Egyptian Informatics Journal 24 (2023) 43–59

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Cyber threat attribution using unstructured reports in cyber threat intelligence

Ehtsham Irshad *, Abdul Basit Siddiqui

Department of Computer Science, Capital University of Science & Technology, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 24 July 2022
Revised 20 November 2022
Accepted 30 November 2022
Available online 9 December 2022

Keywords:
Cyber Threat Actor (CTA)
Tactics Techniques and Procedures (TTP)
Natural Language Processing (NLP)
Cyber Threat Intelligence (CTI)
Cyber-Threat Intelligence Reports (CTIR)
Incident of Compromise (IOC)
Advanced persistent threats (APT)
Structured Threat Information Expression (STIX)
Common Attack Pattern Enumeration and Classification (CAPEC)
Smoothed Binary Vector (SMOBI)
Similarity Based Vector Representation (SIMVER)

ABSTRACT

Cyber-threat attribution is the identification of attacker responsible for a cyber-attack. It is a challenging task as attacker uses different obfuscation and deception techniques to hide its identity. After an attack has occurred, digital forensic investigation is conducted to collect evidence from network/system logs. After investigation and collecting evidence reports are published in multiple formats such as text and PDF. There is no standard format for publishing these reports, so extracting meaningful information from these reports is a challenging task. Manual extraction of features from unstructured cyber-threat intelligence (CTI) is a difficult task. There is a need for an automated mechanism to extract features from unstructured reports and attribute cyber-threat actor (CTA). The aim of this research is to develop a mechanism to attribute or profile cyber threat actors (CTA) by extracting features from CTI reports. Moreover define a methodology to extract features from unstructured CTI reports by using natural language processing (NLP) techniques and then attributing cyber threat actor by using machine learning algorithms. Extracting features i.e., tactics, techniques, tools, malware, target organization/country and application by using novel embedding model known as" Attack2vec" which is trained on domain specific embeddings. Training model on domain specific embedding produces high results as compared to model train on general embeddings specially in the field of cyber security. Results of this novel model is compared with different methods. Machine learning algorithms such as decision tree, random forest, support vector machine is used for classification of CTA. This novel model produces high results as compared to other models with Accuracy of 96%, Precision of 96.4%, Recall of 95.58% and F1-measure of 95.75%.

1. Introduction

Cyber-threat intelligence (CTI) is the knowledge base which include context, behaviour, actions taken and what are the implications of this attack. CTI provides this knowledge base to mitigate against attacks [1 2 3]. It allows organizations to take inform decisions about future attacks and provide an opportunity to access challenges and threats they are facing in cyber space [4 5 6]. Today organizations are focusing on building their own knowledge repository from data available in the world. Based on this knowledge, threat feeds are developed in structured threat information expression (STIX) format [7 8], which is considered standard format in this domain. The CTI life cycle consists of six phases. The first phase is planning and direction. The second phase is the collection of information from different sources. Then third phase is processing of the acquired information. The fourth phase is the analysis of the information. Then fifth phase is dissemination of information. The last phase is the feedback. CTI life cycle is shown in Fig. 1.

Cyber-threat attribution is knowing about the person or organization behind the attack. There are different profiles and various attributes of the attacker [9]. There are also different levels of attribution as shown in Fig. 2. The first level is knowing about the tools, tactics techniques and procedures (TTP) used by the attacker. It is the first step to know about the tools used by the attackers. The second level is to know about the country behind the attack. It tells the motives and objective behind the attack. The third and most important level is knowing about the person who has conducted the attack. It is a very challenging task as attacker uses different

* Corresponding author.
E-mail addresses: ehtsham_irshad@hotmail.com (E. Irshad), abasit.siddiqui@cust.edu.pk (A. Basit Siddiqui).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2022.11.001

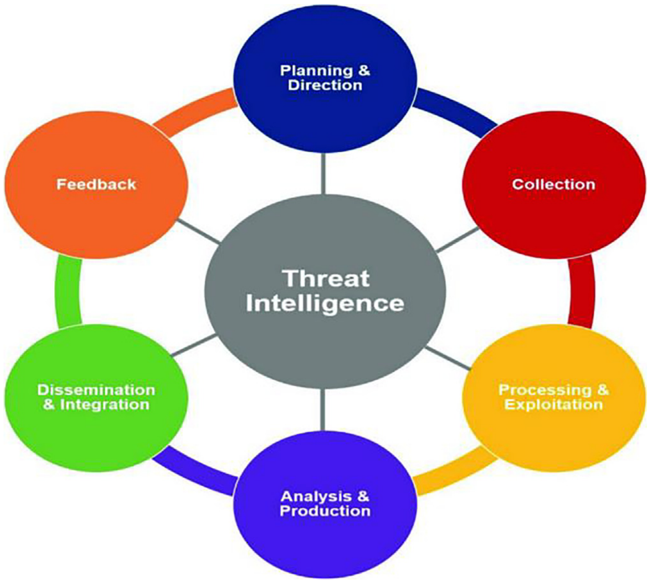


Fig. 1. CTI Life Cycle.

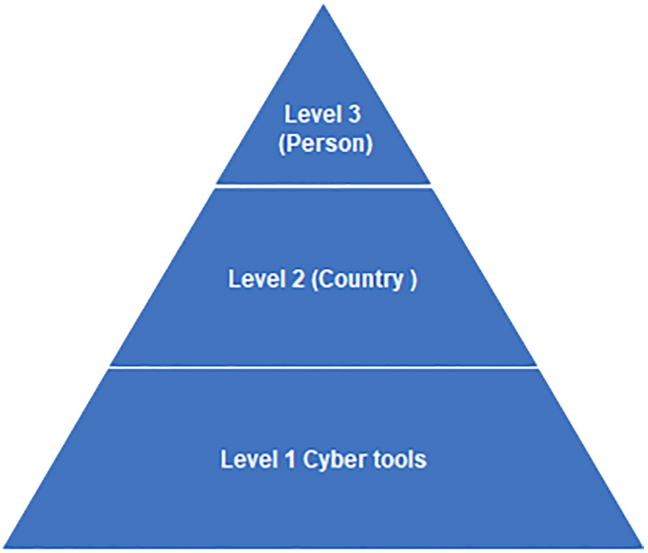


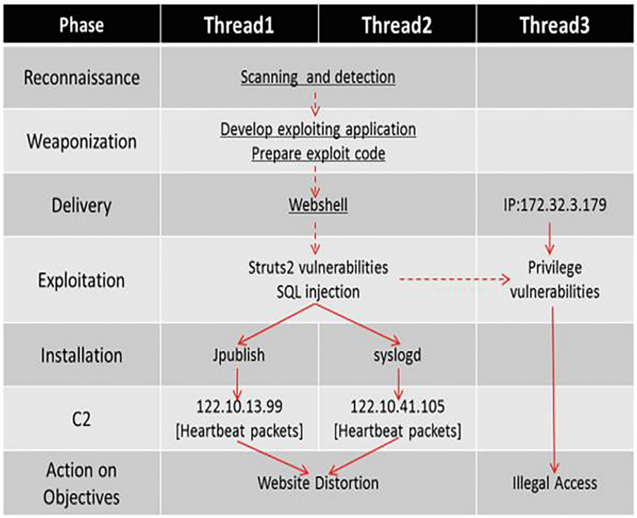
Fig. 2. Levels of Attribution.

deception techniques to hide its identity. Various benchmark frameworks are used in this this research work such as MITRE ATT&CK framework [10 11 12]. MITRE ATT&CK is a knowledge repository for the tactics, techniques procedures and cyber actors on real world attack patterns. It is an open-source repository which is very useful for the research community and industry in utilizing knowledge. It is also providing threat feeds in STIX format for organizations to utilize in security devices. According to MITRE framework, tactics are the objectives which an attacker wants to achieve. Techniques are the mechanism and tools which an attacker use to achieve its objective. Groups are the attacker or cyber threat actor, the perpetrator behind an attack. This framework also provides information how to mitigate these attacks. Its database is continuously updating with the help of research community. Other benchmark framework used in this study are APT Groups and operations [13] and CAPEC [14]. To match target country name, list of country names from Wikipedia [35] is used. For target organizations spglobal document [36] is used. To match list of software's, wikipedia list of software's [37] is used.

Attributing advanced persistent threats (APT) is a difficult task because these attacks remain undected for a longer period. These types of attacks trigger on a specific time. APT are state sponsored attacks, so it makes attribution a challenging task. An important task in cyber-threat attribution is to identify the attack steps used in the attacks by the attackers as shown in Fig. 3. Finding attack steps is an important step, as it tells the whole attack flow. The framework mostly used in cyber-threat attribution before MITRE is cyber kill chain which describes seven phases of an attack (reconnaissance, weaponization, delivery, exploitation, installation, command and control, actions on objectives) [25].

One of the important aspect in this domain is to identify patterns and regularities from the data set [12] as shown in Fig. 4. This pattern identification is helpful in attribution of attacker. It will help in identifying the relationship between different features of the attack. Every attack can provide information about who, what, where why and how it happened. These five parameters tell us about the plan, execution and process involved in the attack as shown in Fig. 5. Who parameter identifies the actual person, organization, and country behind an attack. It is a very important step to know the actual adversary of the attack. This information can be useful to mitigate against future attacks. What parameter tells the overall scope of the attack. It tells what an attacker wants to achieve in this attack. Where parameter identifies the attacker direction. When parameter identifies the timestamp of the attack, at what time the attack has occurred. The why parameter tells the objectives and goals of the attacker. The how parameter identifies the tools, techniques used by the attacker. By knowing the TTP used by the attacker, one can mitigate against the future attacks [60]. Cyber-threat attribution is a complex task as attackers usually take different measures to conceal its identity. By attributing cyber threat actor, organizations can predict the future threats and take preventive measures against these attacks in an effective manner. Effective attribution can obstruct the attack process. Cyber-threat attribution has been done only using limited number of features, this may not provide detailed information about attacker profile. To make precise judgment about attacker, there is a need to focus on the detailed feature set i.e., TTP, tools, malware, target country, target organization, target application. Using of detailed features will improve the cyber-threat attribution process.

Bulk of data is available for CTI, so extracting information of interest is somewhat a challenging task. It is important to identify the sources of data for CTI. The sources of data available are as follows.



Phase	Thread1	Thread2	Thread3
Reconnaissance	Scanning and detection		
Weaponization	Develop exploiting application Prepare exploit code		
Delivery	Webshell		IP:172.32.3.179
Exploitation	Struts2 vulnerabilities SQL injection		Privilege vulnerabilities
Installation	Jpublish	syslogd	
C2	122.10.13.99 [Heartbeat packets]	122.10.41.105 [Heartbeat packets]	
Action on Objectives	Website Distortion		Illegal Access

Fig. 3. Attack Steps.

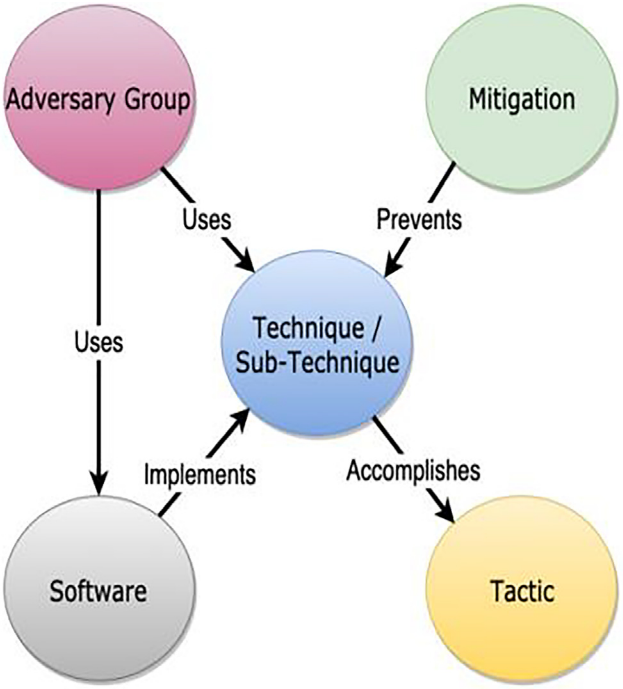


Fig. 4. ATT&CK Model Relationship.

