawler is a python-based automated script, which is first scraps the hidden services and works on the principle of Breadth-first search approach. As the web address of the requisite links are hard to find out for dark web, hence, snowball sampling is most signification method to find the hidden services in dark web. If the number of neighbours k that can be included in the sample is restricted, snowball approach is best fit for such scenario. In this case, Start with a k-node random selection of nodes. The second sample stage is created by adding each of the additional k nodes after that. This keeps on till the required sample size is attained. The Snowball sampling is a sort of sampling by discovery whereby each member of the sample is requested to identify k distinct members of the overall population, where k is a given number; for instance, each member could be requested to list his "k peers". Pseudocode for Snowball-Miner is mentioned in Algorithm 1.

Annotation: Considering set of hidden services S, which may extend up to N, whereas pull of hidden services connected to s ∈ S using the linked edges is E(s). Here initial individual vertex which represents a hidden service is denoted as ego and the identified hidden service connected to all other vertices or hidden services are known as alters. By deliberately choosing a portion of a system's vertices and edges, snowball sampling is a recurrent assessment approach that seeks to disclose architectural data pertaining to a network.

Algorithm 1: Snowball-Crawler Functionality for data collection

Require: Initial identified hidden service s ∈ S
Ensure: Entire set of hidden services S = {s_1, s_2, …s_N}
1: SET initial iteration j =0
2: Choose a random vertex or hidden Service S (0) of vertices from the network.
3: For each j ∈ N as long entre set of hidden services are not discovered do
4: Increment the value of j by 1
5: Request each ego s in $S^{(j-1)}$ to find its linked alters. Considering $S^{(i)}$ all neighbours alter that are not triggered earlier. i.e.- $S^{(i)}$ = U E(s) - $S^{(0)}$ - …. - $S^{(i-1)}$ where s∈ $S^{(i-1)}$
S ← APPEND $S^{(i)}$
Increment j
6: End For

In our research we have scrapped around 90,000 threats oriented data feed from various .onion websites which are categorized into four basic domain such as Energy, Finance, Illicit Activities and Service. Posts related to these domains are categorized based on the related sub-domain as specified in Table I:

Table I: Mapping of Domain and relevant Sub-Domains

Domain	Sub-Domain	Description
Energy Sector	• Power Sector • Oil and Natural Gas sector • Wind and Solar energy	Vulnerabilities released on Darknet applicable for Energy sector
Finance	• Banking fraud • Card credential stolen • Money Laundering	Credit card data, Bank account details stolen and ready to sell at Dark web.
Illicit Activities	• Illegal Pornography • Illegal Political Content • Motivational Militant activities	Ideology of militant activities sspread across dark net
Service	• DDoS-As-a-Service, Ransomware-As-a-Service, Hacking-As-a-Service • Ammunition trading • Drugs trading • Violence, Hitman hiring	Hiring hitman for spreading of violence, various illegal services hiring

In this experiment we have considered only those threat related posts or datafeed that are relevant to these four domains only. Apart from these area darknet caters substantial amount of data related to other perspective such as hosing of fraud websites that contains illegal contents, communal ideology related forums, abusive video regarding animal assassination, human trafficking etc. which were excluded in our experiment. In fact, four cybersecurity experts—three corporate personnel and one senior management professional invested their rigorous effort for meticulously labelling the data gathered by the Snowball-Miner in order to train and test our suggested solution. Total threat documents are categorized into four domain and the count of such documents are depicted in Figure 3. Finally, when our labelled dataset is ready it contains around 11000 labelled descriptions diversified into four categories like, Energy Sector, Finance, Illicit Activities and Service.

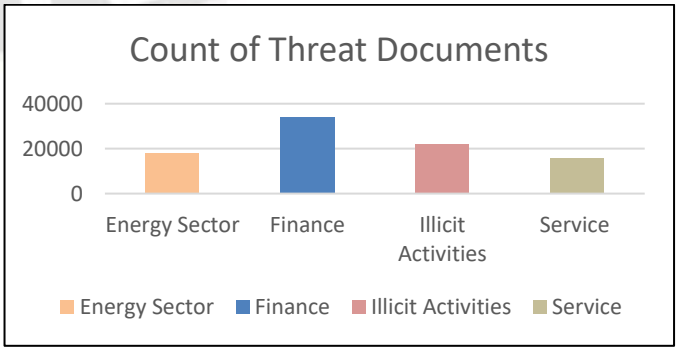

Fig 3: Breakup of threat data feed collection

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

B. Data Preparation:

In order to accomplish unsupervised machine learning, it is essential to collect a substantial quantity of data. Then, we need to identify the obtained dataset as accurate or incorrect to construct the learning data. In our technique, databases are collected using a Snowball-Miner., that specializes in data collection on the dark web. It must be mentioned that there is a risk in collecting the information because the dark web forums may contain malware. We collected 90,000 forum posts relevant to malicious intended offers. Approach of data filtering, filter the collected data to remove noise, irrelevant information, or false positives that may hinder the model's performance. Most of the threat feeds contains irrelevant data, advertisement etc. Removing such unintended data from the final dataset diminishing the dimension of the data at the same time it reduces the noise as well.

Data labeling: Assign appropriate labels or categories to the filtered data, such as gun related, drugs related, malware, phishing, ransomware, child porn or other relevant threat types [37]. This labeling will serve as the ground truth for training the deep learning model and the machine learning model.

Data Cleansing: In our method, stop-words cleaning, duplicate removal, stemming, normalization and tokenization are applied as preprocessing. Tokenization is the process of clarifying the separation between words. The cleaning procedure deletes superfluous characters such as numbers and parentheses () in the text [38]. The normalization process transforms the text into uncommon or uppercase. The stop-words procedure removes unneeded words that appear in any phrase, such as "I".

