A Literature Review on Mining Cyberthreat Intelligence from Unstructured Texts

Md Rayhanur Rahman*, Rezvan Mahdavi-Hezaveh†, Laurie Williams‡
{mrahman*, rmahdav†, lawilli3‡}@ncsu.edu
North Carolina State University, Raleigh, NC, USA

***Abstract*—Cyberthreat defense mechanisms have become more proactive these days, and thus leading to the increasing incorporation of cyberthreat intelligence (CTI). Cybersecurity researchers and vendors are powering the CTI with large volumes of unstructured textual data containing information on threat events, threat techniques, and tactics. Hence, extracting cyberthreat-relevant information through text mining is an effective way to obtain actionable CTI to thwart cyberattacks. The goal of this research is *to aid cybersecurity researchers understand the source, purpose, and approaches for mining cyberthreat intelligence from unstructured text through a literature review of peer-reviewed studies on this topic.* We perform a literature review to identify and analyze existing research on mining CTI. By using search queries in the bibliographic databases, 28,484 articles are found. From those, 38 studies are identified through the filtering criteria which include removing duplicates, non-English, non-peer-reviewed articles, and articles not about mining CTI. We find that the most prominent sources of unstructured threat data are the threat reports, Twitter feeds, and posts from hackers and security experts. We also observe that security researchers mined CTI from unstructured sources to extract Indicator of Compromise (IoC), threat-related topic, and event detection. Finally, natural language processing (NLP) based approaches: topic classification; keyword identification; and semantic relationship extraction among the keywords are mostly availed in the selected studies to mine CTI information from unstructured threat sources.**

I. Introduction

Defending and preventing cyberattacks and breaches have become increasingly difficult as attackers improve their tactics, techniques, and procedures (TTPs) and craft clever schemes to manipulate targeted users with low technical abilities to exploit the vulnerabilities latent in the computing systems [1]. As a result, Information Technology (IT) organizations have to deal with cyberattacks that are better organized, funded, and financially motivated. For example, in July 2019, Capital One, one of the biggest US based credit card providers, experienced a data breach of credit card and social security numbers of 106 million users [2].

These circumstances have necessitated the introduction of proactive defense mechanisms against cyberattacks. The identification of cyberthreat intelligence (CTI) is one of the proactive defense mechanisms against cyberattacks. CTI is referred to the set of organized and collected information about cyberthreats that can be utilized to predict, prevent, or defend cyberattacks [3]. CTI information can also help IT organizations build necessary tactics and strategies to weaken the attacker's methods as well as to build tools and techniques to thwart malicious attempts. Consequently, in recent years, the amount of CTI that can be mined from textual reports has grown. The most usual sources of these CTI reports are available in publicly-accessible online and media artifacts, such as blogs, forums, social platforms, source repositories, issue trackers, news articles, and organizational reports. These reports can help cybersecurity practitioners understand the nature of cyberattacks and motives of the attackers by describing attackers' TTPs, and cyber kill chain phases. However, manually extracting relevant information from these large volumes of unstructured CTI reports is error-prone and inefficient [4].

Cybersecurity researchers [5], [6], [7] have focused on mining CTI that can automatically extract CTI information from unstructured reports. Mining CTI facilitates extracting the TTPs, structure of cyberattacks known as cyber kill chain [8]. This also helps identifying artifacts of operating systems, applications, programs, and networks leading to compromise, intrusion, or breach of a computing system collectively referred as indicators of compromise [9], [10]. Researchers [4], [11] are also working on transforming the extracted CTI to structured formats, such as CVE [12], STIX [13] and MITRE ATT&CK [14]. In this literature review, we identify peer-reviewed publications related to CTI mining and summarize the mining source, purposes, and approaches proposed for the CTI mining.

The goal of this research is *to aid cybersecurity researchers understand the source, purpose, and approaches for mining cyberthreat intelligence from unstructured text through a literature review of peer-reviewed studies on this topic*.

In this literature review, we ask these following research questions (RQs):

- RQ1: What unstructured sources are considered by cybersecurity researchers and practitioners for collecting CTI?
- RQ2: What are the purposes addressed by the researchers to mine CTI?
- RQ3: What approaches are availed to mine CTI from unstructured data?

We list our contribution as following:

1) A list of 38 selected studies on mining CTI from unstructured sources;
2) A list of unstructured sources that are being mined for extracting CTI; and
3) An analysis of the mining purposes and approaches to extract CTI from unstructured sources.

The rest of the paper is organized as follows. Section II provides background on literature reviews in CTI. Section III explains our research methodology. Section IV, V, VI, VII presents the result. Section VIII discusses the limitations of our study. Finally, Section IX concludes the study.

II. Literature Review on mining CTI

In this section, we report on CTI literature reviews. Sauerwein et. al [15] conduct a study of 22 threat intelligence sharing platforms that enable automation and smooth the generation, refinement, and examination of security data. Their study result in eight key findings including: "There is no common definition of threat intelligence sharing platforms"; and "Most platforms focus on data collection instead of analysis".

Tuma et. al [16] perform a Systematic Literature Review (SLR) on 26 methodologies of threat analysis. They compared methodologies based on different aspects such as applicability, characteristics of the analysis procedure, and ease of adoption. They also shed light on the restrictions for adopting the existing approaches and discuss the current state of their adoption in software engineering processes. These authors show that the analysis procedures are not clearly defined and they have lack of quality assurance and tool support.