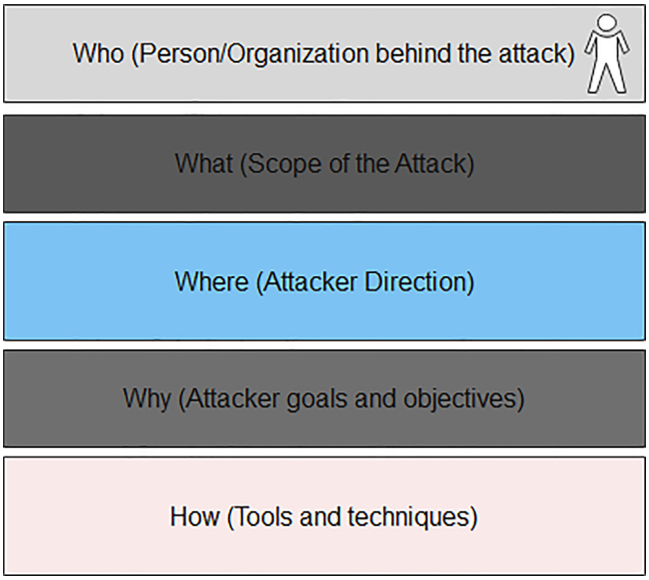

Who (Person/Organization behind the attack)
What (Scope of the Attack)
Where (Attacker Direction)
Why (Attacker goals and objectives)
How (Tools and techniques)

Fig. 5. Parameters of a Cyber-Attack.

- CTI feeds (in STIX format)
- Real Time Data
- Network/Server Logs
- Hacker Forums
- Social Media (Twitter, Facebook etc.)
- Honeypots
- Unstructured CTI Reports
- Common Vulnerabilities and Exposures (CVE)
- National Vulnerability Database (NVD)
- Blogs
- Threat Advisories
- Security Websites
- Clear Web
- Dark Web
- Web Repositories e.g., GitHub

In the proposed study the problem of cyber-threat attribution has been addressed. A framework is proposed for cyber-threat attribution. Features are extracted from unstructured CTI reports to attribute cyber-threat actor. A detailed literature review is conducted to analyse the techniques used for cyber threat attribution. Important aspects in this domain are also highlighted. Extracting meaningful information from these unstructured reports is somewhat difficult task. Detailed feature set i.e., TTP, tools, malware, target country, target organization, target application are extracted from unstructured CTI reports. For extracting features from unstructured CTI reports, a novel embedding model known as “attack2vec” is proposed which is trained on domain specific embeddings, as general embedding model do not produce good results in domain specific fields such as cyber security. So attack2vec model is trained on data sets from field of cyber threat intelligence. Two state of the art embedding model i.e., continuous bag of word (CBOW) model and Skip gram model is used for comparative analysis. Skip gram model produces good results as compared to CBOW. After feature extraction, they are validated against benchmark framework such as MITRE ATT&CK, CAPEC and APT group and operations by using cosine similarity. CTA attribution is done by using classification algorithms such as decision tree, random forest, and support vector machine. Comparative analysis is conducted with state of art models to evaluate the performance of novel model. The comparative analysis showed that proposed model produces good results.

The main contributions of this research work are as follows: -

This proposed study can contribute to more efficient and accurate attribution of cyber-threat actor. First contribution of this proposed study is the development of cyber-threat attribution framework which gives a detailed and accurate profiling mechanism. Furthermore, development of a novel embedding model known as “attack2vec” trained on domain specific datasets is a major contribution of this research work.

Furthermore, inclusion of detailed feature set i.e., TTP, malware, tools, target country, target industry, target application to profile cyber-threat actor is another major contribution of this research work. In earlier research only limited features such as TTP, tools and malware were used. Results showed that inclusion of detailed feature helps in more accurate detection of CTA. Features such as target country, target organization and target application has not been used in the research so far.

The layout of this paper is as follows. Related work is explained in Section 2. Problem statement is in section 3, Objectives and significance is explained in section 4. Proposed methodology is explained in section 5. Experimentation and results are explained in section 6, conclusion, and future work in section 7.

2. Related work

In this research work, proposed a model for cyber-threat attribution. A feature extraction model [32] named similarity-based vector representation (SIMVER) is proposed to extract features from unstructured CTI reports. This model is compared with word2vec and smoothed binary vector (SMOBI). 238 unstructured CTI reports are used. Deep learning models are used for attribution of different cyber threat actors. In this study [15,16], proposed that cyber threat actors use different tools and techniques to conduct attacks on organizations. It is somewhat difficult for attacker to change its tools. Therefore, it is very important to identify attacker from its attack patterns. This will help the organizations to protect against future attacks. Identification of these attack pattern is very helpful especially in fin-tech industry. In this study features are extracted from unstructured CTI reports to attribute cyber-threat actor. In this study [17] proposed that information regarding tactics techniques and procedures (TTP) is mostly present in human readable format. TTP is considered an important feature. As by extracting it from unstructured data, organizations can protect them from future attacks. It forces the attacker to change its tools and techniques, as this is quite a difficult task for attacker. In this research work proposed [18] that threat actions can be extracted from threat related articles. In this method LSI and cosine similarity are used to extract features from the data set. Taxonomy used in this work is MITRE ATT&CK. In this work [19], evaluated different classification approaches to extract features from unstructured text. MITRE ATT&CK is used as a benchmark. A tool named ''Rcatt'' is developed to extract features from unstructured data. This tool generates reports in STIX format which is considered a standard form of representation in this field.

There is lot of raw information present for the cyber threat intelligence [20], extracting this information and convert it into intelligence can be very useful. To extract useful information from raw data and draw patterns, association rule mining technique [34] used. It generates rules to identify different patterns between TTPs. In this work it is explained that anti-malware systems are used to detect a malicious code or activity within the system [21]. It is out of scope of these systems to detect the attacker behind the attack and its intent. A lot of raw data for cyber threat intelligence exists in world today. Manually extracting information from this raw data is nearly impossible because a bulk of raw data exists. Now there is need to design automated mechanism to extract useful information from this data and convert it into intelligence. The studies [22–24] elaborates that the need of information security is increasing with the development of new technologies and infrastructures. Security analysts and experts faces a huge challenge because of increasing security threats. This has led to emergence of a new field called cyber threat intelligence. This field is gaining popularity in the world today and its importance is growing day by day with the increasing number of security threats.

In this study author elaborates that attacker in the modern world have certain level of expertise to conduct cyber-attack in an efficient manner [25]. Attackers use different preventive measures to remain undetected for a longer period. For organizations it a matter of protecting its assets. These attacks can lead to damage their reputation and results in leakage of information. So, for this reason cyber attribution analysis becomes very important and a complex task which requires certain level of expertise. In this research work it is elaborated that there is no fully automatic and online tool available that extracts meaningful and structured information from raw text [26]. In this work" STIXGEN" an online tool for development of structured information in STIX format is proposed. This tool will be helpful for organizations to produce structured information and share meaningful information to share between different organizations. This tool generates structured data in STIX format which will be beneficial especially for the research students. This research work [31] elaborates that there are certain levels of attribution. First level is the host which has started the attack process. Second level is the agent host which has assisted in conducting the attack. Third level is the service provider through which traffic passes. Fourth level is the specific person conducting the attack. Fifth level is the specific organizations and government agencies who has helped in conducting attack. Sixth level is from where the attack originated. Attributing the attacker is a powerful preventive defence against the cyber-attacks. This study [27] elaborates the advantage of domain specific embedding in the field of cyber security. According to the author, by using domain specific embedding in the cyber security field can produce high performance results. In this paper two embedding models are trained to obtain results. One model is trained on 20,000 unstructured cyber threat intelligence reports and second model trained on the web pages crawled from Wikipedia. The results showed that model trained on domain specific embedding produces high results as compared to model train on web pages crawled from Wikipedia. Proposed an automatic extraction method named TTPDrill" [28] which automatically extracts threat related data from unstructured cyber threat intelligence reports. It develops threat feeds in a structured format as STIX. This method customizes some of the NLP techniques and developed a method which can automatically extract threat related data.

This study proposed a novel approach for analysing CTI reports and extracting threat related information from security related corpus [29]. First contribution is the extraction of features from the CTI reports. This study annotates different threat related text used in this domain. A major contribution is the generation of 498,000 tag data set. This study has identified features such as IP hash [30]. Also extracted high level IOC such as tactics and techniques from threat related reports. In this work bias correction methods are used to remove biasness of data collected from different sources. This method is compared with TTPDrill. It outperforms TTPDrill by producing accuracy of 78 %. In the future to extend this work low level IOC are to be extracted from the threat reports. It is the era of digital world [33]. There is huge amount of text data available in the world today. The challenge is to extract useful information from this sparse text. Text classification is the area in which features are extracted from the text and categorize it for normalization of the data. As the data is very sparse and it becomes difficult to handle, so in this regard feature extraction and selection methods are applied. In this work different feature extraction methods as well as machine learning algorithms are reviewed for text classification. In this work [38] a framework is proposed known as IL-CyTIS based on the standard STIX format. This framework is developed by customizing the STIX. The aim of this framework is to extract the threat actions from CTI reports in a more effective way to attribute the cyber threat actor. In the proposed study [39] different threat actions such as tool, industry, file type and organization are extracted from the unstructured CTI reports. Reinforcement algorithms such as BiLSTM-CRF, DTBERT-BiLSTM-CRF, CNN-BiLSTM-CRF are used to evaluate performance measures. In this study [40] threat actions are extracted. Semantic relationship is created between different threat actions to attribute cyber-threat actor. This extraction will help in detecting cyber threat actor more accurately and precisely in the future. Proposed a method [40,41] for identifying malware and extracting threat actions from various datasets such as CTI reports, honeypot, GitHub, NCHC malware knowledge base, GUN open-source project foundation and windows system files. Various machine and deep learning algorithm are used for the classification. CNN is used for classification of malware.

In this work [42] a framework DeLP (Defeasible Logic Programming) is proposed. The aim of this framework is to build a model which can help in attributing cyber-threat more accurately and efficiently. This model will help in cyber threat attribution more efficiently. In this research work [43], an automated mechanism for cyber threat actor is proposed by using TTP that can be extracted from APT reports. Extracted feature is then validated from ATT&CK framework by using cosine similarity. APT reports from 27 different organizations such as US-CERT, CISA, and SOC vendors such as FireEye and Trend-Micro are used. In this study [44], a threat model is proposed for extracting features from cyber-warfare events such as surveillance, data theft, espionage, and misinformation. This model helps in extraction of threat actions. This study [45] proposed a method for extraction of low-level IOC to attribute cyber threat actor. The impact of low-level IOC is low. It is easy for an attacker to change these attributes. e.g., IP spoofing attack can easily be conducted by changing the IP address. Several frameworks such as Chain smith, IOCMiner, Stixgen has been proposed for extraction of threat actions.

In this proposed study [46] a model is proposed for automatic extraction of threat action and then generating TTP. This model extracts threat actions from APT reports. 521 APT reports are used for extracting TTP. BERT-BiLSTM-CRF is used which achieves precision, recall and F1 score of 96 %, 97 %, and 96 %. In this paper [47] a method for cyber-threat intelligence is proposed that used deep learning models which can extract threat actions from Space, Air, Ground, Sea (SAGS) networks. It has three modules, deep pattern extractor, TI-driven detection, and TI-attack type identification technique. In the proposed methods [48,49], study is conducted for cyber-threat attribution. For this purpose, a framework called DLTIF is developed for cyber threat intelligence modelling. The aim is to identify different threat types. For modelling of cyber-threat intelligence an automated framework "DLTIF" [50] is developed which identify threat types. DLTIF is developed for cyber threat intelligence modelling and to identify different threat types. In this study [51] a novel approach for APT attribution is developed. This technique combines two features i.e., code and string feature. Bag of word model is used in this technique to represent vectors. Random forest is used for classification. Dataset used in this work is the collection of CTI reports. This model helps in analysis of network attacks and threat intelligence. This paper [52] proposed a threat modelling technique known as HinCTI. This model extracts high level IOC for CTI data and draw a semantic relationship. This model helps in identification of threat type more accurately and precisely. Comparison with baseline methods is also done to show the performance of this novel model.