Data Stemming: The stemming method determines that derivative words have the same root. Due to fast processing and aggressive nature, we prefer to implement the snowball stemmer approach to stem the words. Snowball Stemmer is component of nltk package-

```
from nltk.stem.snowball import SnowballStemmer
```

Snowball stemming, is the process of reducing a word to its root stem so that terms of the same sort are grouped together under a single stem [39-40]. Some instance of stem words extracted from data feed are-

Vulnerability, vulnerabilities → Vulnerable
attacks, attacked → attack
malwares → malware
wattage → watt

Pseudocode for stemming is mentioned in Algorithm 2:

Algorithm 2: Stemming Pseudocode

Require: Consume a word from the Document corpus w ∈ W
Ensure: Stemmed word vector V = {v_1, v_2, …., v_N}
1: SET initial iteration i =0
2: SET W= w_1
3: For each i up to N where N is the total count of word exists in the corpus Do:
4: IF w = "Stop-words": RETURN
5: Run Normalization Filter
6: Run Stemming Filter
7: IF length of word > 2:
8: V_i←APPEND (w_1)
9: END IF
10: Increment the value of j by 1
11: End For

C. Word Vectorization:

Because forum postings are textual data, artificial language processing needs to be conducted to utilize the material as input for machine learning. Additionally, it is important to identify characteristics suitable for categorization as a preparatory phase of artificial learning. Our solution employs doc2vec for conversational language processing and characteristic extraction [41]. The attributes acquired here represent the meaning of the term in context. We use the newly gathered dark web postings as feeds to our model as unobserved data. Natural phrase analysis is carried out on the gathered data with doc2vec to obtain the features. Consequently, the characteristics are given as a stimulus to the model, and the collected postings are categorized into each cluster. Pre-processing of conversational speech processing is important for appropriately vectorizing documents and getting features. Parameters passed in doc2vec() are enlisted below-

vector_size ==250	***Window***=8	***epochs***=50	***min_count***=5

D. Designing of Domain Classification Model:

In this research we have implemented hybrid CNN-LSTM and CNN-GRU classifiers to classify the domain of the threat data feed [42-43]. CNN is significantly effective to the forecasting of entities, which is designed based on the concept of feedforwarding neural network. CNN is used to minimized the count of attributes using the feature of weight distribution and perception layer. Pooling layer and convolution layers are two core components of CNN whereas convolutions layer comprises of kernels. Extraction of features through the convolutional layers is in large scale. In order to reduce the dimension of the features, a pooling layer has been included in this experiment. Particularly, we used feature map 256 and kernel= 5 with pooling size= 3.

$$l_i = \tanh\left(x_t * k_t + b_t\right) \quad \text{…….. (1)}$$

Where, li= convolutional outcome; bt = bias value, kt = Weight and xt = Input value.

In further stage CNN output age feed into the layer of LSTM layer to maintain the uninterrupted traffic flow for sequence prediction. LSTM is another variant of RNN removing the hidden layer with memory cell for faster processing of data feed. Final out put extracted from LSTM layer are calculated using the formulas:

$$f_t = \sigma\left(W_f, \left[h_{t-1}, x_t\right] + b_f\right) \quad \text{…….. (2)}$$
$$i_t = \sigma\left(W_i, \left[h_{t-1}, x_t\right] + b_t\right) \quad \text{…….. (3)}$$
$$C_i = tanh\left(W_c, \left[h_{c-1}, x_c\right] + b_c\right) \quad \text{…….. (4)}$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

$$C_t = f_t * C_{t-1} + i_t + C_t \quad \text{..... (5)}$$
$$O_i = \sigma\,(W_o\,, [h_{t-1}, x_t] + b_o) \quad \text{..... (6)}$$
$$h_t = O_t * \tanh(C_t) \quad \text{..... (7)}$$

An improved version of the LSTM model, GRU has fewer parameters than LSTM, an easier design, and a greater effectiveness for training. As a result, it performs certain tasks more quickly. During learning long-term patterns, the "vanishing gradient" issue is addressed with GRUs. Architecture of CNN-GRU model has been illustrated in Figure 4.

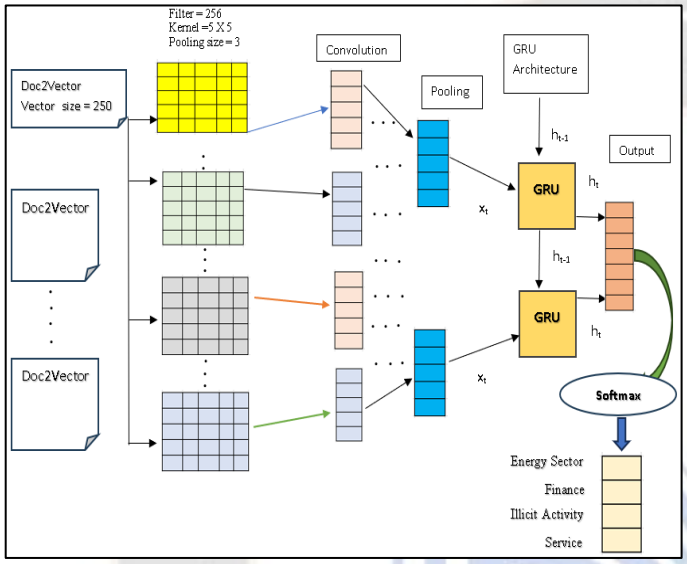

Fig 4: Design of CNN-GRU Model

We have evaluated the result of our model with the supervised classification algorithms like- Naïve Bayes, XGB and Logistic Regression. NB classifier is used as a probabilistic classifier that operates under the presumption that the input features are uncorrelated [44]. The Extreme Gradient Boosted Trees algorithm is a supervised learning approach, which is based on function approximation by optimizing certain loss functions and employing a number of regularization approaches. XGBoost is one of the most well-known and effective variants of this algorithm [45]. The gradient boosted decision tree implementation known as XGBoost was created for speed and performance when applied to hierarchical and table-based information. An ensemble modelling method called XGBoost is significantly effective with big, complex datasets. XGB objective function comprises of loss function and regularization for individual iteration t which is intended to minimize is the following:

$$\boldsymbol{\omega^t} = \sum_{i=1}^{n} l\left(z_i\,, \hat{z}_{i_i}^{(t-1)}\right) + f_t\,(x_i)) + \varphi\,(f_t) \quad \text{..... (8)}$$

Where l is Classification and Regression Trees (CART) learner function, which is considered as a summation of nth and (n-1) th additive trees.