Moreover, Xiong et. al [17] perform an SLR on threat modeling review articles and identify three types of article among these: (i) articles which are making a contribution to the field of threat modeling; (ii) articles which are using an existing threat modeling method; and (iii) articles which are presenting work related to the threat modeling process. They observe that most threat modeling works are done manually with a limited assurance of their validations.

Bridges et. al [18] evaluate existing methods of automatic extraction of security entities from text. They focused on finding security related entities. After comparing the existing approaches, the authors draw conclusions that the existing methods have a low recall, and no large publicly available data set of security documents is available.

Wagner et. al [19] explore the state-of-the-art approaches in the field of CTI sharing and investigate different problem areas. They use articles from academic and gray literature and focus on challenges. They list Collaboration, Trust, and Privacy & Anonymity the most focused topics in the literature.

Overall, these aforementioned researches focused on cyberthreat intelligence from various perspectives such as privacy, sharing, modeling and performance. However, in this study, we focus on extracting CTI from unstructured textual reports, the used sources, the purposes, and their approaches in this literature review.

III. Methodology

Our literature review on mining CTI is performed according to the guidelines prepared by Kitchenham [20]. In this section, we describe how we conducted our literature review process. The process has two phases: (i) search, and (ii) inclusion and exclusion criteria. In the following subsections, we will discuss each of these aforementioned steps.

A. Search

The first step to start any literature review process is to search for relevant research articles in scholar databases. We select six scholar databases to conduct our search: Institute of Electrical and Electronics Engineers (IEEE) Xplore [21], Association for Computing Machinery (ACM) Digital Library [22], ScienceDirect [23], SpringerLink [24], Wiley Online Library [25], and DBLP [26]. We select these databases because they are the recommended scholar databases for performing literature review in the computing science domain [27]. Our next step is to define the search strings that will find the relevant research articles from these aforementioned databases. We construct a set of search strings to identify relevant publications in the scholar databases.

1) (threat *OR* cyber*) *ONEAR/2* (intelligence *OR* action* *OR* advisories)
2) (threat *OR* cyber* *OR* security) *ONEAR/2* (report* *OR* article* *OR* information *OR* threat*)
3) "hacker forum*" *OR* "dark*" *OR* "cti" *OR* "tactics, techniques and procedures" *OR* "apt attack*"

We finally search each of the six scholar databases using the aforementioned search queries. The search process results in a set of research articles that are filtered using inclusion and exclusion criteria which are described in Section IV.

We use three search queries that also returns results not related to CTI mining, which motivates us to use validation techniques on the effectiveness of our search queries. We use the quasi sensitivity metrics (QSM) proposed in [28] which validates whether the set of search queries are sufficient enough to identify research publications on a specific topic from scholar databases. The QSM score is calculated as below. Let *PQGS* as *the publications from search strings which included in quasi-gold set* and *QGS* is *the publications in quasi-gold set*.

$$QSM = \frac{\#PQGS}{\#QGS} \tag{1}$$

Here, the calculation of QSM needs a set of publications referred to as a *quasi-gold set* (QGS) of publications. In [28], the QGS is defined as a collection of studies which is published in well-known conferences and journals recognized by the research community in the corresponding domain of knowledge within a relevant time span. We identify the quasi-gold set of paper using the following techniques.

1) We use the Google Scholar database to find the top journals and conference venues in cybersecurity research domains. This list can be found in [29] where all of the venues and journals are sorted in descending order based on the h5-index and h5-median. These venues are: a) ACM Symposium on Computer and Communications Security b) IEEE Transactions on Information Forensics and Security c) USENIX Security Symposium d) IEEE Symposium on Security and Privacy e) Network and Distributed System Security Symposium (NDSS) f) International Conference on Theory and Applications of

Cryptographic Techniques (EUROCRYPT) g) Computers & Security h) IEEE Transactions on Dependable and Secure Computing i) International Cryptology Conference (CRYPTO) j) International Conference on Financial Cryptography and Data Security k) International Conference on The Theory and Application of Cryptology and Information Security (ASIACRYPT) l) Security and Communication Networks m) Theory of Cryptography n) ACM on Asia Conference on Computer and Communications Security o) Proceedings on Privacy Enhancing Technologies p) IEEE European Symposium on Security and Privacy q) Designs, Codes and Cryptography r) European Conference on Research in Computer Security s) IEEE Security & Privacy t) Journal of Information Security and Applications

2) We scan titles of all the articles and proceedings published in the first 20 venues and journals and look for relevant papers in the mining CTI domain in the last 20 years, from January 2000 to March 2020.
3) Next, we apply forward and backward snowballing technique [30] on the collected set of papers from the prior step. Forward snowballing captures publications that cite the publications found in the prior step. Backward snowballing finds the publications that are cited by the publication found in the prior step.
4) We only consider publications that are peer-reviewed and written in English.
5) We exclude publications not related to mining CTI by reading the titles of the collected publications. If we cannot determine the relevance from the title, we read the abstract, and if still we are unable to determine, then we read the whole publication.

After going through all of these steps, we find publications which are our quasi-gold set. Then, we determine the QSM score of our search queries which can be computed using equation 1.

B. Inclusion and Exclusion Criteria

Our search results include publications that are not relevant to the studies of mining unstructured text for CTI. Hence, we establish inclusion and exclusion criteria to filter the irrelevant publications. Our inclusion and exclusion criteria are described below.