Different cyber security platforms [53] offer situational awareness about different parameters such as threat actor and actions. A lot of useful information is present in the dark web. In this approach, Azure Hacker Asset portal is presented to gather CTI data. This approach makes analysis of reports on dark web to collect insight into CTI for more efficient utilization of CTI. Different techniques [54] for cyber-threat attribution has been proposed in the literature by using machine learning techniques. In the proposed method a honeypot is deployed on Amazon web service to collect data of interest. After text pre-processing different machine learning algorithms are used to attribute cyber-threat actors. Out of the used model SVM produces high accuracy of 94.7 %. In this research work [55] analysis of various types of data for cyber-threat intelligence is conducted to protect organization from cyber-attacks. It is important to analyse various type of cyber threat intelligence data. It is now essential to formulate contextual semantic relation in cyber threat text. In this approach a model known as OSIF is developed to analyse CTI unstructured data. CVE data set is used for cyber actor profiling. In this approach [56], a model known as TIMiner is proposed for sharing CTI data gathered from social media. Convolutional neural network (CNN) is used to classify various types of IOC from the data set. This model generates CTI with domain tags. It a challenging task to attribute cyber-threat actor [57]. As attacker mostly conduct attacks behind proxies so it becomes difficult to identify the initiator of the attack. In the proposed technique problem of CTA attribution is identified. Cyber-threat actor is identified from the attack patterns. In this study [58], a literature review is conducted on the techniques that extract useful information from unstructured text. Total of 28,484 articles are collected. From the analysis of collected information, it is identified that most useful keywords in the field of NLP are topic classification, keyword identification and semantic relationship. In this proposed method [59], a system known as feature smith is used to extract features for android malware. This system improves overall accuracy of extracting threat actions. Advanced persistent threats have become a major threat to countries and organizations in the recent past. As it is somewhat difficult for organizations to detect such type of attacks. In this proposed study [63] a model know as triangle model is proposed. This model uses three attributes such as TTP, sector and tools to draw a relationship for attributing cyber-threat actor. The benchmark framework used for drawing relationship is MITRE ATT&CK. The proposed model will help in attribution of cyber-threat actor more accurately.

Challenges

Availability of CTI reports is a major challenge in this domain. As the forensic investigation is conducted by security vendors, privacy of users is kept secret , so availability of these reports publicly is a major challenge. For this reason, results are generated on imbalanced data sets which can affect the accuracy and performance. Different benchmark frameworks are used for validation of features. A unified benchmark may be used for producing accurate and precise results. Another challenge is extraction of useful information, a report contains a lot of irrevalant information and only few sentences in a report contains information about attack patterns, so extracting useful information from these bulk of data is a challenging task. Another challenge is the availability of reliable reports. If reports are biased, or not of reliable vendors, then results may be biased and inaccurate.

Another challenge is the unstructured format of these reports, as there is no standard format. Different security vendors publish reports according to their own formats, even reports published by a security vendor is in different formats. So, it becomes a challenging task to extract meaningful information. A major challenge is huge amount of data available (unstructured reports, blogs, threat advisories, hacker forums, social media, CVE, NVD, dark web), so extracting useful information from this huge amount of data is a challenging task. A challenge in cyber-threat attribution is the design of a fully automated mechanism which can draw a complete automated flow of an attack. It is a challenging task to design a complete automated framework for cyber-threat attribution as attacker goes through different stages and mechanisms to complete an attack. There exist semi-automated mechanisms for describing an attack flow which helps in the attribution of a threat actor. Identification of relationship between different Incident of Compromise such as TTP malware, tools is a quite a challenge in this domain.

Critical analysis

Limited number of features (TTP, tools) are extracted for CTA attribution so far in the research. There are other important attributes such as target organization, target country, target application which may improve the cyber-threat attribution process. This may provide a detailed information about the attacker profile. L. Perry [27] proposed an embedding model known as SMOBI. Data set used for building domain specific embedding model is not enough. U. Noor [15] technique extracted tools and TTP from unstructured reports. The data set reliability cannot be verified. LSI is modified according to the author but is not explained in detail. S. Naveen [32] extracted tools and TTP from unstructured reports. Proposed an embedding model known as SIMVER, after removing some reports used by L. Perry, results are generated on filtered reports. Detailed feature set not used in this work. Various research work generates results on different data sets, so results of different techniques cannot be compared. So far, no fully automated mechanism has been designed for the cyber-threat attribution. There is semi-automated mechanism which design attack flow process. Cyber-threat attribution has been evaluated on different frameworks such as diamond model, CKC, F2T2EA, MITRE framework. There is a need for single benchmark framework, based on which experimentation may be performed for comparative analysis. In these techniques relationship between different TTP is drawn. There is a need to draw relationship between other features such as tools/-software, malware. It will help in attribution and can provide a more detailed relationship between different attributes. Important aspects in this domain are highlighted below such as NLP and machine learning techniques, features, results generated, performance metrics, cyber-threat actors, tools and framework generated used in this domain are highlighted. These aspects are as follows.

Q1. Which Natural Language Processing (NLP) techniques have been effective in this domain?NLP techniques are used for text cleaning, processes such as removal of stop words, punctuation, tokenization, stemming, lemmatization, and extraction of features. Effective techniques used in this domain are frequency based i.e., TF-IDF and context based i.e., latent semantic indexing (LSI) etc. Some novel models have also been developed by the researchers for extraction of features. After literature survey it is identified that term frequency index document frequency (TF-IDF), latent semantic indexing (LSI) and named entity recognition (NER) are the most used models in this domain. TF-IDF is a frequency-based technique that extracts a word on basis of how frequent it is in the corpus. LSI extracts words based on the context. NER extract words and identify person, location, organization from the corpus. SMOBI is a novel technique based on word2Vec. It searches word in the vocabulary with similar embedding. SIMVER is a novel model which use similar words, if word available in the dataset, assign current index in the matrix. The above stated techniques are most used and are considered effective in this domain for extraction of features.

Q2. Which machine/deep learning models have been effective in this domain?

In this research question, machine/deep learning models used in this domain has been highlighted. The effective techniques used in the literature are random forest, deep learning neural network, decision tree, LSTM and SVM . The above models are effective in producing high results. Classification problem is either multi-class or multi-label classification. In multi-class, there are more than two classes in the datasets, while in multi-label classification, there are more than two labels for a feature. To solve this problem in the literature various techniques such as BR-SVM, BR-DT, LP-SVM and LP-Naive Bayes ha been used

Q3. What kind of performance metrics have been used in literature & which metric is most used?The performance metrics mostly used in this domain are accuracy, precision, recall, f-measure, confidence, support, and lift. Precision is the most used and effective metric used in the literature. Other used metrics are recall, f-measure and accuracy. Confidence, support, and lift are metrics for finding regularities and patterns from the datasets.

Q4. Which data sets have been used in the literature?There has been limitation of data sets in this domain. Identifying data sets will be helpful for research community for future research, without the availability of data sets from reliable sources future research will be difficult in this domain. The datasets identified are unstructured CTI reports (in text and PDF format. They are shown in Table 1.

Table 1
Datasets.

Sr. #	Reports	Year
1.	327 Reports	2019
2.	249 Reports	2019
3.	238 Reports	2020
4.	20,630 Reports	2019
5.	Google programmable search engine	2019
6.	160 Reports	2020
7.	227 Reports	2020
8.	18,257 Reports	2018

Q5. Which features have been used in this domain & are considered most important for cyber threat attribution?

There are two types of features commonly used in this domain. High-Level Incident of Compromise (IOC) and Low-Level IOC. Tactics techniques and procedures (TTP), malware and tools are High level IOC. Low-level IOC are IP, URL, hash, domain name, source/destination port, timestamp, and infection type. After literature review, it is evident that TTP is the most used feature in this domain. In the experiments, now researchers are mostly focusing on high level IOC, as their impact is high and everlasting. Identifying them can force the attackers to change their tools which is very difficult task.

Q6. What results have been generated by different techniques?

The results of different techniques are shown in Table 2. For CTA attribution highest results are 86.5 % accuracy, 95.4 % precision, 83.3 % recall and 87.9 % f-measure.**a.** Which ML and Deep learning models outperforms other methods in literature?Deep learning neural networks out-performs machine learning models in this domain. Random forest and Support Vector Machine (SVM) also produces high results for machine learning models.

Q7. Which feature selection techniques have been used in literature?Information Gain has been mostly used in this area. It is the commonly used technique identified for feature selection [15]. Its role is to identify the most effective feature from the data set.

Q8. What are most used benchmark frameworks in the literature & which framework is mostly used?

Different benchmarks frameworks used in the literature have been identified in this research question. MITRE ATT&CK is most used framework in this domain. The purpose of these benchmark frameworks is to validate a feature. CKC is also used by the researchers before development of MITRE framework. CVE database is also used for extraction of features. Threat actor encyclopaedia published in 2020 by Thai-CERT describes the attacker goals and motivations.

Q9. Which cyber-threat actors (CTA) are mostly used in the research?

Cyber-threat actors are the attacker or person behind a cyberattack. In the literature a cyber-threat actor has been used by different names, so aliases are also identified by the researchers. Cyber threat actor mostly used

in the research are APT28, Lazarus, Turla, Oil Rig, APT17, Fin7, APT29, menu Pass, Deep panda, APT1, admin338, Rocket Kitten, APT12, APT16, APT18, APT30, APT 32, APT34, Equation, FIN5, FIN6, Gameredon, Rocket Kitten, CGMAN, Group5, Ke3chang, Lotus Blossom, Magic Hound, Moafee, Winntie, APT3, APT17, APT28, Molerats, Bronze Butler, Carbanak, Cleaver, Dark hotel, Copy Kittens, Dragonfly, Dragon OK, Dust Storm, Fin10, Copy Kittens.

Q10. What are the most important tools, standards, expressions, and information sharing platforms used in the literature?

In this research question, identification of tools, expressions and information sharing platforms used in the literature has been identified. STIX is considered the most used expression for cyber threat intelligence. Trusted Automated Exchange of Indicator information (TAXII) and Open-IOC are the platforms for extracting threat feeds. Open-source intelligence (OSNIT) is also used by the researchers. It is an open-source repository for collecting information about the attacker.

Table 2
Cyber Threat Attribution Results.

Author/Ref	Accuracy	Precision	Recall	F-Measure
S. Naveen [32]	86.5 %	95.4	83.3 %	87.9 %
U. Noor [15]	94 %	92 %	89 %	89 %
L.Perry [27]	58.4 %	55 %	52.4 %	–

Q11. What are the novel frameworks/tools developed in this domain?

In this research question identified the novel frameworks and tools developed. Some of the commonly used frameworks are explained below.STIXGEN: - It is a tool developed for generation of CTI threat feed in a more detailed and comprehensive manner from raw text data. It will ensure sharing and availability of CTI feeds among various organizations.Action-Miner: - The goal of this tool is to extract low level IOC from CTI feeds more efficiently and accurately as compared to other tools.

ATIS: - Automated threat intelligence fusion framework considers different sources such as cuckoo malware database and tries to create intelligence from these data sources. It is a collection tool to collect meaningful information and draw relationship from this data.

Six-gill: - It is a tool used in dark web which collect hacker information from different sources. The aim of this tool is to extract features from the dark web.

TTP-Drill: - This tool extracts threat-action, then convert threat-action into STIX format from unstructured CTI reports.

IOCMiner: - It is a framework of extracting IOC from unstructured text such as twitter.

Feature Smith: - It is a system to generate a feature set for detecting malwares of android platform.

SMOBI: - It is an improved bag of word model. It assigns weights to each entry in the model. Then it finds word in the vocabulary with similar embedding based on cosine similarity.

SIMVER: - It is a way of representing neural embeddings. Use similar words, if word available in the dataset, assign current index in the matrix. It uses skip gram model.

Q12. Which Security Vendor and external sources are mostly used by the researchers?

Identified the security vendors and external sources used in this domain. It is important to know the reliable vendors and sources refereed by the researchers. Symantec, Fire-eye, Crowd strike, Trend Micro are mostly used, while for extraction of threat actions from raw data twitter stream is mostly used.

3. Problem statement

Manual extraction of attack pattern from cyber threat reports is a tedious and challenging task. Cyber-threat attribution has been carried out by extracting technical features i.e., TTP, tools, malware. These features are not accurately identifying and detecting the modern-day sophisticated attackers. These features are not providing detailed information about attacker profile. Due to complex nature of attackers, there is need to include detail feature set which may also include features such as target country, organization, and application. So, there is a need to include these features and analyse their impact on the attribution process.

4. Objectives and significance

Cyber-threat attribution is a challenging task. The aim is to know the attacker who has conducted the attack. It is important step in formulating responses to cyber-attacks. The objective of this research work is to profile or attribute cyber-threat actor from their attack patterns i.e. TTP, tools, techniques, malware, target organization, target country and target application. The aim in this work is to profile the attacker to who has conducted attacks. This attribution helps in finding identity of the attacker, what tools and techniques they had used for the attack, what are the target country/organization and target application. Based on this knowledge, organizations can protect them from future cyber-attacks. It is a challenging task to extract features from unstructured data and convert it into intelligence. This work will help security vendors to attribute or profile cyber threat actor on basis of attack patterns identified from the attacks. Profiling attacker with detailed features will give comprehensive and detailed information about the attacker. The major difference between proposed method and earlier ones is the inclusion of detailed feature set. This will help security analysts to know about the detailed profile of the attacker.