E. Extraction of IoC and Threat Keywords:

a) *RegEx Parser:* Threat data documents contains a substantial sources of threat intelligence in the form of IoCs and threat keywords [46]. An NLP approach is employed to extract the paragraphs "d" that are supposed to contain the threat keywords [47-50]. Spurious data as well as irrelevant data such as advertisements, pop up contents are discarded in our research. However, stop words confused the conventional NLP to extract the threat data content. In order to avoid such backdrops, we used a RegEx parser which is able to filtrate the special patterned IoCs viz.- malicious IP address, SPAM email, corrupted file executives, infected hash files, threat domains and websites as individualized below in Table II –

Table II: IoC extracted using Regular Expression

Text in threat report (T ∈ p)	IoC extracted using RegEx parser	IoC Type
Ryuk ransomware attack launched from 172.16.23.197	172.16.23.197	IP Address
Hash File	9b81bad2111312e669697b69b9f121a1f9519da61cd5d37689e38381c1ffad28	Hash File
Website http://37.44.238.213/nekoY/netgear contains malicious activities	http://37.44.238.213/nekoY/netgear	URL
Data breach incident identified in uidaigov.co	uidaigov.co	Domain
1jt0477@gmail.com populates bulk mails to the SCADA systems	1jt0477@gmail.com	E-mail
A command injection vulnerability has been discovered in COMFAST. The affected version is COMFAST CF-XR11 V2.7.2. CVE-2023-38866 (Critical)	CVE-2023-38866	CVE
Privileged escalation conducted by creating a boot file boot_17_create.exe	boot_17_create.exe	Executable Files

Pattern identification of IoCs using the RegEx parser is mentioned in Figure 5

Regular expression samples of recognizing IOC.

IOC TYPE	Regular Expression
CVE	CVE-[0-9]{4}-[0-9]{4,6}
MD5	[a-f 0-9]{32}|[A-F 0-9]{32}
SHA1	[a-f 0-9]{40}|[A-F 0-9]{40}
Email	[a-z][_a-z0-9_]+[a-z0-9]+.[a-z]
Register	[HKLM|HKCU]\\[A-F 0-9]{40}
IP	\d{1, 3}.\d{1, 3}.\d{1, 3}.\d{1, 3}

Fig 5: Regex Parser approach

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

b) Threat Action Identification using NLP and Subject-Object-Verb (SVO) combination:

Threat descriptions contains TTPs and non-TTPS articles (Such as help, advertisement etc.). We considered only TTP articles to extract the threat actions. Threat actions have a typical structure, comprising of Subject, Object and Verb (SOV). Standford dependency representation is suitable for extracting these SOV structure which are referred as the threat action candidate keys. For instance, ransomware "Cryptolocker" is referred as Subject, destination like 65.38.5.19 is the object and "compromised" is the action. In graphical representation originating node is called as primary and the target node is considered as the secondary node [26].

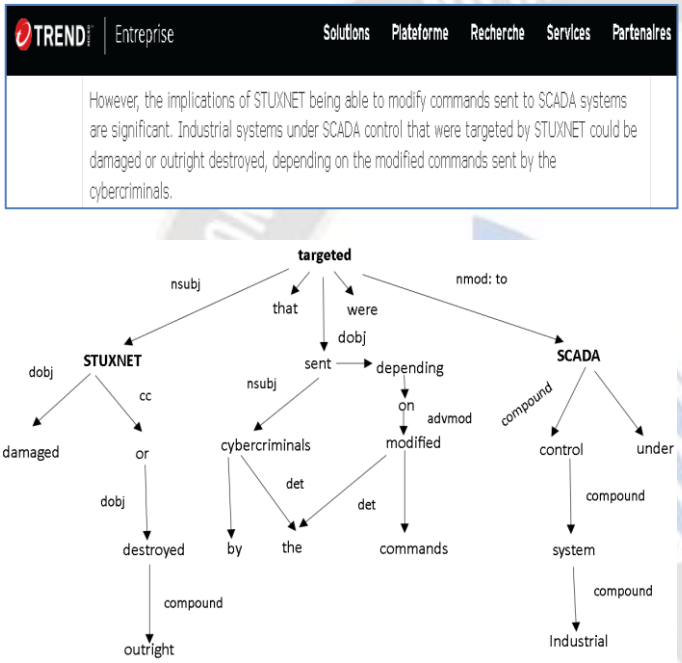

Fig 6: Syntactics analysis of the type dependency and the relation among the primary and secondary nodes

Syntactics analysis of the type dependency and the relation among the primary and secondary nodes are individualizes in Figure 6 for the sentence "Industrial systems under SCADA control that were targeted by STUXNET could be damaged or outright destroyed, depending on the modified commands sent by the cybercriminals". “STUXNET” and “targeted” with nsubj, where “STUXNET” is the noun clause and “targeted” is the verb and the noun is controlled by the verb by means of which “STUXNET targeted” word will be extracted as the threat action [52]. Stanford typed dependencies for identification of threat actions used in this study is illustrated in Table III.

Table III: Stanford typed dependencies for identification of threat actions

Dependency type	Threat Action
nsubj	primary<verb>, secondary<subj>
dobj	primary<verb>, secondary<obj>
nmod: agent	primary<verb>, secondary<subj>
nmod: to	primary<verb> to, secondary<obj>
nmod: under	primary<verb> to, secondary<obj>
det	determiner
cc	coordination
prep	preposition
advmod	adverb modifier
nmod: using	primary<verb> using, secondary<obj>
nsubjpass	primary<verb>, secondary<obj>

Threat Action/Description: Finally, we have designed an ontology of threat action candidate keys as mentioned in the Table IV.