1) Exclusion Criteria:

- Publications that are not peer-reviewed: keynote abstracts, call for papers, and presentations;
- Publications that were published before 2000;
- Publications that are written in languages except in English

2) Inclusion Criteria:

- Publications must be available for downloading or reading on the web; and
- Title, Keywords, Abstract, and Introduction of the paper which explicitly indicates that the publication is related to mining unstructured text for CTI.

TABLE I
SEARCH RESULTS FOR SCHOLAR DATABASES

scholar Database	Count
IEEE	7,294
ACM Digital Library	6,389
ScienceDirect	3,992
Wiley Online Library	1,486
DBLP	8,890
Springer Link	431

After applying the aforementioned inclusion and exclusion criteria, we obtained the set of publications that we use for our literature review.

IV. RESULTS

In this section, we discuss the selected studies that we have found through our search process.

A. Selected Studies

First, in Table I, we report the number of publications found in each scholar database. In total, we find 28,484 publications in six scholar databases. We exported the search results from each database as *csv* or *bib* files in March 2020. We then eliminate the duplicate publication entries and get 20,938 publications. Finally, by applying the exclusion and inclusion criteria mentioned in Section III, we get 38 articles. In Table IV from Section X, these publications are listed in alphabetical order. We refer to each of the selected publications as *P#*, for example, *P1* refers to the article *Acing the ioc game: Toward automatic discovery and analysis of open-source cyberthreat intelligence*. We also calculate the QSM score using equation 1. The number of *quasi gold set* papers identified is 15 and the set of selected publications contains 11 of the papers in *quasi gold set*. Hence, the QSM score is approximately 73%. This score is acceptable as defined in [28].

B. Overview of the selected studies

The 38 selected publications have been published during the time period of 2007 to 2019. We find that 35 of the selected studies are conference and workshop papers, 3 studies are journal papers, and 1 study is published as a book chapter. We observe that 40% of the selected studies are published in three conferences. Among the selected studies, 8 studies are published in *IEEE International Conference on Intelligence and Security Informatics*; 4 studies are published in *IEEE International Conference on Big Data*; and 3 studies are published in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. The rest of the venues are associated with only one paper. We also observe that most CTI mining studies started from 2015. The period of 2018 to 2019 contributes 26 studies in total. However, we find two studies in 2007 and 2009 as well. Overall, mining CTI from unstructured sources has increased in last the two years.

TABLE II
CTI MINING SOURCES

CTI Mining Source	Studies	Count
Threat Reports	P2, P3, P5, P6, P8, P9, P12, P14, P31, P33, P34, P35, P36, P37, P38	15
Forums	P10, P13, P17, P20, P21, P22, P23, P24, P26, P27, P28	11
Twitter	P4, P15, P16, P18, P20, P21, P25, P32	8
Blogs	P1, P11, P19, P20	4
Version Control Repositories	P29	1
System and Application Logs	P7	1
Darknet Marketplace	P30	1

V. RQ1: WHAT UNSTRUCTURED SOURCES ARE CONSIDERED BY CYBERSECURITY RESEARCHERS AND PRACTITIONERS FOR COLLECTING CTI?

From the selected 38 studies, we observe that 7 types of sources where CTI mining has been applied. These sources are listed in the following subsections. As reported in Table II, threat reports have been explored most by the cybersecurity researchers for mining CTI, approximately 37% times. Hacker forums, Twitter feeds and security blogs have also been considered as valuable sources for threat intelligence mining, 27%, 20%, and 10% cases, respectively. Darknet marketplaces, version control repositories, and logs from systems and applications have each been used in one study. The URLs for the sources are given in Table V from Section X. We provide additional information on each of these source types in the following subsections.

A. Threat Reports, N=15

Cybersecurity organizations and researchers publish threat intelligence reports on cybersecurity incidents, system vulnerabilities, exploits, cyberattacks, and data breaches. These reports also contain the TTPs of the attackers along with tools and strategies they deploy to compromise the system's security. These threat reports contain information on the attacker's tactics and techniques, malware details, tools and procedures, vulnerabilities, and exploits. The selected studies have explored threat reports published by security vendors, such as Fireye [31], Kaspersky [32]; security research group such as Citizen Lab [33], expert forums such as BrightTalk [34] and APT reports [35]. From Table II, we also observe that researchers mined CTI from threat reports and advisories for extracting IoCs and TTPs of malicious users. For example, in P12, the authors develop a tool named *ChainSmith* where the tool automatically extracted 24,653 IoCs from a set of 14,155 threat reports. In P37, the authors develop a tool called *TTPDrill* where they analyze the threat reports from Symantec [36] to extract threat actions and TTPs and store the TTPs into STIX format.

B. Forums, N=11

Online forums maintained by security vendors, security experts, malicious users, hackers, penetration testers contain resources regarding attack TTPs, vulnerabilities, exploits, security news, troubleshooting, best practices, and malware analysis. These forums also work as a communication platform for security experts and hackers. Forum posts are used for hacker asset analysis and CTI topic identification. For example, in P10, the authors use data hacker forums to extract the assets from the hacker conversation and source code snippets. In P13, the authors identify and cluster the most relevant hacker topics in the forums.