5. Proposed methodology

For cyber-threat attribution a framework is proposed. It consists of three phases i.e., data collection, feature extraction and cyber-threat attribution. These phases are explained in detail below: -.

5.1. Data Collection

In this phase aim is to collect unstructured CTI reports from different sources. The sources of data collection are reports published by research community, security vendors and a google programmable search engine. There is limitation of data sets in this domain. Most of the experiments have been performed on the imbalanced data sets which can affect the results. So, aim of this phase is collect data set to perform experiments. Availability of more data sets will help the future investigations. The data set collected in this phase for cyber threat attribution are shown in Table 3..

5.2. Feature Extraction

Feature extraction phase consists of three steps. First step is text pre-processing phase or known as text cleaning. Second step is feature extraction in which novel embedding model known as attack2vec is trained on domain specific embeddings from different cyber-threat intelligence data sets. The third step is semantic mapping or feature validation phase in which cosine similarity is used to find similarity between different documents and sentences. In this step feature extracted are validated from benchmark frameworks.

5.2.1. Text Pre-processing

The first phase in feature extraction is to clean the text, known as text pre-processing. For text pre-processing processes such as converting text to lower case, removal of stop words, punctuation, special characters, tokenization, lemmatization have been performed. These words may affect the performance of the model, so it is necessary to remove common words and clean the text. For removal stop words NLTK library is used in python. The process of text pre-processing is shown in Fig. 6.

Table 3
Dataset.

Reports	Description
327 Reports [15]	2019
248 Reports [27]	2019
238 Reports [32]	2020
20,630 Reports [27]	2019
18,257 Reports [30]	2018
17,000 Reports [28]	2019
160 Reports [29]	2020
227 Reports [16]	2020
[15]	Google Programmable Search engine (google)

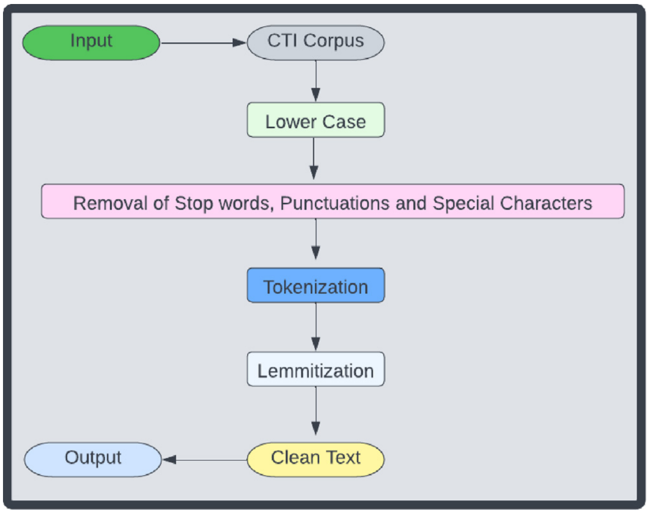


Fig. 6. Text Pre-processing.

5.2.2. Attack2vec embedding model

After cleaning the text, next task is to extract features from unstructured CTI reports. For this, developed a novel embedding model known as "attack2vec". It is based on state-of-the-art word2vec model. According to [61,62] general embedding model do not produce good results in domain specific models such as cyber-security. So, there is a need to build a model that is trained on domain specific embeddings. As word2vec model is trained on Wikipedia pages, that does not produce good results in domain specific fields. To overcome this limitation "attack2vec" is trained on domain specific embeddings. To train attack2vec on domain specific embeddings, datasets from cyber-security fields are collected for training model. Vocabulary size of embedding model is 2 million words. Attack2vec model consists of input layer, hidden and output layer. Weights are assigned. Attack2vec algorithm working is shown below and then model input to neural network is shown in Fig. 7.

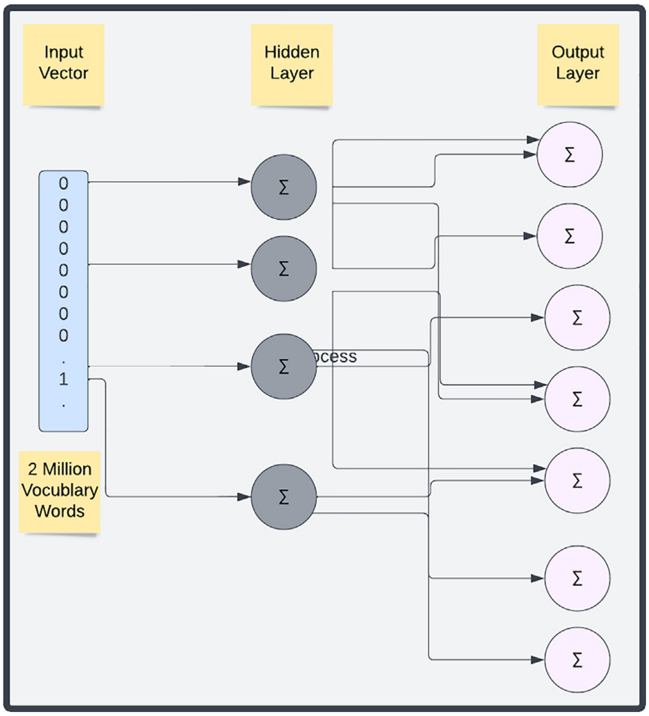


Fig. 7. Attack2vec Neural Network.

The data sets used in training of embedding model are shown below.

a. 20,630 CTI reports [27].
b. 18,257 CTI reports [30].
c. 17,000 CTI reports [28].
d. CVE database [64].
e. Malware Sample reports [51,65–68].

Algorithm: Attack2Vec (Trained on domain specific Corpus)

Input: -
F - CTI Corpus
W_F - Word Corpus, set of words in corpus
W_D - Word Corpus after text pre-processing
Wu – Unique words Vocabulary
Output: -
V: Vector Representation

```
Initialize W_F of size |max| with words from text files F
x: = 0
W_D → NLTK (W_F [max])
For i: = 0 to |max| do
temp: = W_D [i]
If temp: = € S
S_[x]:= temp
x: = x + 1
i → i + 1
V → S
```

There are 2 types of embedding models. Continuous bag of word model (CBOW) and skip gram as shown in Fig. 8. In CBOW model, target word is searched from context words. This model is faster and used for larger data sets. In skip gram model context word is searched from corresponding target words. It is preferred where finding of context is important. In this analysis used both models for testing. Different window size i.e., n = 3, 5 and 7 is used. Corresponding vector size is 100. Skip gram model shows good results as compared to CBOW, so it is preferred over CBOW. After the model is trained on domain specific data sets, the next task is to extract features. The feature set extracted in this work are shown in Table 4..

To know the performance of embeddings model with different machine learning algorithms such as decision tree, random forest & support vector machine, performed various experiments with different window size of n = 3, 5 and 7. When CBOW is used with decision tree, for n = 5 accuracy, precision, recall and f1score of 84 %, 85 %, 83.8 % and 84 % is recorded. When CBOW is used with random forest classifier, for n = 5 we get Accuracy, precision, recall and f1score of 87 %,88.6 %,87.1 % and 87 %. With CBOW with SVM for n = 5, accuracy, precision, recall and f1score of 85 %, 83 %, 85.9 % and 89.5 % is recorded. The performance of skip gram model is relatively better with different machine learning algorithms. When decision tree is implemented, for n = 5 highest accuracy, precision, recall and f1-score of 85 %, 83 %, 85.9 % and 89.5 % is seen. When Skip gram is used with random forest it produces highest results overall. The accuracy, precision, recall and f1-score are 96 %, 96.6 %, 95.58 % and 95.75 % is recorded. When skip gram model is used with SVM, for n = 5 highest accuracy, precision, recall and f1-score of 89 %, 91.2 %, 91.5 % and 90.3 % is recorded. Results of these experimentation is shown in Figs. 9-14.

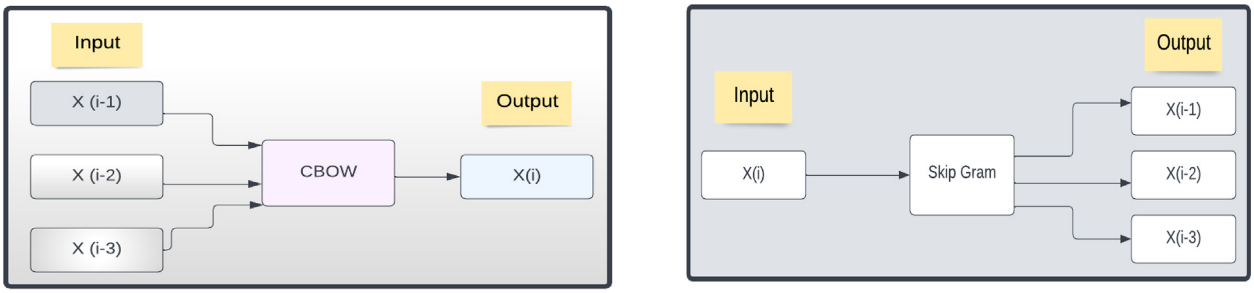


Fig. 8. CBOW & Skip Gram Model.

Table 4
Features Extracted.

Sr. #	Features	Remarks
1	Cyber Threat Actor	Class
2	TTP	High level IOC
3	Malware	High level IOC
4	Tools	High level IOC
5	Target Organization	–
6	Target Country	–
7	Target Application	–

5.2.3. Semantic mapping

The task in this step is to validate extracted features against benchmark frameworks such as MITRE ATT&CK, CAPEC, APT group and Operations, Global Industry Standards, alternative country names from Wikipedia, application names from Wikipedia are used. Extracted feature is validated against benchmarks frameworks. If the feature is validated from benchmark framework, then it is included in the corpus. TTP, malware is validated from MITRE, CAPEC, APT group and operations. The feature target industry is

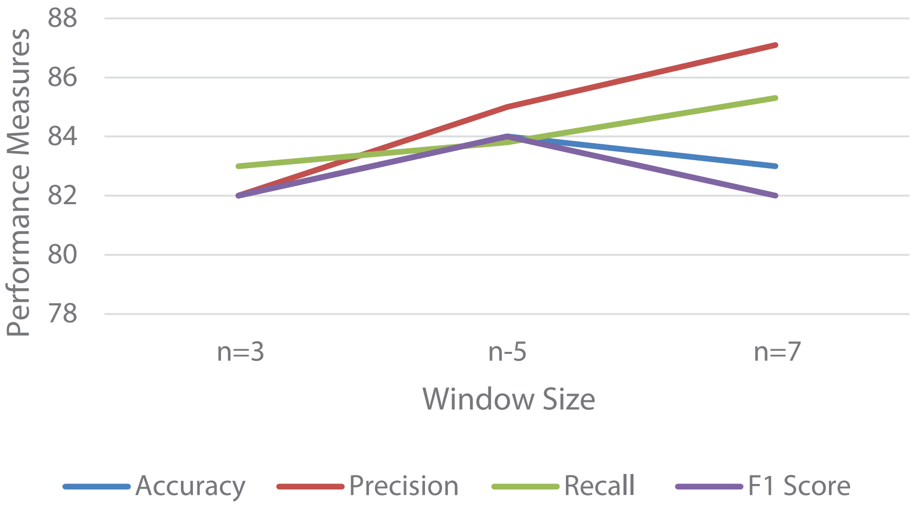


Fig. 9. CBOW with Decision Tree.

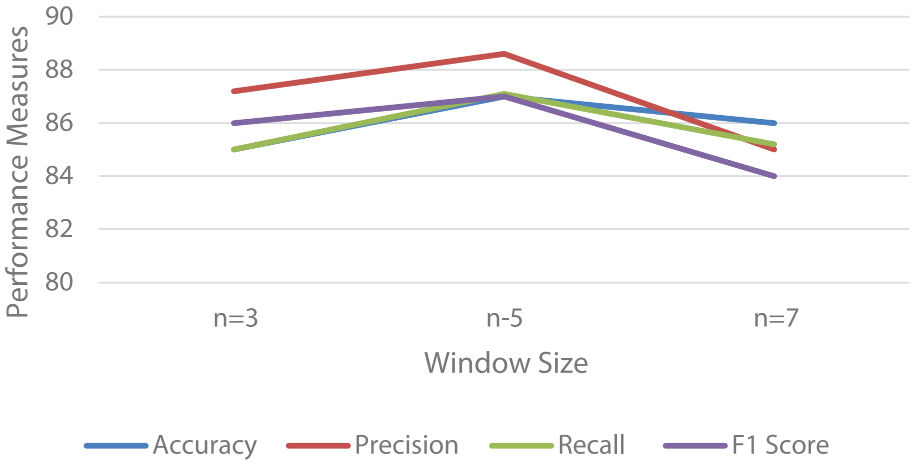


Fig. 10. CBOW with Random Forest.

matched against (https://www.spglobal.com/marketintelligence/en/documents/112727-gics-mapbook2018v3letter digitalspreads.pdf), country names (https://en.wikipedia.org/wiki/List of alternative country names) and software are matched from Wikipedia (https://en.wikipedia.org/wiki/Category:Lists of software). For validating feature set against the corpus, cosine similarity is used. The process is shown in Fig. 15.