Table IV: ontology of threat action candidate keys

Threat Action document	Threat Action candidate key
Industrial systems under SCADA control that were targeted by STUXNET could be damaged or outright destroyed, depending on the modified commands sent by the cybercriminals.	STUXNET
	STUXNET targeted Industrial system
	STUXNET damaged Industrial system
	STUXNET targeted SCADA
	STUXNET destroy SCADA
	Cybercriminals
	Cybercriminals sent modified command
	Modified command destroyed SCADA

Similarity index: The TF-IDF and BM25 computations are used within this segment to determine how identical threat documents d is to one another. In the process of text vectorization, the attribute element frequency is often determined using the TF-IDF approach [17]. Consider P as the count of all words in the threat data document d, P(k) is the entire set of words k in the threat data document d and weight of an attribute item is measured by the formula-

$$\mathrm{TF} - \mathrm{IDF}(k) = \frac{P}{P(k)} * (\log(\frac{P+1}{P(k)+1}) + 1) \quad \text{…….. (9)}$$

Whereas, BM25 classifier is an extended variant of TF-IDF and the expansion ceiling of the TF number is limited by adding an integer to the TF-IDF, and the threat document size is used to assess the priority of the threat candidate key. Cumulative correlative index of BM25 classifier is mentioned below considering-

tf	represents the frequency of each word
idf	the inverse word frequency of each word
L	is the length of the text
and k_1, k_2, and b	the adjustment factors

$$Similarity\ index = idf * \frac{(k_1+1)*tf}{k_2*(1-b+b*L)+tf} \quad \text{…….. (10)}$$

To finalize the entire list of Threat keywords, we have extracted all similar keywords based on probabilistic approach by mapping Threat candidate keys with the threat documents using similarity index.

F. Integration of Cyber Threat Intelligence respective Domains to generate domain specific TI and improves the quality of threat intelligence:

In this segment we will extract the domain oriented cyber threat intelligence. In order to perform the task, firstly, we have to extract only the threat keywords from the threat action candidate keys and append all the threat keywords to the list TKi = {k1, k2, k3…ki}, where (1 ≤i ≤N). Thereafter, all IoCs extracted using RegEx parser are amalgamated with TKi to

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

create a IoC vector IOC cand id ate such that IOCcandidate= TKi U IOCi and form a comprehensive list of all IoCs inclusive of all threat keywords-

$$IOC_{cand\ id\ ate} = \{ioc_1, ioc_2, \ldots, ioc_i\} \text{ where } 1 \leq i \leq N \quad \textbf{…….. (11)}$$

In our experiment threat documents are D = {d1, d2,,…., di} where 1 ≤i ≤N and we create a corpus of verbs that initiate a threat action against di
is explained in Algorithm 3.

$$V_{Verb} = \{s_1, s_2, \ldots s_i\} \text{ where } 1 \leq i \leq N \quad \textbf{…….. (12)}$$

Now, considering every threat document d_i, based on strong linguistic correlation with each verb s_i we have extracted the relevant ioc_i and mapped each IoCs of di to the corresponding domain tag. Pseudocode for integrating domain tag with the IOCs

Algorithm 3: Pseudocode for Integration of Domian wise Threat Intelligence

Require: Threat Document D={d_1,d_2,…d_i}
Domain Label $\tau = \{\tau_1, \tau_2, \ldots \tau_i\}$
Ensure: Threat Intelligence integrated with Domain Tag = {Energy, Finance, illicit Activities, Services}

```
For Each di ∈ D Do
    Doc2vector ← doc2vec (data cleansing (di))
    For each epoch for CNN Model Do
        Attributes ← maxpooling of convolution of doc2vector
        For each epoch for GRU Model Do
            Ẑi = max (Softmax of linked attributes)
        End For
        Evaluate the - ∇ of Σ Zi log Ẑi where 1 ≤i ≤N and assign to C (Zi, Ẑi)
    End For
    d_τi ← Tag di ⊂ τ
    For each d_τi Do
        IOC cand id ate ←Threat_Keyword (TKi) U IoC obtained from RegEx Parser
        Vverb ← Corpus of verb that initiate a threat action
        For each Si ∈ Vverb Do
            For each ioci ∈ IOC cand id ate Do
                Check for semantic relationship between ioci and vi and if dependencies are found then Append (Xi, τi)
            End For
        End For
    End For
End For
```

With regard to the increasing volume of CTI, it is now obvious to extract the quality CTI for the different domains. In this section we have focused R-dimension (Relevance) for evaluation of quality dimension of extracted CTI-

R- Dimension: IOCs are useful for mitigation of cyber-attack risks, when it has been integrated with security measures of appropriate domains. For instance, an SOC is in place for a state-of-art SCADA system and the IOCs are integrated to the SOC of SCADA may not effective to that extend for a financial institution and vice-versa. Thus, implanting of domain relevant IOCs to the security measures solutions confirms the relevance quality dimension of CTIs. In order to achieve such quality CTI, we need to segregate the CTI to domain specific [54-55]. Here we focused four domains in our experiment such as Energy, Finance, illicit Activities, Services and we tag the CTIs into these four domains as per Algorithm 3.

V. Result and Discussion

In our experiment we have classified the threat documents into four domains such as – Energy sector, Finance, illicit Activities, Services. Threat documents are collected through a Snowball-crawler, which is a core component of snowball-miner. Here, we have monitored around 62 distributed onion sites to collect our intended threat documents using 48 nos of crawlers. These onion sites yield 90,000 forum posts relevant to malicious intended offers. Using CNN-LSTM and CNN-GRU model to domain wise categorization of the threat documents We have evaluated the precision, recall and accuracy of the model. In order to access the performance of the domain classifier we also employed supervised machine learning based classifiers like Naïve Bayes, XGB and Logistic Regression. In our setup we have tested our models' performance on 80-20 and 70-30 ration of train and test data set. We have calculated the Accuracy, Precision and recall parameters (Figure 7) of the designed model for evaluating the classification of threat domain.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

		Prediction		
		+	-	
Actual	+	TP True Positive	TN True Negative	Recall = $\frac{TP}{TP + FN}$
	-	FP False Positive	FN False Negative	
		Presion = $\frac{TP}{TP + FP}$	Accuracy = $\frac{TP + TN}{TP + FP + FN + TN}$	

Fig 7: Performance Evaluation Parameters

Initially, we have tested the classification process through various ML model with two-fold train test ratio viz- 80-20 and 70-30 and among three tested ML model XGB outperforms the other model NB and LR in 80-20 train-test ratio. In order to classify the domain-oriented threat documents XGB model return 93.87% in terms of accuracy whereas, LR return 88.73% of accuracy for the same. Detail of the comparison of the performance metrices for threat domain wise threat categorization is illustrated in Table V.