C. Twitter Feeds, N=8

Twitter [37] has become an important source of cutting edge CTI [38]. News and updates are made available through Twitter such as zero day attacks, new vulnerabilities, and data breaches. Reputed security experts and organizations, such as Brian Krebs [39] and InterSec World Wide [40], regularly posts feeds related to threat intelligence and security incidents. Twitter feeds are written in natural languages (mostly English) which have been used by the security researchers to extract CTI. Selected studies that use Twitter feeds as unstructured source mostly to mine the threat feeds to classify threat-related topics and keywords, and identify past or future threat events. For example, in P15, the authors use deep learning techniques to identify cyberthreat related Twitter feeds. In P16, the authors develop a tool named *CyberTwitter* which can extract vulnerability concepts from the feeds, build a graph-based knowledge base, and predict future threat events related to the identified vulnerabilities.

D. Web Blogs, N=4

Blogs written by security experts or white hat hackers are rich sources of technical information on exploits, vulnerabilities, attacks, and security-related news. Textual information obtained from these web blogs to mine threat intelligence in the selected studies. Blogs have been explored by the researchers for IoC extraction, threat event, and topic identification. For example, in P1, 9 million IoCs have been extracted from 71,000 blog articles.

E. Version Control Repositories, N=1

Github [41], a version control repository, is used by software practitioners for tracking source code versions, documentation, wiki, issues, bug reports, and commit logs. Commit messages, bugs and issues are important sources for vulnerable code and packages, exploits, security measures to tackle threats against malicious users. These are mostly written in natural language such as English. In P29, commit logs, bugs and issue reports from open source repositories in Github have been explored for mining vulnerability information.

TABLE III
CTI SOURCES, PURPOSE AND HIGH LEVEL APPROACHES

Studies	CTI Mining Source	**CTI Mining Purpose**	**Approaches**
P1	Blogs	IoC extraction	IoC related topic classification, IoC term extraction, IoC relationship extraction
P2	Threat Reports	Attack attribution	Semantic concept extraction of CTI terms
P3	Threat Reports	TTPs extraction	Semantic concept extraction of CTI terms
P4	Twitter	Threat event identification	CTI topic classification, named entity recognition of CTI terms
P5	Threat Reports	TTPs extraction	CTI terms identification, named entity recognition of CTI terms
P6	Threat Reports	IoC extraction	Custom open source tool for IoC parsing
P7	Logs	TTPs extraction	Custom open source tool for IoC parsing
P8	Threat Reports	TTPs extraction	CTI terms identification
P9	Threat Reports	Threat topic analysis	Semantic concept extraction among CTI terms
P10	Forums	Hacker asset Analysis	CTI topic classification
P11	Blogs	Threat topic analysis	CTI topic classification, CTI terms identification
P12	Threat Reports	IoC extraction	CTI semantic concept extraction, named entity recognition of CTI terms
P13	Forums	Threat topic analysis	CTI topic classification
P14	Threat Reports	IoC extraction	semantic concept extraction of CTI terms
P15	Twitter	Threat topic analysis	CTI topic classification
P16	Twitter	Threat event identification	Vulnerability keyword extraction, semantic concept extraction of CTI terms, knowledge graph, threat ontology
P17	Forums	Hacker asset Analysis	CTI terms identification
P18	Twitter	Threat event identification	CTI topic classification
P19	Blogs	Threat event identification	CTI terms identification, semantic concept extraction of CTI terms
P20	Forums, Twitter, Blogs	Threat event identification	CTI terms identification, semantic concept extraction of CTI terms
P21	Forums, Twitter	Threat event identification	CTI terms identification, semantic concept extraction of CTI terms
P22	Forums	Threat topic analysis	CTI topic classification
P23	Forums	Hacker asset analysis	CTI topic classification, CTI terms identification
P24	Forums	Threat topic analysis	CTI topic classification
P25	Twitter	Threat topic analysis	CTI topic classification
P26	Forums	Threat topic analysis	CTI topic classification
P27	Forums	IoC extraction	CTI terms identification, semantic concept extraction of CTI terms
P28	Forums	IoC extraction	CTI topic classification
P29	Version Control Repositories	Software vulnerability extraction	Semantic concept extraction of CTI terms, knowledge graph, threat ontology
P30	Darknet Marketplace	Threat event identification	CTI terms identification, semantic concept extraction of CTI terms
P31	Threat Reports	IoC extraction	CTI terms identification
P32	Twitter	Threat topic analysis	CTI topic classification, semantic concept extraction of CTI terms
P33	Threat Reports	Threat event identification	CTI topic classification, text summarization of CTI topics
P34	Threat Reports	TTPs extraction	CTI terms identification, semantic concept extraction of CTI terms
P35	Threat Reports	IoC extraction	Named entity recognition of CTI terms, semantic concept extraction of CTI terms
P36	Threat Reports	Threat event identification	CTI topic classification, semantic concept extraction of CTI terms
P37	Threat Reports	TTPs extraction	Threat ontology, semantic concept extraction of CTI terms
P38	Threat Reports	TTPs extraction	semantic concept extraction of CTI terms

F. Systems and Application Logs, N=1

Logs written in the unstructured format generated from operating systems, software, network applications contain information, such as error and debug warning, system and application activities. In the study P7, researchers extract TTPs from these log information.

G. Darknet MarketPlace, N=1

The darknet marketplace often contains information on illegal cybersecurity activities, such as hacking tools; malware; spyware; ransomware; pawned passwords; and leaked data from software services. In P30, researchers mine CTI from information extracted from darknet marketplaces. In P30, researchers analyze darknet marketplace data to identify future cyberthreat activities.