5.3. Cyber-Threat attribution

The next phase is the cyber-threat attribution, knowing the actor behind a cyber-attack. In the recent times role of machine learning in the cyber-threat attribution has been increased. The role of machine learning is becoming very important, and it is the need of hour also. As there is a need for development of accu-

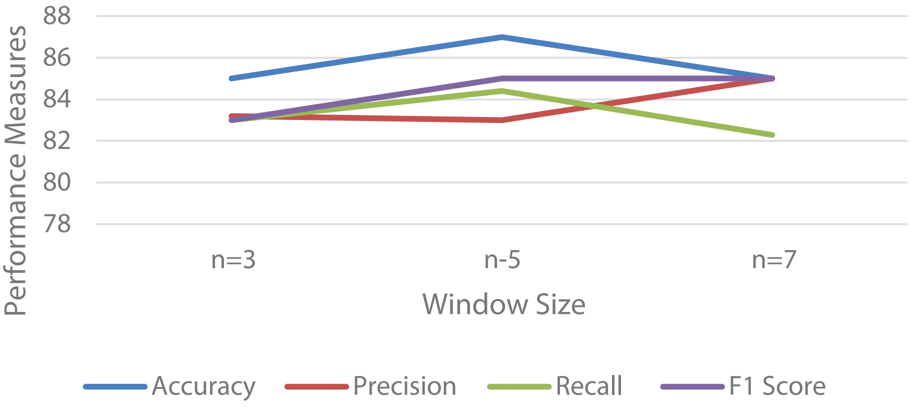


Fig. 11. CBOW with SVM.

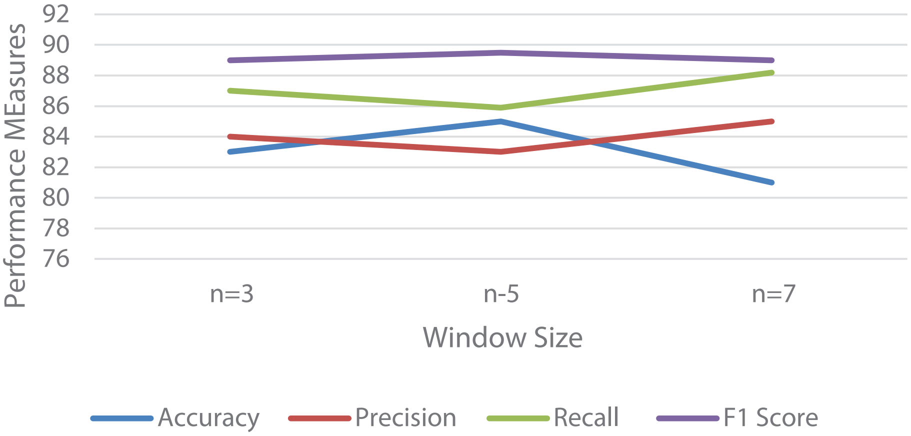


Fig. 12. SKIP gram with Decision Tree.

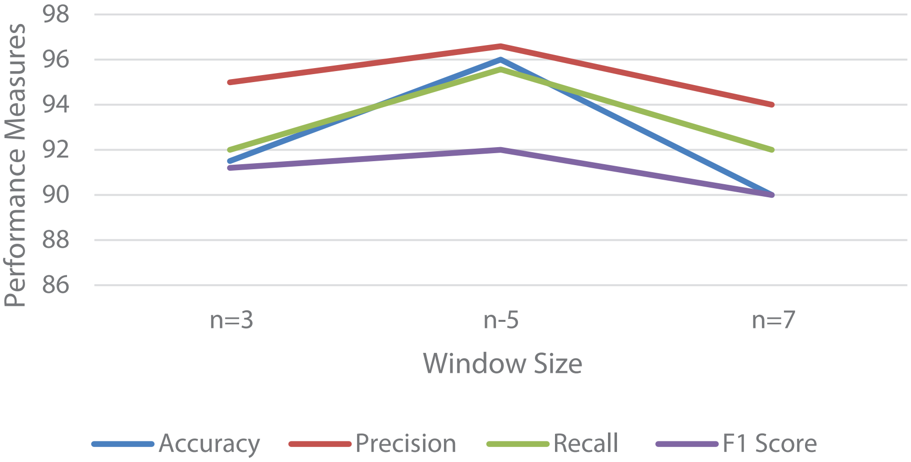


Fig. 13. SKIP gram with Random Forest.

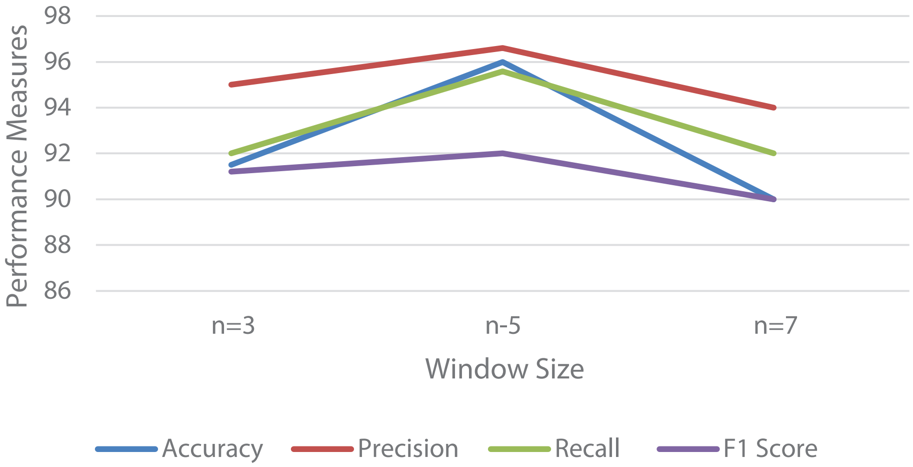


Fig. 14. SKIP gram with SVM.

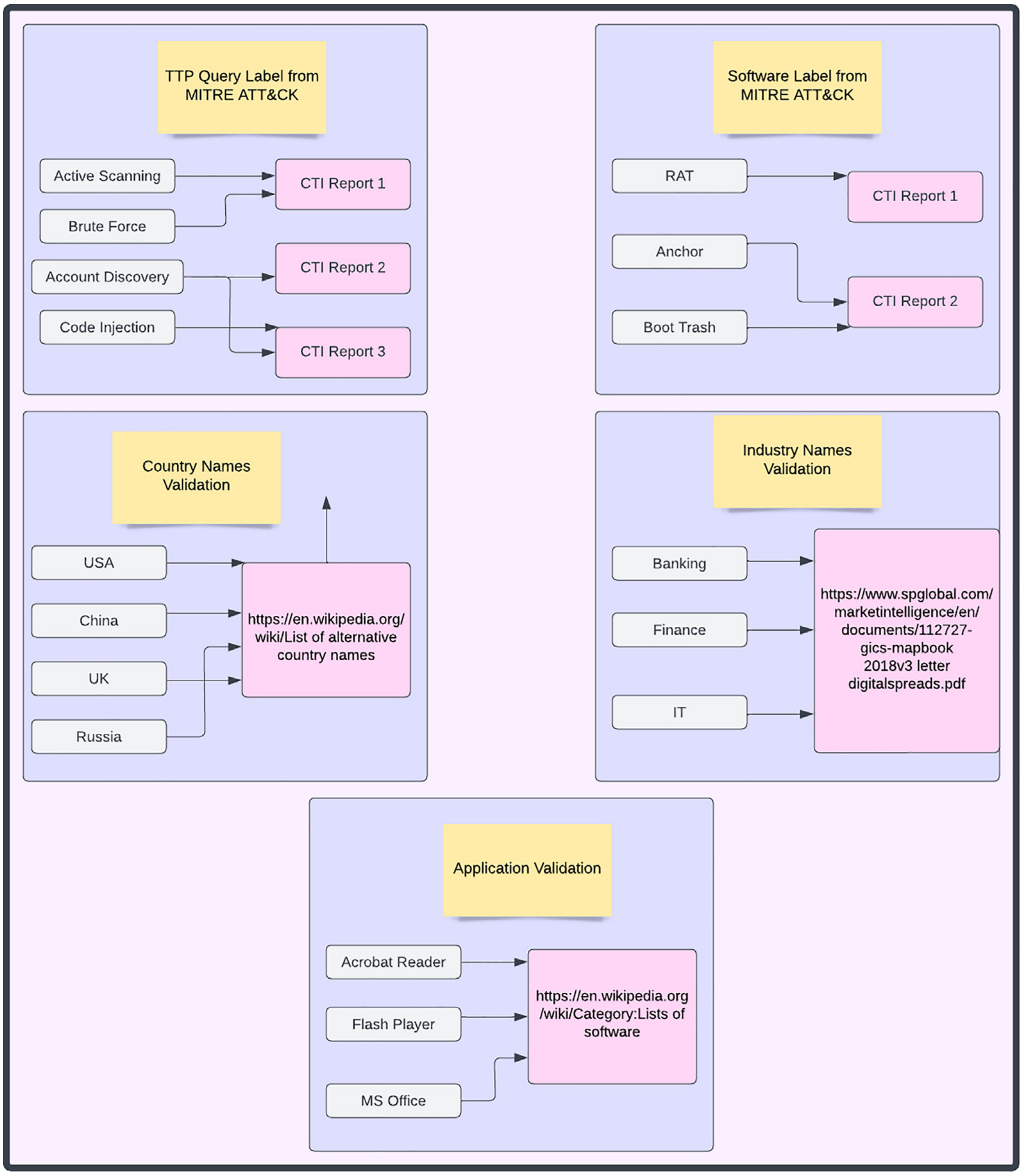

TTP Query Label from MITRE ATT&CK
Active Scanning
Brute Force
Account Discovery
Code Injection
CTI Report 1
CTI Report 2
CTI Report 3
Software Label from MITRE ATT&CK
RAT
Anchor
Boot Trash
CTI Report 1
CTI Report 2
Country Names Validation
USA
China
UK
Russia
https://en.wikipedia.org/wiki/List of alternative country names
Industry Names Validation
Banking
Finance
IT
https://www.spglobal.com/marketintelligence/en/documents/112727-gics-mapbook 2018v3 letter digitalspreads.pdf
Application Validation
Acrobat Reader
Flash Player
MS Office
https://en.wikipedia.org/wiki/Category:Lists of software

Fig. 15. Semantic Mapping.

rate defense techniques to protect organizations from future attacks. To make a thorough judgment on the detection of attacks, it is necessary to diversify a mechanism to detect attacks with greater precision. Despite the considerable efforts of researchers, the cyber-threat attribution continues to face challenges in improving detection accuracy, so there is a need of approaches to better classify attribution. Machine Learning algorithms are used by many researchers to evaluate their performance. Machine Learning algorithms used in three ways, single, hybrid and ensemble. The use of machine learning will help in accurate prediction. In this phase dataset build in the last phase is used for attribution. Classification is a two-step process. First step is the learning, and second stage is the prediction stage. In this phase different machine algorithms are used. This problem is multi-class classification problem as there are more than 2 classes. In this work number of classes are twelve. The classification algorithms used in the proposed study are Decision tree, Random Forest, and Support Vector Machine. These algorithms are selected because they produce good results on textual data. Data flow diagram is shown in Fig. 16. In this flow CTI reports are the input. Then text preprocessing such as removal of stop words, punctuation, lemmatization is done to clean the text. Then features are extracted by using novel embedding model attack2vec. After extraction of features these features are validated from benchmark frameworks such as MITRE ATT&CK, CAPEC, APT group and operations by using cosine similarity. If a feature is matched against benchmark frameworks, then it is included in the corpus otherwise feature is discarded and searching for next word is performed. After feature validation the next step is the CTA attribution. Classification is done to know the actors behind the attack. For classification algorithm the data is divided in to training and testing data. Then CTA attribution is performed.