Table V: comparison of the performance metrices for threat domain wise threat categorization

ML Model	Train-Test Ratio	Precision	Recall	Accuracy
Naïve Bayes	80-20 Ratio	84.63	87.92	91.07
	70-30 Ratio	85.02	88.19	90.69
Extreme Gradient Boost	80-20 Ratio	88.32	90.89	93.87
	70-30 Ratio	82.27	89.23	92.17
Logistic Regression	80-20 Ratio	79.82	82.33	88.73
	70-30 Ratio	76.53	80.65	88.92

We have employed two hybrids deep learning convolutional model such as CNN-GRU and CNN-LSTM. Entire threat documents are split into 80-20 and 70-30 ratio for training and test data set. Out of these models CNN-LSTM is 96.37% accurate for predicting the domain specific threat documents which is shown in Figure 8 and Figure 9.

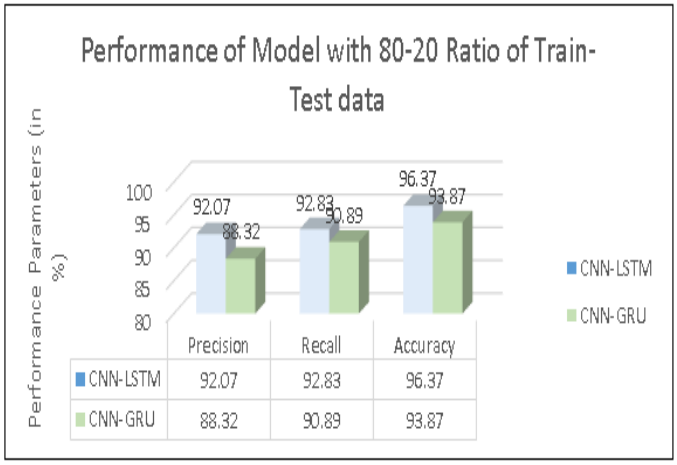

Fig 8: Performance of CNN-GRU and CNN-LSTM in 80-20 Ratio

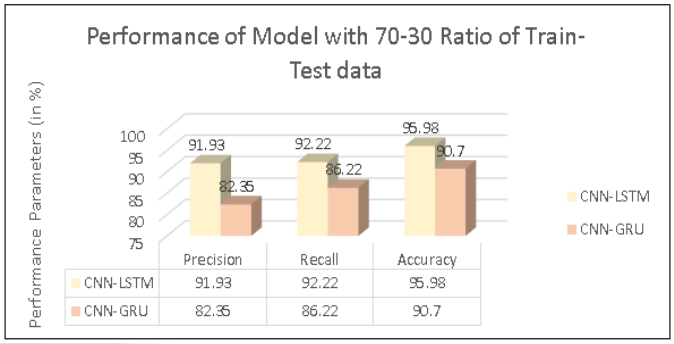

Fig 9: Performance of CNN-GRU and CNN-LSTM in 70-30 Ratio

Entire experiment was set up in development environment with the following configuration as specified in Table VI. To gather the data, we used snowball-crawler which is a python-based script executed on the principle of Breadth-first search approach which was executed on Jupyter notebook and spyder environment. Additionally, surface web are accessible from any normal web browser whereas traditional web browser are unable to access the deep web site or .onion web pages. In our experiment we used Tor browser and DuckDuckGo, haystak search engine to accelerate the snowball approach to find the new dark web links.

Table VI: Development Environment Setup

Environment	Jupyter Notebook, Spyder v5.4.1
OS	Windows 11 Professional 22H2 pack
Programming Language	Python 3.7.4
CPU	12th Gen Intel(R) Core (TM) i7-12700 2.10 GHz
Memory	16 GB
Graphics	Nvidia Graphics
VM	Virtual Box
Browser and Search Engine	Tor, Google Chrome, DuckDuckGo, haystak

In respect model training time LR consumes least timing to train the system whereas CNN-LSTM spent highest span i.e. 23.37 Minutes for build the training model as mentioned in Figure 10.

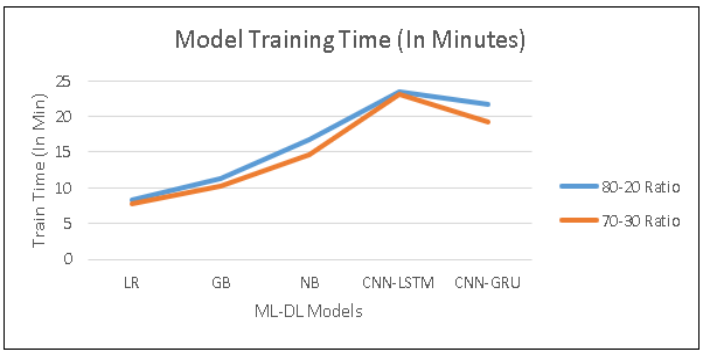

Fig 10: Time consumption to develop the Model

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Implementing Tor browser with additional VPN security measures in VM environment we scarped various deep web pages along with surface website for collecting threat related documents. Few of the instance of such .onion pages are mentioned in Table VII.