VI. RQ2: WHAT ARE THE PURPOSES ADDRESSED BY THE RESEARCHERS TO MINE CTI?

We report our findings on the purposes of mining CTI in the papers in Table III. We find 7 types of purposes targeted by the researchers to mine CTI from unstructured sources. We

find that researchers mine CTI for mainly four purposes: threat event identification (9 times), threat topic analysis (9 times), IoC extraction (8 times), and TTPs extraction (7 times). The other three CTI mining purposes appear in selected studies fewer times than the aforementioned four purposes: hacker asset analysis (3 times), attack attribution (once), and software vulnerability extraction (once). These purposes are described below.

- **Threat event identification (N=9):** Authors in the studies mine CTI to gain proactive information on likely future cyberattacks or analyze the textual data on previous cyberattacks. Twitter feeds are the most used dataset for identifying CTI threats as it has been used 5 (P4, P16, P18, P20, P21) times out of 9 in the studies. Threat reports (twice in P33 and P36), blogs (twice in P19, P20), forums (twice in P20, P21), and darknet marketplaces (once in P30) are also used for the dataset by the authors.
- **Threat topic analysis (N=9):** Textual information on cyberthreat topics contains details on vulnerabilities, exposures, cyberthreat events, tools, and techniques used by the hackers. From these texts, threat events can be identified through the classification on threat-related keywords, topics, and semantic relation among CTI terms. Forum posts and Twitter feeds are the most used dataset for identifying threat events as they have been used respectively 4 (P13, P22, P24, P26) and 3 (P15, P25, P32) times out of 9 in the studies. Threat reports (P9), and blog posts (P11) are also used once for the dataset in the studies.
- **IoC extraction (N=8):** Indicators of Compromise (IoC) refers to the artifacts of operating systems, applications, programs, and networks which indicate compromise, intrusion, or breach of computing systems [9]. Textual information on cyberthreat topics contains details on vulnerabilities, exposures, cyberthreat events, tools, and techniques used by the hackers which is mined for extracting IoCs. Threat reports are the most used dataset for identifying CTI threats as it has been used 5 times (P6, P12, P14, P31, P35) out of 8 in the studies. Forum (P28) and blog posts (P1) are also used for the dataset in the studies.
- **TTPs extraction (N=7):** TTPs refers to the cyberattack tactics, techniques and procedures [11]. Information on TTPs is written in unstructured texts in threat-related topics. Selected studies mostly used threat reports from cybersecurity vendors as the dataset for extracting TTPs, 6 times (P3, P5, P8, P34, P37, P38) out of 7. Log data from systems and applications is used once (P7) as a dataset.
- **Hacker asset analysis (N=3):** Darknet and hacker forums often contain textual descriptions on hacking tools, documents, scripts, source codes, and online resources regularly used by the attackers' community. In the selected studies, all of the 3 work (P10, P17, P23) have used forum posts and attachments as the dataset for hacker asset categorization, storage, search, and visualization.
- **Attack attribution (N=1):** Threat related topics may contain information on cyberthreat incidents, and associated cyberattack actors such as their roles, strategies, and procedures. The researchers of P2 mined threat reports from security vendors to map cyberthreat actors to the cyberthreat incidents based on the attack patterns.
- **Software vulnerability extraction (N=1):** Open source repositories from version control systems contain information on software vulnerabilities, exploits, bugs, issues, and comments. In P29, these repository information are mined to collect software vulnerability information, dependent packages, libraries, framework, secure alternative packages.

VII. RQ3: What approaches are availed to mine CTI from unstructured data?

We report our findings on the mining approaches of the selected studies in Table III. As the unstructured threat topics are written in natural languages, natural language processing (NLP) techniques have been used by the authors in these studies. The approaches used in the studies, as single or combination of approaches, are reported in Table III. We observe that NLP techniques related to topic classification, keyword identification, and semantic relation among the keywords and topics are mostly used in the selected studies. The most used technique is to find the semantic relationship among the CTI keywords which is 18 times followed by keyword identification (17 times) and topic classification (16 times). These aforementioned approaches include techniques such as tokenization, parts of speech tagging, feature extraction, latent semantic analysis, latent dirichlet allocation, named entity recognition, and sequence labeling. Threat ontology-based techniques have also been used twice in the studies to map the extracted keywords to match patterns with the TTPs. Text summarization techniques have been used once to identify threat events related topics. Finally, in two studies, a custom open-source tool named *cti-python-stix2* [42] is used by the authors to mine IoCs from threat related topics.

VIII. Threats to validity

We discuss the limitation of our literature review as following. The search process of finding the relevant papers may not be comprehensive as we use six scholar databases as sources. However, other scholar databases may contain more studies. As the search process is done by March 2020. We do not include papers published after this date. The process of searching, and applying inclusion and exclusion criteria is subjective and may miss relevant studies. The findings in our paper are based on the 38 selected studies. As mentioned earlier in this section, the generality of our findings may be limiting as the process of searching and documenting the findings from each study is a subjective process

IX. Conclusion

Cyberthreat intelligence (CTI) can be extracted from unstructured texts on threat-related topics written by cybersecurity researchers, organizations, and vendors. In this literature review, we identify 38 peer-reviewed studies on mining CTI from unstructured texts from a set of 28,484 publications collected from 6 scholar databases. We report the source, purpose, and approaches for mining CTI in the selected studies. We observe that the utilized sources of unstructured threat data are the threat reports, Twitter feeds, and posts from hackers and security experts. We also find that authors in the studies identified Indicator of Compromise (IoC), classified threat-related topics, and detected past and future cyberthreat events through mining CTI. Finally, the majority of the studies used natural language processing (NLP) based approaches such as CTI topic classification, CTI terms identification along with semantic relationship extraction among the CTI terms.