Decision tree

Decision Tree is one of the simplest classification algorithm. It is well suited for categorial type of data sets. It is a supervised machine learning algorithm which is used to solve classification and regression problems. In this algorithm a tree is made which has two types of node, one is root node and other is leaf node. Prediction starts from the root node. A major challenge in this algorithm is the selection of root node. Different algorithms such as ID3, CHAID, C4.5, CART, MAR can be used in it depending on the classification problem. The logic of this algorithm is easy, so it is easy to understand. It starts from root node and goes down to leaf node for the selection of an attribute. In it for attribute selection measure information gain and Gini index is used. It requires less data cleaning, and it helps in identifying all possible outcomes.

Random forest

It is a supervised machine learning algorithm that is used to solve classification as well as regression problems. This algorithm is constructed from decision tree algorithm. A random forest consists of many decision tree. The forest that is created is trained through different bagging and bootstrap aggregation techniques. This algorithm is more accurate than decision tree. It provides a mechanism to handle missing values. A major advantage of this algorithm over decision tree is that it can solve the problem of overfitting. It takes less time to train the model and it can handle large data sets.

Support vector machine

It is a supervised machine learning algorithm which is used to solve both classification and regression problems. It is suitable for text classification problems. There are two types of this algorithm linear and non-linear SVM.

The proposed framework is shown in Fig. 17. It consists of three phases. In the first phase data is collected in the form of unstructured CTI reports. For comparative analysis the CTI reports referred by S. Naveen are used. These reports are placed in the CTI corpus manager. The second phase is the feature analytics or feature extraction phase. Different text cleaning processes such as shifting to lower case, removal of stop words, punctuations, special characters, tokenization, lemmatization is performed. For removing stop words nltk library is downloaded in the python. Then next step is feature extraction engine phase. In this phase the features are extracted from text by using novel embedding model attack2vev trained on domain specific embeddings. The vocabulary size of

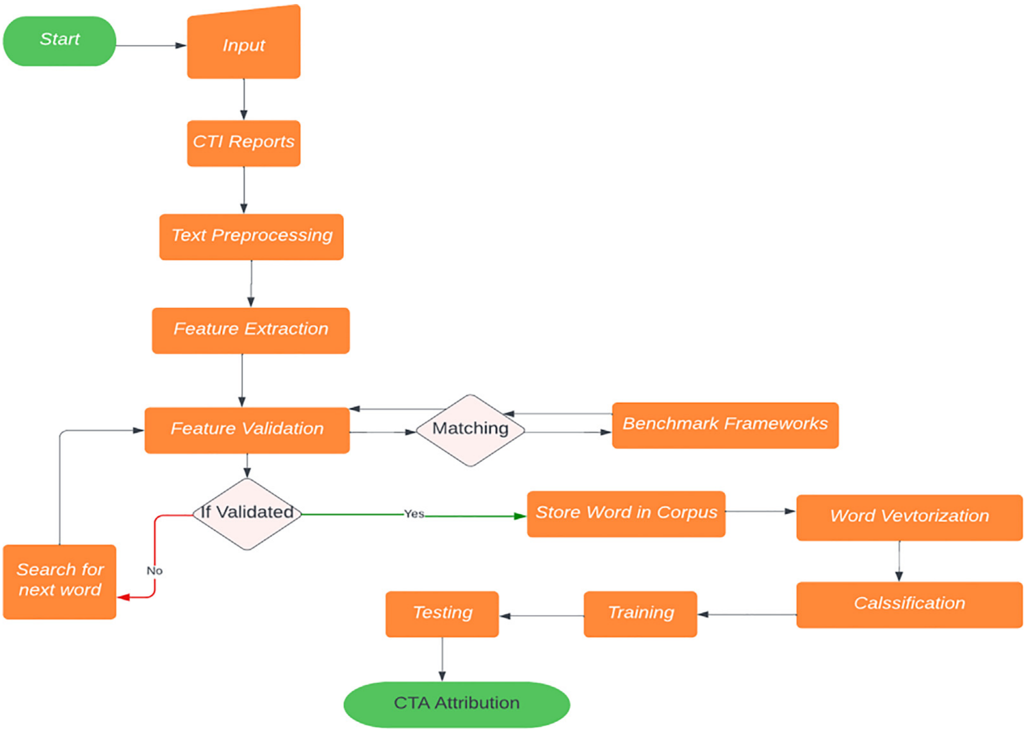
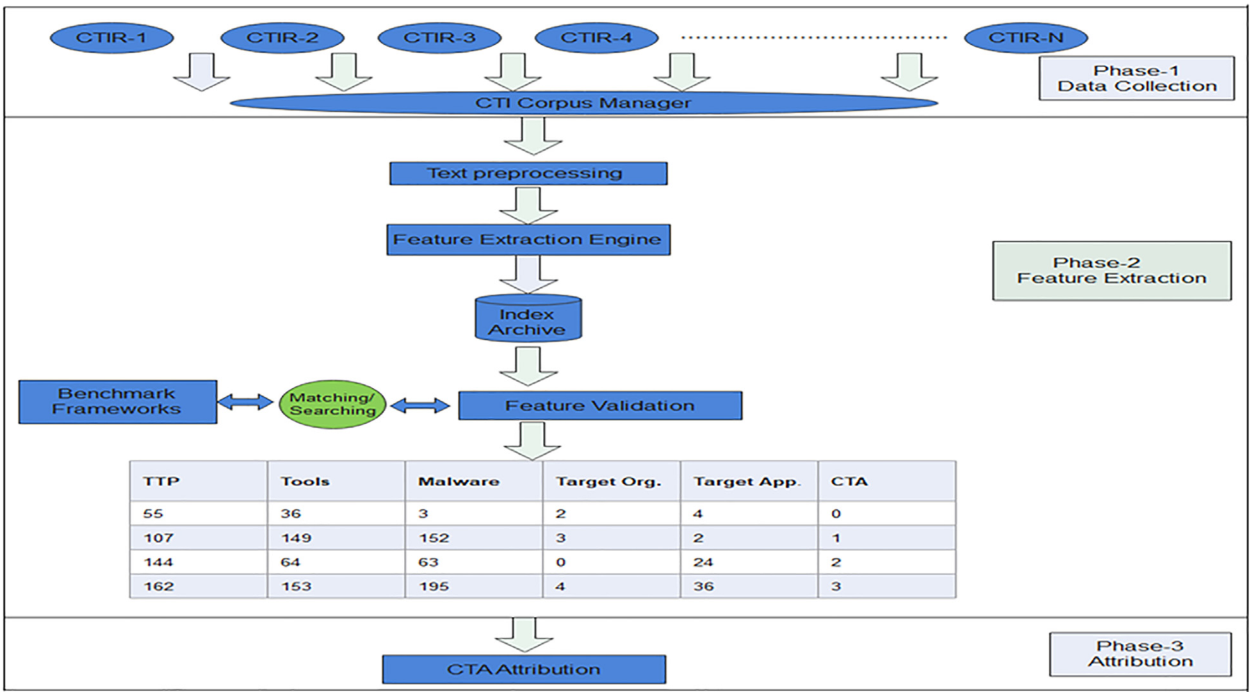


Fig. 16. Data Flow Diagram.



TTP	Tools	Malware	Target Org.	Target App.	CTA
55	36	3	2	4	0
107	149	152	3	2	1
144	64	63	0	24	2
162	153	195	4	36	3

Fig. 17. Proposed Framework.

embedding model is around two million. The extracted features are stored in index archive after using attack2vec model. These extracted features are matched against benchmark frameworks such as MITRE ATT&CK, CAPEC and APT group and operations by using cosine similarity. It is used because it is considered one of the leading method for document classification. If a document contains features its value is high, means it is most similar. If a document contains less matched features, then its value is low. After feature validation, the CTA attribution phase comes in which cyber threat actor is extracted by using different classification algorithms such as decision tree, random forest, and support vector machine. The aim of CTA attribution is to know about the actual attacker behind the attack.

6. Experiments & results

The aim of this research work is to attribute cyber threat actor based on attack patterns extracted from unstructured CTI reports. The evaluation metrics used in this work are accuracy, precision, recall and F1-measure. Accuracy tells us the overall performance

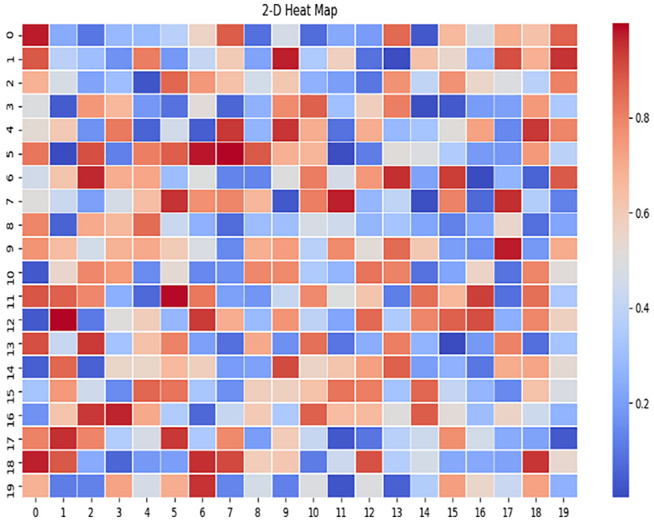
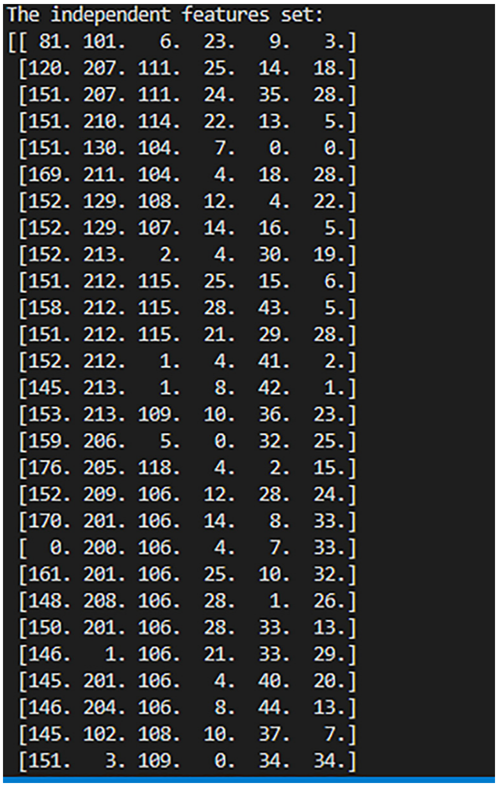
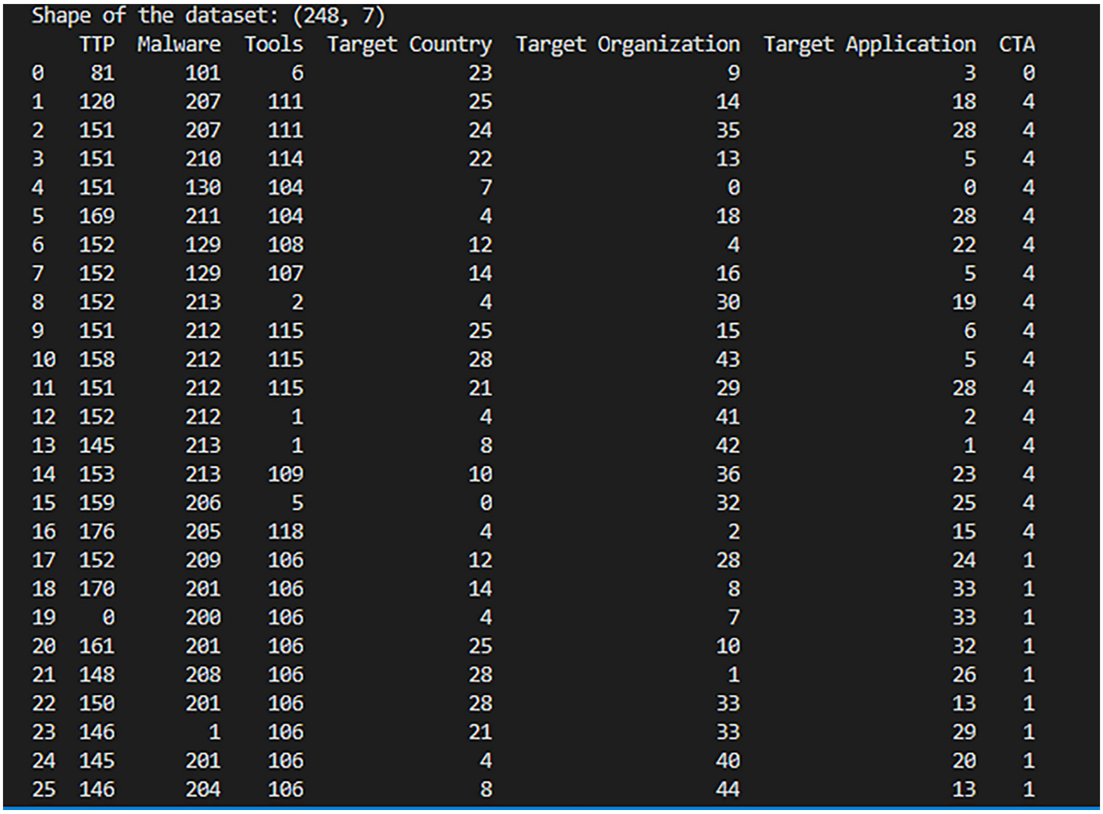