Table VII: List of Dark web URLs and category

Marketplace/ Group Identity	URL of Darkweb	Domains of Threat Documents
WeAreAMSTERDAM	amster4mdi4ls27irjhe3ul6hzha5itf6naifanpfbdbfybclbnqipyd.onion	Illegal Drug
Kraken market	ns2kzt6hgke6xblt3tet3yq2q5zuawx3e56qrwlomeg3jobbqnzbcoid.onion	Illegal Drug
Incognito Market	zguhsmugwcb7iivknafp7ph3twup57olx5zulhbjacgtetihrvg7maad.onion	Illegal Casino
Mitnicklens Hackers	zjuhnrnaz573362iwpqm2peoozwd37yfeohyzdymmn2xlewr5li3zpqd.onion	Hire a hacker
Deep web Besa Mafia	zsyvom262oiaoc6es7bgg66xieyil6nqkh7jn5ntraghpqgudbcl3vad.onion	Hire a hacker
18th Street Gang Hitmen Marketplace	h4gca3vb6v37awux.onion	Hitman Hiring
CyberTeam	m26hbhcbee73o42wyltb7dcog6d5jjzelbib64ojcew6cso625tpxwqd.onion	DDOS as a service
Pundit Hacker	vscvkdcnjpwkdumrrnxsfhmx5shkztqzehnkvelpfrzj7sqkra7bcjid.onion	Reputation Damage
Red Onion	rwgfulliwadfagfeookhqsdsfdsomiiowcs.onion	Power plant attack group
Identity Documents Forge Master	Fobgemanrtpvbxok.onion	Counterfeit documents
	gcardstguk366hru.onion	Cards Credit Card

DISCUSSION: In our experiment we have observed that, Deep learning model CNN-LSTM is 96.37% accurate for predicting the domain specific threat documents which is significantly remarkable outcome compared to the other deep learning models and the NB, XGB, LR based Machine learning model. On large datasets, deep learning models outperform other models in terms of accuracy in the result. In our case, threat documents are large in collection at the same time these are unstructured in nature. CNN-LSTM hybrid model is perfectly suitable for such large volume of data which are unstructured in nature.

XGB outperforms to the other ML based classifier when we distribute the data into 80-20 ratio. Despite of the fact that, LR classifier is best fit algorithms for traditional classification problem, this experiment reflects (Table V) that LR is worst fit algorithms when the dataset is large in size and unstructured in nature. In this experiment we have observed that, Accuracy rate is proportional to the train data size for XGB algorithm whereas Accuracy rate is inversely proportional to the train data size for LR algorithm. However, accuracy rate is independent to the train data size for probabilistic Naïve Bayes model.

CNN-LSTM is employed to categorise each and every single traffic network, thus time taken to train the model will be highest compared to the other hybrid convolutional model. Time consumption for train the model in 80-20 ratio is organized as LR<GB<NB<CNN-GRU<CNN-LSTM. However, CNN-LSTM is independent to the train data size in respect of time consumption to train the model. CNN-LSTM takes 23 minutes for both the train size i.e., 80% and 70%.

Both of our suggested LSTM- and GRU-based models performed noticeably better than the current models. In principle, when contrasting the CNN-LSTM and CNN-GRU models, the CNN-GRU model can slightly surpass the CNN-LSTM model when there is less training data. Though CNN-LSTM computes much slower due to its higher complexity than the CNN-GRU formula, at the same time data are voluminous in nature and continuous update on previous time-steps, we prefer CNN-LSTM model for accurate prediction of classification of threat documents into appropriate domain.

VI. Conclusion and Future Work

Main goal of our research work is to leverage dark web resources to generate domain wise threat intelligence. In order to achieve such aims, we have considered the dark web for collecting substantial data source. State-of-the-art research emerged that, majority of the threat extraction experiments was carried out on the surface web platform like security vendor blogs, forums etc. In our research, we have scrapped the dark web market places to gather threat related documents and implementing various ML and DL model we have classified the threat documents into four domains- Energy sector, Finance, illicit Activities, Services. CNN-LSTM outperforms as 96.37% accurate for predicting the domain specific threat documents. Simultaneously, we have employed NLP on the threat documents and extracted the relevant IoCs and threat keywords from the documents using RegEx parser and subject, Object and Verb (SOV) semantics dependency analysis. Thereafter, we have integrated the IoCs with the respective domain s to generate domain wise threat intelligence. This approach also improves the quality of the extracted threat intelligence in terms of R-Dimension (Relevancy), as these IoCs /threat keywords are domain oriented which enhanced the effectiveness of the intelligence.

- In the future work, we will explore the heterogeneous corpus of data collection from surface web and dark web to measure the performance metrices implementing the convolutional network with the hybrid convolutional model.

- Other quality parameters like ETC-Dimension (Extensiveness, Timeliness and Completeness) will be explored to enhance the quality of the threat intelligence.

REFERENCES

[1] Kim, K., Alfouzan, F.A. and Kim, H. (2021). Cyber-Attack Scoring Model Based on the Offensive Cybersecurity Framework. Applied Sciences, 11(16), p.7738. doi:https://doi.org/10.3390/app11167738.

[2] Liu, K., Xu, S., Xu, G., Zhang, M., Sun, D. and Liu, H. (2020). A Review of Android Malware Detection Approaches Based on Machine Learning. IEEE Access, 8, pp.124579–124607. doi:https://doi.org/10.1109/access.2020.3006143.

[3] Sahoo, K.S., Puthal, D., Tiwary, M., Rodrigues, J.J.P.C., Sahoo, B. and Dash, R. (2018). An early detection of low rate DDoS attack to SDN based data center networks using information distance metrics. Future Generation Computer Systems, 89, pp.685–697. doi:https://doi.org/10.1016/j.future.2018.07.017.

[4] Lehto, M. (2022). Cyber-Attacks Against Critical Infrastructure. Computational Methods in Applied Sciences, pp.3–42. doi:https://doi.org/10.1007/978-3-030-91293-2_1.

[5] Nicholson, A., Webber, S., Dyer, S., Patel, T. and Janicke, H. (2012). SCADA security in the light of Cyber-Warfare. Computers & Security, 31(4), pp.418–436. doi:https://doi.org/10.1016/j.cose.2012.02.009.

[6] Qassim, Q.S., Jamil, N., Mahdi, M.N. and Abdul Rahim, A.A. (2020). Towards SCADA Threat Intelligence based on Intrusion Detection Systems - A Short Review. 2020 8th International Conference on Information Technology and Multimedia (ICIMU). doi:https://doi.org/10.1109/icimu49871.2020.9243337.