Acknowledgement

This work is supported by the NSA Science of Security Lablet and the NSA Laboratory for Analytic Sciences.

References

[1] J. Kutscher, "M-Trends 2017: A View From the Front Lines." https://www.fireeye.com/blog/threat-research/2017/03/m-trends-2017.html. [Online; accessed 22-August-2020].

[2] R. Mcmilan, "Capital One Breach Casts Shadow Over Cloud Security." https://www.wsj.com/articles/capital-one-breach-casts-shadow-over-cloud-security-11564516541. [Online; accessed 22-August-2020].

[3] V. S. M. Legoy, "Retrieving att&ck tactics and techniques in cyber threat reports," Master's thesis, University of Twente, 2019.

[4] U. Noor, Z. Anwar, T. Amjad, and K.-K. R. Choo, "A machine learning-based fintech cyber threat attribution framework using high-level indicators of compromise," *Future Generation Computer Systems*, vol. 96, pp. 227–242, 2019.

[5] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 755–766, 2016.

[6] Z. Zhu and T. Dumitras, "Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports," in *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 458–472, IEEE, 2018.

[7] K.-C. Lee, C.-H. Hsieh, L.-J. Wei, C.-H. Mao, J.-H. Dai, and Y.-T. Kuang, "Sec-buzzer: cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation," *Soft Computing*, vol. 21, no. 11, pp. 2883–2896, 2017.

[8] J. Greenert and M. Welsh, "Breaking the kill chain," *Foreign Policy*, vol. 16, 2013.

[9] W. Gragido, "Understanding indicators of compromise (IOC) part I." https://web.archive.org/web/20170914034202/https://blogs.rsa.com/understanding-indicators-of-compromise-ioc-part-i/. [Online; accessed 22-August-2020].

[10] S. Samtani, H. Zhu, and H. Chen, "Proactively identifying emerging hacker threats from the dark web: A diachronic graph embedding framework (d-gef)," *ACM Transactions on Privacy and Security (TOPS)*, vol. 23, no. 4, pp. 1–33, 2020.

[11] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and 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*, pp. 103–115, 2017.

[12] MITRE, " Common Vulnerabilities and Exposures." https://cve.mitre.org/. [Online; accessed 22-August-2020].

[13] MITRE, "Structured Threat Information eXpression." https://stixproject.github.io/. [Online; accessed 22-August-2020].

[14] MITRE, "MITRE ATTCK." https://attack.mitre.org/. [Online; accessed 22-August-2020].

[15] C. Sauerwein, C. Sillaber, A. Mussmann, and R. Breu, "Threat intelligence sharing platforms: An exploratory study of software vendors and research perspectives," 2017.

[16] K. Tuma, G. Çalikli, and R. Scandariato, "Threat analysis of software systems: A systematic literature review," *Journal of Systems and Software*, vol. 144, pp. 275–294, 2018.

[17] W. Xiong and R. Lagerström, "Threat modeling–a systematic literature review," *Computers & Security*, vol. 84, pp. 53–69, 2019.

[18] R. A. Bridges, K. M. Huffer, C. L. Jones, M. D. Iannacone, and J. R. Goodall, "Cybersecurity automated information extraction techniques: Drawbacks of current methods, and enhanced extractors," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 437–442, IEEE, 2017.

[19] T. D. Wagner, K. Mahbub, E. Palomar, and A. E. Abdallah, "Cyber threat intelligence sharing: Survey and research directions," *Computers & Security*, vol. 87, p. 101589, 2019.

[20] S. Keele *et al.*, "Guidelines for performing systematic literature reviews in software engineering," tech. rep., Technical report, Ver. 2.3 EBSE Technical Report. EBSE, 2007.

[21] "Ieee xplore." [Online]. Available: https://ieeexplore.ieee.org/ Accessed 22 Aug 2020.

[22] "Acm digital library." [Online]. Available: https://dl.acm.org/ Accessed 22 Aug 2020.

[23] "Sciencedirect." [Online]. Available: https://sciencedirect.com/ Accessed 22 Aug 2020.

[24] "Springer." [Online]. Available: https://link.springer.com/ Accessed 22 Aug 2020.

[25] "Wiley online library." [Online]. Available: https://onlinelibrary.wiley.com/ Accessed 22 Aug 2020.

[26] "dblp, computer science bibliography." [Online]. Available: https://dblp.uni-trier.de/ Accessed 22 Aug 2020.

[27] M. Kuhrmann, D. M. Fernández, and M. Daneva, "On the pragmatic design of literature studies in software engineering: an experience-based guideline," *Empirical software engineering*, vol. 22, no. 6, pp. 2852–2891, 2017.

[28] H. Zhang, M. A. Babar, and P. Tell, "Identifying relevant studies in software engineering," *Information and Software Technology*, vol. 53, no. 6, pp. 625–637, 2011.

[29] G. Scholar, "Google Scholar — Top Publications — Compyter Security and Cryptography." https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computersecuritycryptography. [Online; accessed 22-August-2020].

[30] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[31] F. Inc, " Cyber threat intelligence on advanced attack groups and technology vulnerabilities ." https://www.fireeye.com/current-threats/threat-intelligence-reports.html. [Online; accessed 22-August-2020].

[32] K. Lab, " Securelist — Kaspersky's cyberthreat research and reports ." https://securelist.com/. [Online; accessed 22-August-2020].