Fig. 18. Heat Map.



```
The independent features set:
[[ 81. 101.   6.  23.   9.   3.]
 [120. 207. 111.  25.  14.  18.]
 [151. 207. 111.  24.  35.  28.]
 [151. 210. 114.  22.  13.   5.]
 [151. 130. 104.   7.   0.   0.]
 [169. 211. 104.   4.  18.  28.]
 [152. 129. 108.  12.   4.  22.]
 [152. 129. 107.  14.  16.   5.]
 [152. 213.   2.   4.  30.  19.]
 [151. 212. 115.  25.  15.   6.]
 [158. 212. 115.  28.  43.   5.]
 [151. 212. 115.  21.  29.  28.]
 [152. 212.   1.   4.  41.   2.]
 [145. 213.   1.   8.  42.   1.]
 [153. 213. 109.  10.  36.  23.]
 [159. 206.   5.   0.  32.  25.]
 [176. 205. 118.   4.   2.  15.]
 [152. 209. 106.  12.  28.  24.]
 [170. 201. 106.  14.   8.  33.]
 [  0. 200. 106.   4.   7.  33.]
 [161. 201. 106.  25.  10.  32.]
 [148. 208. 106.  28.   1.  26.]
 [150. 201. 106.  28.  33.  13.]
 [146.   1. 106.  21.  33.  29.]
 [145. 201. 106.   4.  40.  20.]
 [146. 204. 106.   8.  44.  13.]
 [145. 102. 108.  10.  37.   7.]
 [151.   3. 109.   0.  34.  34.]
```

Fig. 19. Independent and dependent Feature set.



```
Shape of the dataset: (248, 7)
    TTP  Malware  Tools  Target Country  Target Organization  Target Application  CTA
0    81      101      6              23                    9                   3    0
1   120      207    111              25                   14                  18    4
2   151      207    111              24                   35                  28    4
3   151      210    114              22                   13                   5    4
4   151      130    104               7                    0                   0    4
5   169      211    104               4                   18                  28    4
6   152      129    108              12                    4                  22    4
7   152      129    107              14                   16                   5    4
8   152      213      2               4                   30                  19    4
9   151      212    115              25                   15                   6    4
10  158      212    115              28                   43                   5    4
11  151      212    115              21                   29                  28    4
12  152      212      1               4                   41                   2    4
13  145      213      1               8                   42                   1    4
14  153      213    109              10                   36                  23    4
15  159      206      5               0                   32                  25    4
16  176      205    118               4                    2                  15    4
17  152      209    106              12                   28                  24    1
18  170      201    106              14                    8                  33    1
19    0      200    106               4                    7                  33    1
20  161      201    106              25                   10                  32    1
21  148      208    106              28                    1                  26    1
22  150      201    106              28                   33                  13    1
23  146        1    106              21                   33                  29    1
24  145      201    106               4                   40                  20    1
25  146      204    106               8                   44                  13    1
```

Fig. 20. Dataset shape.

of the model. Precision tells us that what percentage of tuples that classifier label as positive is positive. Recall is also known as sensitivity; it tells us that what percentage of relevant results are predicted correctly by the classifier. F1 measure is the harmonic mean of precision and recall. The formulas used to calculate these metrics are shown below. Ratio of training and testing data is 80:20. Implementation of heat map of the model is shown in Fig. 18. Independent and dependent feature sets in python implementation are shown in Fig. 19. The sample shape of data set in python implementation is shown in Fig. 20. In this work cross validation process is used. It divides original data set in to 2 parts. Dataset is divided into k-parts and the value of k = 10 is used. Individual cyber threat actor with number of reports is shown in Table 5.

In this analysis, compared the performance of attack2vec with different models used in the research as shown in Table 6. The per-

Table 5
Cyber Threat Actors.

Group	Aliases	No. of Reports
APT3	Gothic Panda, Pirpi	16
APT17	Deputy Dog, Axiom	16
APT28	Sednit, Fancy Bear	55
APT29	Euroapt, CozyBear	14
Deep Panda	Shell Crew, Webmasters	10
Fin7	Fin7	16
Lazarus	Hidden Cobra, Zinc	34
menu Pass	APT10, Hogfish	12
Oilrig	APT34, Crambus	19
Rocket Kitten	Shamoon, Magic Hound	13
Turla	Krypton, Iron hunter	22
Winntie	Blackfly	11
	Total	**238**

Table 6
Different Models Performance.

Model	Accuracy	Precision	Recall	F1-Measure
SMOBI	54.4 %	63.3 %	50.8 %	53.4 %
Word2vec	84.9 %	90.9 %	82.3 %	85.3 %
SIMVER	86.5 %	95.4 %	83.3 %	87.9 %
Modified LSI	94 %	92 %	89 %	89 %
Attack2Vec	96 %	96.6 %	95.8 %	95.75 %

Table 7
Performance of Each Threat Actor.

Class	Actor	Precision	Recall	F1-Measure
0	APT3	100 %	100 %	100 %
1	APT17	100 %	100 %	100 %
2	APT28	100 %	100 %	100 %
3	APT29	100 %	100 %	100 %
4	Deep Panda	100 %	100 %	100 %
5	Fin7	100 %	100 %	100 %
6	Lazarus	100 %	100 %	100 %
7	Menu pass	100 %	100 %	100 %
8	Oilrig	100 %	100 %	100 %
9	Rocket kitten	80 %	100 %	89 %
10	Turla	80 %	80 %	80 %
11	Winntie	100 %	67 %	80 %

Table 8
Results.

Algorithm	Accuracy	Precision	Recall	F1- Score
Random Forest	96 %	96.6 %	95.58 %	95.75 %
Decision Tree	81 %	84 %	81 %	83 %
SVM	81 %	84 %	81 %	82 %

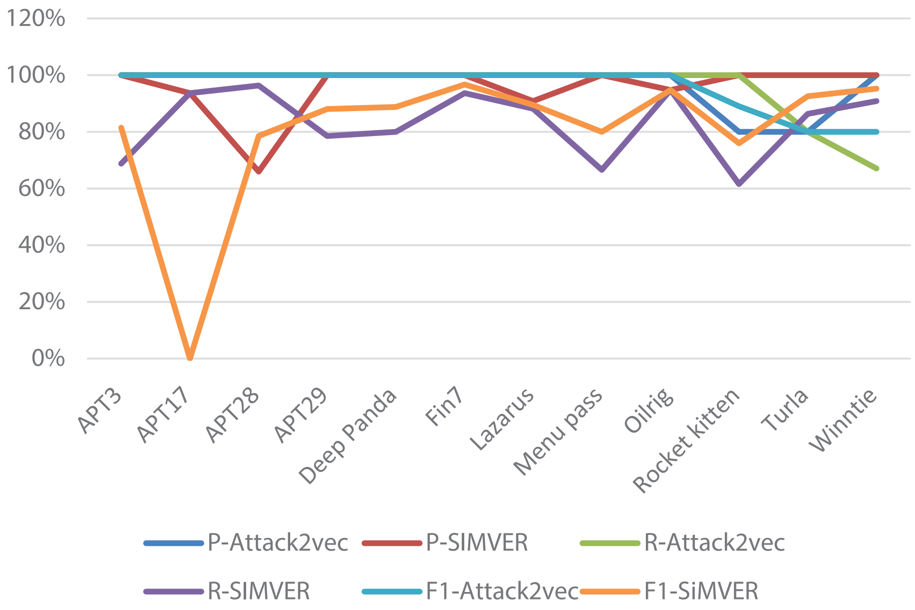


Fig. 21. Individual CTA Results (Attack2Vec vs SIMVER).

formance parameters used for comparison are accuracy, precision, recall and F1-score. The models used for comparative analysis are SMOBI, Word2vec, SIMVERmodified LSI. From the results it is clear attack2vec model outperforms other models.

In this research work different classifiers are to attribute CTA. Classifiers used are random forest, decision tree and support vector machine. Out of three classifiers used random forest has maximum (Tables 7 and 8) accuracy, precision, recall and F1-measure of 96 %, 96.6 %, 95.58 % and 95.75 %. The individual performance of threat actors shows that most of the threat actors are predicted accurately. APT3, APT17, APT28, APT29, Deep-panda, Fin7, Lazarus, Menu pass, Oilrig has 100 % precision, recall and F1 measure. Three classifiers i.e., Rocket kitten has 80 % precision, 100 % recall and 89 % F1 measure, while Turla threat actor has 80 % precision, 80 % recall and 80 % F1 measure, while Winntie threat actor has 100 % precision, 67 % recall and 80 % F1 measure. These results show that using detailed feature in attributing cyber threat actor improves the overall performance of the model. The test bed for implementation is python. The system used for implementation is Core I-5 with 8 GB RAM. The results are shown in Table 6. In Fig. 21 individual result comparison of SIMVER vs attack2vec is shown. The figure shows that attack2vec outperforms SIMVER and the performance measures (Accuracy, Precision, Recall and F1-score) of attack2vec is higher than SIMVER embedding model for individual threat actors.

7. Conclusion & future guidelines

There is lot of CTI information available in the world. Sources include threat feeds, hacker forums, social media, dark web, security websites, threat advisories, honeypots, CVE, NVD and unstructured CTI reports. Unstructured CTI reports contains a lot of useful information, it is considered one of the most useful source of CTI information. Manual extraction of meaningful information from this data is a challenging task. In this research work attributed CTA by extracting features from unstructured CTI reports. Detailed feature set i.e., TTP, tools, malware, target organization, country and application have been used. This detailed feature set has provided a comprehensive overview of attacker profile and it helped in attribution more accurately and precisely. A novel model "Attack2vec" has been developed, which is trained on domain specific embeddings. Results showed that this novel model produces high results as compared to other models. This novel model achieves accuracy, precision, recall and F1-score of 96 %, 96.5 %, 95.58 % and 95.75 % respectively. Machine learning models such as decision tree, random forest, support vector machine has been used for classification. Using detailed feature set has improved the classification results.

This work has certain limitations. There is limitation of data sets in this domain. It is difficult to find data sets for experimentation in this field. There is no standard format of reports (unstructured reports), so it is difficult to extract useful information. Imbalanced Data sets. No benchmark data set in this field. Different research work is done on different data sets, so it is difficult to compare different techniques with each other. It is difficult to extract all features from a report, there can be missing values.

Future work in this area is to include behavioural features for cyber threat actors. The inclusion of behavioural feature might improve attribution of CTA. The major challenge in using behavioural attributes is the availability of datasets. How to extract behavioural attributes is still a major challenge in this domain. Future aim to find the optimal features for cyber-threat attribution by using hybrid features i.e., technical and behavioural features.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Wagner TD, Mahbub K, Palomar E, Abdallah AE. Cyber threat intelligence sharing: Survey and research directions. Comput Secur 2019;87:101589.

[2] Wagner TD, Palomar E, Mahbub K, Abdallah AE. "A novel trust taxonomy for shared cyber threat intelligence", Security and ommunication. Networks 2018;2018.

[3] V. Mavroeidis, S. Bromander," Cyber threat intelligence model: an evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence," in 2017 European Intelligence and Security Informatics Conference (EISIC). IEEE, 2017, pp. 91-98..

[4] Conti M, Dargahi T, Dehghantanha A. Cyber threat intelligence: challenges and opportunities. In: Cyber Threat Intelligence. Springer; 2018. p. 1–6.