[7] Ajmal, A.B., Alam, M., Khaliq, A.A., Khan, S., Qadir, Z. and Mahmud, M.A.P. (2021). Last Line of Defense: Reliability Through Inducing Cyber Threat Hunting With Deception in SCADA Networks. IEEE Access, 9, pp.126789–126800. doi:https://doi.org/10.1109/access.2021.3111420.

[8] Alghassab, M. (2021). Analyzing the Impact of Cybersecurity on Monitoring and Control Systems in the Energy Sector. Energies, 15(1), p.218. doi:https://doi.org/10.3390/en15010218.

[9] A. Yeboah-Ofori, S. Islam and E. Yeboah-Boateng, "Cyber Threat Intelligence for Improving Cyber Supply Chain Security," 2019 International Conference on Cyber Security and Internet of Things (ICSIoT), Accra, Ghana, 2019, pp. 28-33, doi: 10.1109/ICSIoT47925.2019.00012.

[10] ieeexplore.ieee.org. (n.d.). Cyber Security of Smart Grids in the Context of Big Data and Machine Learning. [online] Available at: https://ieeexplore.ieee.org/abstract/document/8745044/ [Accessed 12 Oct. 2023].

[11] Lan, J., Liu, X., Li, B., Li, Y. and Geng, T. (2022). DarknetSec: A novel self-attentive deep learning method for darknet traffic classification and application identification. Computers & Security, 116, p.102663. doi:https://doi.org/10.1016/j.cose.2022.102663.

[12] Habibi Lashkari, A., Kaur, G. and Rahali, A. (2020). DIDarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic using Deep Image Learning. 2020 the 10th International Conference on Communication and Network Security. doi:https://doi.org/10.1145/3442520.3442521.

[13] Moore, D. and Rid, T. (2016). Cryptopolitik and the Darknet. Survival, 58(1), pp.7–38. doi:https://doi.org/10.1080/00396338.2016.1142085.

[14] Enoch, S.Y., Huang, Z., Moon, C.Y., Lee, D., Ahn, M.K. and Kim, D.S. (2020). HARMer: Cyber-Attacks Automation and Evaluation. IEEE Access, 8, pp.129397–129414. doi:https://doi.org/10.1109/access.2020.3009748.

[15] Samtani, S., Li, W., Benjamin, V. and Chen, H. (2021). Informing Cyber Threat Intelligence through Dark Web Situational Awareness: The AZSecure Hacker Assets Portal. Digital Threats: Research and Practice, 2(4), pp.1–10. doi:https://doi.org/10.1145/3450972.

[16] Zhao, J., Yan, Q., Li, J., Shao, M., He, Z. and Li, B. (2020). TIMiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data. Computers & Security, 95, p.101867. doi:https://doi.org/10.1016/j.cose.2020.101867.

[17] Zhang, H., Shen, G., Guo, C., Cui, Y. and Jiang, C. (2021). EX-Action: Automatically Extracting Threat Actions from Cyber Threat Intelligence Report Based on Multimodal Learning. Security and Communication Networks, 2021, pp.1–12. doi:https://doi.org/10.1155/2021/5586335.

[18] misq.umn.edu. (n.d.). Linking Exploits from the Dark Web to Known Vulnerabilities for Proactive Cyber Threat Intelligence: An Attention-Based Deep Structured Semantic Model. [online] Available at: https://misq.umn.edu/linking-exploits-from-the-dark-web-to-known-vulnerabilities-for-proactive-cyber-threat-intelligence-an-attention-based-deep-structured-semantic-model.html.

[19] Griffioen, H., Booij, T. and Doerr, C. (2020). Quality Evaluation of Cyber Threat Intelligence Feeds. Applied Cryptography and Network Security, pp.277–296. doi:https://doi.org/10.1007/978-3-030-57878-7_14.

[20] Moraliyage, H., Sumanasena, V., De Silva, D., Nawaratne, R., Sun, L. and Alahakoon, D. (2022). Multimodal Classification of Onion Services for Proactive Cyber Threat Intelligence Using Explainable Deep Learning. IEEE Access, 10, pp.56044–56056. doi:https://doi.org/10.1109/access.2022.3176965.

[21] Connolly, K.M., Klempay, A., McCann, M. and Brenner, P. (2023). Dark Web Marketplaces: Data for Collaborative Threat Intelligence. Digital threats. doi:https://doi.org/10.1145/3615666.

[22] ieeexplore.ieee.org. (n.d.). Extraction of Actionable Threat Intelligence from Dark Web Data. [online] Available at: https://ieeexplore.ieee.org/document/10165477/ [Accessed 19 Oct. 2023].

[23] Apurv Singh Gautam, Yamini Gahlot and Kamat, P. (2019). Hacker Forum Exploit and Classification for Proactive Cyber Threat Intelligence. Lecture notes in networks and systems, pp.279–285. doi:https://doi.org/10.1007/978-3-030-33846-6_32.

[24] Azene Zenebe, Mufaro Shumba, Carillo, A. and Cuenca, S. (2019). Cyber Threat Discovery from Dark Web. doi:https://doi.org/10.29007/nkfk.

[25] ieeexplore.ieee.org. (n.d.). A Method for Extracting Unstructured Threat Intelligence Based on Dictionary Template and Reinforcement Learning. [online] Available at: https://ieeexplore.ieee.org/document/9437858.

[26] Husari, G., Al-Shaer, E., Ahmed, M., Chu, B. and Niu, X. (2017). TTPDrill. Proceedings of the 33rd Annual Computer Security Applications Conference. doi:https://doi.org/10.1145/3134600.3134646

[27] Kanti Singh Sangher, Singh, A., Hari Mohan Pandey and Kumar, V. (2023). Towards Safe Cyber Practices: Developing a Proactive Cyber-Threat Intelligence System for Dark Web Forum Content by Identifying Cybercrimes. 14(6), pp.349–349. doi:https://doi.org/10.3390/info14060349.Rawat, R., Mahor, V., Chirgaiya, S., Shaw, R.N. and Ghosh, A. (2021). Analysis of Darknet Traffic for Criminal Activities Detection Using TF-IDF and Light Gradient Boosted Machine Learning Algorithm. Lecture Notes in Electrical Engineering, pp.671–681. doi:https://doi.org/10.1007/978-981-16-0749-3_53.