[33] C. Lab, " Targeted Threat Archives ." https://citizenlab.ca/category/research/targeted-threats/. [Online; accessed 22-August-2020].

[34] BrightTALK, " Threat Intelligence - BrightTALK ." https://www.brighttalk.com/topic/threat-intelligence/. [Online; accessed 22-August-2020].

[35] K. Lab, " Kaspersky's APT Intelligence Reports ." https://usa.kaspersky.com/enterprise-security/apt-intelligence-reporting. [Online; accessed 22-August-2020].

[36] Symantec, "Symantec." [Online]. Available: https://securitycloud.symantec.com/cc/#/landing Accessed 22 Aug 2020.

[37] Twitter, " Twitter ." https://twitter.com. [Online; accessed 22-August-2020].

[38] S. Mittal, P. K. Das, V. Mulwad, A. Joshi, and T. Finin, "Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 860–867, IEEE, 2016.

[39] krebsonsecurity, "krebsonsecurity." [Online]. Available: https://krebsonsecurity.com/ Accessed 22 Aug 2020.

[40] Intersec Worldwide, "Intersec wordlwide." [Online]. Available: https://intersecworldwide.com/ Accessed 22 Aug 2020.

[41] Github, " Github ." https://github.com. [Online; accessed 22-August-2020].

[42] oasis-open, "oasis-open/cti-python-stix2." [Online]. Available: https://github.com/oasis-open/cti-python-stix2 Accessed 22 Aug 2020.

[43] A. Niakanlahiji, J. Wei, and B.-T. Chu, "A natural language processing based trend analysis of advanced persistent threat techniques," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2995–3000, IEEE, 2018.

[44] A. Bose, V. Behzadan, C. Aguirre, and W. H. Hsu, "A novel approach for detection and ranking of trendy and emerging cyber threat events in twitter streams," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 871–878, IEEE, 2019.

[45] Y. Ghazi, Z. Anwar, R. Mumtaz, S. Saleem, and A. Tahir, "A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources," in *2018 International Conference on Frontiers of Information Technology (FIT)*, pp. 129–134, IEEE, 2018.

[46] D. Kim and H. K. Kim, "Automated dataset generation system for collaborative research of cyber threat analysis," *Security and Communication Networks*, vol. 2019, 2019.

[47] F. Sadique, S. Cheung, I. Vakilinia, S. Badsha, and S. Sengupta, "Automated structured threat information expression (stix) document generation with privacy preservation," in *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 847–853, IEEE, 2018.

[48] G. Ayoade, S. Chandra, L. Khan, K. Hamlen, and B. Thuraisingham, "Automated threat report classification over multi-source data," in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, pp. 236–245, IEEE, 2018.

[49] T. Wang and K. P. Chow, "Automatic tagging of cyber threat intelligence unstructured data using semantics extraction," in *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 197–199, IEEE, 2019.

[50] S. Samtani, K. Chinn, C. Larson, and H. Chen, "Azsecure hacker assets portal: Cyber threat intelligence and malware analysis," in *2016 IEEE conference on intelligence and security informatics (ISI)*, pp. 19–24, Ieee, 2016.

[51] F. S. Tsai and K. L. Chan, "Blog data mining for cyber security threats," in *Data Mining for Business Applications*, pp. 169–182, Springer, 2009.

[52] I. Deliu, C. Leichter, and K. Franke, "Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent dirichlet allocation," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5008–5013, IEEE, 2018.

[53] Z. Long, L. Tan, S. Zhou, C. He, and X. Liu, "Collecting indicators of compromise from unstructured text of cybersecurity articles using neural-based sequence labelling," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[54] V. Behzadan, C. Aguirre, A. Bose, and W. Hsu, "Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5002–5007, IEEE, 2018.

[55] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart, and P. Shakarian, "Darknet and deepnet mining for proactive cybersecurity threat intelligence," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pp. 7–12, IEEE, 2016.

[56] F. Alves, P. M. Ferreira, and A. Bessani, "Design of a classification model for a twitter-based streaming threat monitor," in *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pp. 9–14, IEEE, 2019.

[57] F. S. Tsai and K. L. Chan, "Detecting cyber security threats in weblogs using probabilistic models," in *Pacific-Asia Workshop on Intelligence and Security Informatics*, pp. 46–57, Springer, 2007.

[58] A. Sapienza, S. K. Ernala, A. Bessi, K. Lerman, and E. Ferrara, "Discover: Mining online chatter for emerging cyber threats," in *Companion Proceedings of the The Web Conference 2018*, pp. 983–990, 2018.

[59] A. Sapienza, A. Bessi, S. Damodaran, P. Shakarian, K. Lerman, and E. Ferrara, "Early warnings of cyber threats in online discussions," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 667–674, IEEE, 2017.

[60] M. Kadoguchi, S. Hayashi, M. Hashimoto, and A. Otsuka, "Exploring the dark web for cyber threat intelligence using machine leaning," in *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 200–202, IEEE, 2019.

[61] V. Benjamin, W. Li, T. Holt, and H. Chen, "Exploring threats and vulnerabilities in hacker web: Forums, irc and carding shops," in *2015 IEEE international conference on intelligence and security informatics (ISI)*, pp. 85–90, IEEE, 2015.

[62] I. Deliu, C. Leichter, and K. Franke, "Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks," in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 3648–3656, IEEE, 2017.

[63] B.-D. Le, G. Wang, M. Nasim, and M. A. Babar, "Gathering cyber threat intelligence from twitter using novelty classification," in *2019 International Conference on Cyberworlds (CW)*, pp. 316–323, IEEE, 2019.