[5] Gartner, 2021 Gartner," https://www.gartner.com, 2021.
[6] Brown R, Lee RM. The evolution of cyber threat intelligence (cti)": 2019 sans cti survey. SANS Institute: Singapore; 2019.
[7] Mitre," The Mitre corporation," https://stixproject.github.io, 2018.
[8] E. W. Burger, M. D. Goodman, P. Kampanakis, K. A. Zhu," Taxonomy model for cyber threat intelligence information exchange technologies," in Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security, 2014, pp. 51-60.
[9] Doynikova E, Novikova E, Kotenko I. Attacker behaviour forecasting using methods of intelligent data analysis: A comparative review and prospects". Information 2020;11(3):168.
[10] M. ATT&CK," Mitre", https://attack.mitre.org, 2021.
[11] Al-Shaer R, Spring JM, Christou E. Learning the associations of mitre att&ck adversarial techniques". In: In 2020 IEEE Conference on Communications and Network Security (CNS). IEEE; 2020. p. 1–9.
[12] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, C. B. Thomas," Mitre att&ck: Design and philosophy," Technical report, 2018.
[13] " Apt group and operations", https://airtable.com, 2018.
[14] C. Corporation," Mitre att&ck," https://capec.mitre.org, 2021.
[15] Noor U, Anwar Z, Amjad T, Choo K-K R. A machine learning based fintech cyber threat attribution framework using high-level indicators of compromise. Futur Gener Comput Syst 2019;96:227–42.
[16] C. Gossi," Identifying attackers by using machine learning on unstructured cyber threat intelligence," 2020.
[17] Landauer M, Skopik F, Wurzenberger M, Hotwagner W, Rauber A. A framework for cyber threat intelligence extraction from raw log data. In: In 2019 IEEE International Conference on Big Data (Big Data). IEEE; 2019. p. 3200–9.
[18] Li M, Zheng R, Liu L, Yang P. Extraction of threat actions from threat-related articles using multi-label machine learning classification method. In: in 2019 2nd International Conference on Safety Produce Informatization (IICSPI). IEEE; 2019. p. 428–31.
[19] V. Legoy, M. Caselli, C. Seifert, A. Peter," Automated retrieval of att&ck tactics and techniques for cyber threat reports," arXiv preprint arXiv:2004.14322; 2020.
[20] M. S. Abu, A. Ariffin, S. R. Selamat, R. Yusof," An attribution of cyberattack using association rule mining (arm).".
[21] Modi A, Sun Z, Panwar A, Khairnar T, Zhao Z, Doupe̒ A, Ahn G-J, Black P. Towards automated threat intelligence fusion. In: in 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC). IEEE; 2016. p. 408–16.
[22] R. R. Ramnani, K. Shivaram, and S. Sengupta," Semi-automated information extraction from unstructured threat advisories," in Proceedings of the 10th Innovations in Software Engineering Conference, 2017, pp. 181-187.
[23] Ghazi Y, Anwar Z, Mumtaz R, Saleem S, Tahir A. A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources. In: in 2018 International Conference on Frontiers of Information Technology (FIT). IEEE; 2018. p. 129–34.
[24] Niakanlahiji A, Safarnejad L, Harper R, Chu B-T. IOCMiner: Automatic extraction of indicators of compromise from twitter. In: in 2019 IEEE International Conference on Big Data (Big Data). IEEE; 2019. p. 4747–54.
[25] Qiang L, Zeming Y, Baoxu L, Zhengwei J, Jian Y. Framework of cyber-attack attribution based on threat intelligence". In: Interoperability, Safety and Security in IoT. Springer; 2016. p. 92–103.
[26] Iqbal Z, Anwar Z, Mumtaz R. Stixgen-a novel framework for automatic generation of structured cyber threat information. In: in 2018 International Conference on Frontiers of Information Technology (FIT). IEEE; 2018. p. 241–326.
[27] Perry L, Shapira B, Puzis R. No-doubt: Attack attribution based on threat intelligence reports. In: in 2019 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE; 2019. p. 80–5.
[28] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, X. Niu," TTPDrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources," in Proceedings of the 33rd Annual Computer Security Applications Conference, 2017, pp. 103-115.
[29] Kim G, Lee C, Jo J, Lim H. Automatic extraction of named entities of cyber threats using a deep bi-lstm-crf network". Int J Mach Learn Cybern 2020;11 (10):2341–3255.
[30] Ayoade G, Chandra S, Khan L, Hamlen K, Thuraisingham B. Automated threat report classification over multi-source data". In: in 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC). IEEE; 2018. p. 236–45.
[31] Qiang L, Ze-Ming Y, Bao-Xu L, Zheng-Wei J. A reasoning method of cyber-attack attribution based on threat intelligence. Int J Comput Syst Eng 2016;10 (5):920–4.
[32] Naveen S, Puzis R, Angappan K. Deep learning for threat actor attribution from threat reports". In: in 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP). IEEE; 2020. p. 1–6.
[33] F. P. Shah, V. Patel," A review on feature selection and feature extraction for text classification," in 2016 international conference on wireless communications, signal processing and networking (WiSPNET). IEEE, 2016, pp. 2264-2268.
[34] Noor U, Anwar Z, Rashid Z. An association rule mining-based framework for profiling regularities in tactics techniques and procedures of cyber threat actors. In: in 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE). IEEE; 2018. p. 1–6.
[35] Wikipedia, The Free Encyclopedia.Online: https: https://en.wikipedia.org/wiki/List of alternative country names.
[36] https://www.spglobal.com/marketintelligence/en/documents/112727-gics-mapbook 2018 v3 letter digitalspreads.pdf.
[37] https://en.wikipedia.org/wiki/Category:Lists of software.
[38] Kim, Keecheon and an, Jung Hyun and Yoo, Joseph, "A design of IL-CyTIS for automated cyber threat detection", 2018 International Conference on Information Networking (ICOIN), IEEE, pages 689-693, 2018.
[39] Wang, Xuren and Chen, Rong and Song, Binghua and Yang, Jie, and Jiang, Zhengwei and Zhang, Xiaoqing, and Li, Xiaomeng and Ao, Shengqin, "A Method for Extracting Unstructured Threat Intelligence Based on Dictionary Template and Reinforcement Learning", 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, pages 262-267, 2021.
[40] Kim BI, Kim N, Lee S, Cho H, Park J. A study on a cyber threat intelligence analysis (CTI) platform for the proactive detection of cyber-attacks based on automated analysis. 2018 International Conference on Platform Technology and Service (PlatCon). IEEE; 2018.
[41] Settanni, Giuseppe and Shovgenya, Yegor and Skopik, Florian and Graf, Roman and Wurzenberger, Markus and Fiedler, Roman, "Acquiring cyber threat intelligence through security information correlation" 2017 3rd IEEE International Conference on Cybernetics (CYBCONF)", IEEE, pages=1-7, 2017.
[42] Chen, Chia-Mei and Wang, Shi-Hao and Wen, Dan-Wei, and Lai, Gu-Hsin and Sun, Ming-Kung, "Applying convolutional neural network for malware detection", 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), IEEE, pages 1-5, 2019.
[43] Nunes, Eric and Shakarian, Paulo, and Simari, Gerardo I and Ruef, Andrew, "Argumentation models for cyber-attribution", 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE pages 837-844, 2016.
[44] Shin Y, Kim K, Lee JJ, Lee K. "ART. IEEE; 2021. p. 15–20. Automated Reclassification for Threat Actors based on ATT&CK Matrix Similarity", 2021 World Automation Congress (WAC.
[45] Goel S, Nussbaum B. Attribution Across Cyber Attack Types: Network Intrusions and Information Operations. IEEE Open Journal of the Communications Society, IEEE 2021:1082–93.
[46] Wang, Tianyi and Chow, Kam Pu, "Automatic tagging of cyber threat intelligence unstructured data using semantics extraction", 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), IEEE, pages 197-199, 2019.
[47] Zongxun, Li and Yujun, Li and Haojie, Zhang and Juan, Li, "Construction of TTPS From APT Reports Using Bert", 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pages-260–263, 2021.
[48] Al-Hawawreh M, Moustafa N, Garg S, Hossain MS. Deep learning-enabled threat intelligence scheme in the Internet of Things networks. IEEE Transactions on Network Science and Engineering, IEEE 2020:2968–81.
[49] Jaafar, Fehmi and Avellaneda, Florent and Alikacem, El-Hackemi, "Demystifying the cyber attribution: An exploratory study, 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress IDASC/PiCom/CBDCom/Cyber SciTech)", IEEE, pages 35-40,2020.
[50] Kumar P, Gupta GP, Tripathi R, Garg S, Hassan MM. DLTIF: Deep Learning-Driven Cyber Threat Intelligence Modelling and Identification Framework in IoT-Enabled Maritime Transportation Systems. IEEE Transactions on Intelligent Transportation Systems, IEEE 2021.
[51] Wang, Qinqin and Yan, Hanbing and Han, Zhihui, "Explainable APT Attribution for Malware Using NLP Techniques", 2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS), IEEE, 70-80, 2021.
[52] Gao Y, Xiaoyong LI, Hao PENG, Fang B, Yu P. HinCTI: A cyber threat intelligence modelling and identification system based on heterogeneous information network. IEEE Transactions on Knowledge and Data Engineering, IEEE 2020.
[53] Samtani, Sagar, and Li, Weifeng and Benjamin, Victor, and Chen, Hsinchun, "Informing Cyber Threat Intelligence through Dark Web Situational Awareness: The AZSecure Hacker Assets Portal", Digital Threats: Research and Practice (DTRAP), ACM, pages 1-10, year 2021.
[54] Alkaabi, Maimonah and Olatunji, Sunday, "Modeling Cyber-Attribution Using Machine Learning Techniques", 2020 30th International Conference on Computer Theory and Applications (ICCTA), IEEE, pages 10-15,2020.
[55] Li, Ke and Wen, Hui and Li, Hong and Zhu, Hong song and Sun, Limin, "Security OSIF: Toward automatic discovery and analysis of event based cyber threat intelligence", 2018 IEEE Smart World, Ubiquitous Intelligence \& Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation Smart World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), IEEE, pages 741-747, 2018.
[56] Zhao J. and Yan, Qiben and Li, Jianxin and Shao, Minglai and He, Zuti and Li, Bo, "TIMiner: Automatically extracting and analysing categorized cyber threat intelligence from social data". Computers & Security: Elsevier; 2020.
[57] Avellaneda, Florent and Alikacem, El-Hackemi and Jaafar, Femi, "Using Attack Pattern for Cyber Attack Attribution", 2019 International Conference on Cybersecurity (ICoCSec), IEEE, pages 1-6, 2019.
[58] Rahman, Md Rayhanur and Mahdavi-Hezaveh, Rezvan and Williams, Laurie" A Literature Review on Mining Cyber-threat Intelligence from Unstructured

Texts", International Conference on Data Mining Workshops (ICDMW), IEEE, pages 516—525, 2020.

[59] Zhu, Ziyun and Dumitra, Tudor, "Feature smith: Automatically engineering features for malware detection by mining the security literature", Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pages 767-778, 2016.

[60] Mavroeidis, Vasileios and Hohimer, Ryan and Casey, Tim and Jesang, Audun, "Threat Actor Type Inference and Characterization within Cyber Threat Intelligence", 2021 13th International Conference on Cyber Conflict (CyCon), pages 327-352,2021, IEEE.

[61] Roy, Arpita and Park, Youngja and Pan, Shimei, "Learning domain-specific word embeddings from sparse cybersecurity texts", arXiv preprint arXiv:1709.07470, 2017.

[62] Mumtaz S, Rodriguez C, Benatallah B, Al-Banna M, Zamanirad S. Learning word representation for the cyber security vulnerability domain. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE; 2020. p. 1–8.

[63] Warikoo A. The triangle model for cyber threat attribution. Journal of Cyber Security Technology 2021:118.

[64] Mitre," The Mitre corporation," https://cve.mitre.org, 2022.

[65] Rosenberg I, Sicard G, David EO. Deep-apt: nation-state apt attribution using end-to-end deep neural networks". In: International Conference on Artificial Neural Networks. Springer; 2017. p. 91–9.

[66] A. Roy, Y. Park, and S. Pan," Learning domain specific word embeddings from sparse cyber security texts," arXiv preprint arXiv: 1709.07470, 2017.

[67] Husari G, Niu Xi, Chu B, AlShaer, Ehab. Using entropy and mutual information to extract threat actions from cyber threat intelligence". In: 2018 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE; 2018. p. 16.

[68] Chen, Chia-Mei and Wang, Shi-Hao Wen, Dan-Wei, and Lai, Gu-Hsin and Sun, Ming-Kung, "Applying convolutional neural network for malware detection", 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), IEEE, pages 1-5, 2019.