[28] Rajawat, A.S., Bedi, P., Goyal, S.B., Kautish, S., Xihua, Z., Aljuaid, H. and Mohamed, A.W. (2022). Dark Web Data Classification Using Neural Network. Computational Intelligence and Neuroscience, 2022, pp.1–11. doi:https://doi.org/10.1155/2022/8393318.

[29] Chang, W., Xu, Z., Zhou, S. and Cao, W. (2018). Research on detection methods based on Doc2vec abnormal comments. Future Generation Computer Systems, [online] 86, pp.656–662. doi:https://doi.org/10.1016/j.future.2018.04.059.

[30] Gao, P., Shao, F., Liu, X., Xiao, X., Qin, Z., Xu, F., Mittal, P., Kulkarni, S.R. and Song, D. (2021). Enabling Efficient Cyber Threat Hunting With Cyber Threat Intelligence. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICDE51399.2021.00024.

[31] Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y. and Zhang, J. (2023). Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives. IEEE Communications Surveys & Tutorials, pp.1–1. doi:https://doi.org/10.1109/comst.2023.3273282.

[32] Varghese, V., S, M. and Kb, S. (2023). Extraction of Actionable Threat Intelligence from Dark Web Data. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICCC57789.2023.10165477.

[33] Connolly, K.M., Klempay, A., McCann, M. and Brenner, P. (2023). Dark Web Marketplaces: Data for Collaborative Threat Intelligence. Digital threats. doi:https://doi.org/10.1145/3615666.

[34] Huete Trujillo, D.L. and Ruiz-Martínez, A. (2021). Tor Hidden Services: A Systematic Literature Review. Journal of Cybersecurity and Privacy, 1(3), pp.496–518. doi:https://doi.org/10.3390/jcp1030025.

[35] Rawat, R., Rajawat, A.S., Mahor, V., Shaw, R.N. and Ghosh, A. (2021). Dark Web—Onion Hidden Service Discovery and Crawling for Profiling Morphing, Unstructured Crime and Vulnerabilities Prediction. Lecture Notes in Electrical Engineering, pp.717–734. doi:https://doi.org/10.1007/978-981-16-0749-3_57.

[36] Zhang, S., Jafari, O. and Nagarkar, P. (2021). A Survey on Machine Learning Techniques for Auto Labeling of Video, Audio, and Text Data. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2109.03784.

[37] Krishnan, S., Franklin, M.J., Goldberg, K., Wang, J. and Wu, E. (2016). ActiveClean. Proceedings of the 2016 International Conference on Management of Data. doi:https://doi.org/10.1145/2882903.2899409.

[38] Anjali, M. and Jivani, G. (n.d.). A Comparative Study of Stemming Algorithms. [online] Available at: https://kenbenoit.net/assets/courses/tcd2014qta/readings/Jivani_ijcta2011020632.pdf.

[39] 40. snowball.tartarus.org. (n.d.). Snowball: A language for stemming algorithms. [online] Available at: http://snowball.tartarus.org/texts/introduction.html, (accessed on 27 September 2023).

[40] Kim, D., Seo, D., Cho, S. and Kang, P. (2019). Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. Information Sciences, 477, pp.15–29. doi:https://doi.org/10.1016/j.ins.2018.10.006.

[41] Sarwar, M.B., Hanif, M.K., Talib, R., Younas, M. and Sarwar, M.U. (2021). DarkDetect: Darknet Traffic Detection and Categorization Using Modified Convolution-Long Short-Term Memory. IEEE Access, 9, pp.113705–113713. doi:https://doi.org/10.1109/access.2021.3105000.

[42] Zhai, F., Yang, T., Chen, H., He, B. and Li, S. (2023). Intrusion Detection Method Based on CNN–GRU–FL in a Smart Grid Environment. Electronics, [online] 12(5), p.1164. doi:https://doi.org/10.3390/electronics12051164.

[43] Chen, H., Hu, S., Hua, R. and Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. EURASIP Journal on Advances in Signal Processing, 2021(1). doi:https://doi.org/10.1186/s13634-021-00742-6.

[44] ieeexplore.ieee.org. (n.d.). Detecting Cybersecurity Attacks Using Different Network Features with LightGBM and XGBoost Learners. [online] Available at: https://ieeexplore.ieee.org/abstract/document/9319392 [Accessed 16 Oct. 2023].

[45] Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y. and Zhang, J. (2023). Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives. IEEE Communications Surveys & Tutorials, pp.1–1. doi:https://doi.org/10.1109/comst.2023.3273282.

[46] MITRE, "Common Vulnerabilities and Exposures, Accessed on: 25th September 2023. Available: https://cve.mitre.org.

[47] MITRE, Common Weakness Enumeration, Accessed on: 25th September 2023. Available: https://cwe.mitre.org/about/index.html.

[48] MITRE, ATT&CK®, Accessed on: 28th September 2023. Available: https://attack.mitre.org/

[49] MITRE, Scoring CWEs, Accessed on: 29th September 2023. Available: https://cwe.mitre.org/scoring/index.html.

[50] ieeexplore.ieee.org. (n.d.). A Method for Extracting Unstructured Threat Intelligence Based on Dictionary Template and Reinforcement Learning. [online] Available at: https://ieeexplore.ieee.org/document/9437858 [Accessed 17 Oct. 2023].

[51] Schaberreiter, T., Kupfersberger, V., Rantos, K., Spyros, A., Papanikolaou, A., Ilioudis, C. and Quirchmayr, G. (2019). A Quantitative Evaluation of Trust in the Quality of Cyber Threat Intelligence Sources. Proceedings of the 14th International Conference on Availability, Reliability and Security - ARES '19. doi:https://doi.org/10.1145/3339252.3342112.

[52] Schlette, D., Böhm, F., Caselli, M. and Pernul, G. (2020). Measuring and visualizing cyber threat intelligence quality. International Journal of Information Security, 20(1).