[64] A. S. Gautam, Y. Gahlot, and P. Kamat, "Hacker forum exploit and classification for proactive cyber threat intelligence," in *International Conference on Inventive Computation Technologies*, pp. 279–285, Springer, 2019.

[65] M. Macdonald, R. Frank, J. Mei, and B. Monk, "Identifying digital threats in a hacker web forum," in *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015*, pp. 926–933, 2015.

[66] R. Williams, S. Samtani, M. Patton, and H. Chen, "Incremental hacker forum exploit collection and classification for proactive cyber threat intelligence: An exploratory study," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 94–99, IEEE, 2018.

[67] L. Neil and A. Joshi, "Mining threat intelligence about open-source projects and libraries from code repository issues and bug reports," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 7–12, IEEE, 2018.

[68] F. Dong, S. Yuan, H. Ou, and L. Liu, "New cyber threat discovery from darknet marketplaces," in *2018 IEEE Conference on Big Data and Analytics (ICBDA)*, pp. 62–67, IEEE, 2018.

[69] R. Azevedo, I. Medeiros, and A. Bessani, "Pure: Generating quality threat intelligence by clustering and correlating osint," in *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pp. 483–490, IEEE, 2019.

[70] K. Li, H. Wen, H. Li, H. Zhu, and L. Sun, "Security osif: Toward automatic discovery and analysis of event based cyber threat intelligence," in *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp. 741–747, IEEE, 2018.

[71] 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*, pp. 181–187, 2017.

[72] N. Kim, M. Kim, S. Lee, H. Cho, B.-i. Kim, J.-h. Park, and M. Jun, "Study of natural language processing for collecting cyber threat intelligence using syntaxnet," in *International Symposium of Information and Internet Technology*, pp. 10–18, Springer, 2018.

[73] T. Bo, Y. Chen, C. Wang, Y. Zhao, K.-Y. Lam, C.-H. Chi, and H. Tian, "Tom: A threat operating model for early warning of cyber security threats," in *International Conference on Advanced Data Mining and Applications*, pp. 696–711, Springer, 2019.

[74] G. Husari, X. Niu, B. Chu, and E. Al-Shaer, "Using entropy and mutual information to extract threat actions from cyber threat intelligence," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 1–6, IEEE, 2018.

X. APPENDIX

TABLE IV: List of publications selected for Systematic Literature Review

P1 [5]	Liao, Xiaojing, et al. "Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence." 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016.
P2 [4]	Noor, Umara, et al. "A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise." Future Generation Computer Systems 96 (2019): 227-242.
P3 [43]	Niakanlahiji, Amirreza, et al. "A natural language processing based trend analysis of advanced persistent threat techniques." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.
P4 [44]	Bose, Avishek, et al. "A novel approach for detection and ranking of trendy and emerging cyber threat events in twitter streams." 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2019.
P5 [45]	Ghazi, Yumna, et al. "A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources." 2018 International Conference on Frontiers of Information Technology (FIT). IEEE, 2018.
P6 [46]	Kim, Daegeon, et al. "Automated Dataset Generation System for Collaborative Research of Cyber Threat Analysis." Security and Communication Networks 2019 (2019).
P7 [47]	Sadique, Farhan, et al. "Automated structured threat information expression (STIX) document generation with privacy preservation." 2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, 2018.
P8 [48]	Ayoade, Gbadebo, et al. "Automated threat report classification over multi-source data." 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC). IEEE, 2018.
P9 [49]	Wang, Tianyi, et al. "Automatic Tagging of Cyber Threat Intelligence Unstructured Data using Semantics Extraction." 2019 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2019.
P10 [50]	Samtani, Sagar, et al. "Azsecure hacker assets portal: Cyber threat intelligence and malware analysis." 2016 IEEE conference on intelligence and security informatics (ISI). Ieee, 2016.
P11 [51]	Tsai, Flora S., et al. "Blog data mining for cyber security threats." Data Mining for Business Applications. Springer, Boston, MA, 2009. 169-182.
P12 [6]	Zhu, Ziyun, et al. "Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports." 2018 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2018.
P13 [52]	Deliu, Isuf, et al. "Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent dirichlet allocation." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.
P14 [53]	Long, Zi, et al. "Collecting Indicators of Compromise from Unstructured Text of Cybersecurity Articles using Neural-Based Sequence Labelling." 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019.
P15 [54]	Behzadan, Vahid, et al. "Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.
P16 [38]	Mittal, Sudip, et al. "Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities." 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2016.
P17 [55]	Nunes, Eric, et al. "Darknet and deepnet mining for proactive cybersecurity threat intelligence." 2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, 2016.
P18 [56]	Alves, Fernando, et al. "Design of a Classification Model for a Twitter-based Streaming Threat Monitor." 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). IEEE, 2019.
P19 [57]	Tsai, Flora S., et al. "Detecting cyber security threats in weblogs using probabilistic models." Pacific-Asia Workshop on Intelligence and Security Informatics. Springer, Berlin, Heidelberg, 2007.
P20 [58]	Sapienza, Anna, et al. "Discover: Mining online chatter for emerging cyber threats." Companion Proceedings of the The Web Conference 2018. 2018.
P21 [59]	Sapienza, Anna, et al. "Early warnings of cyber threats in online discussions." 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017.
P22 [60]	Kadoguchi, Masashi, et al. "Exploring the Dark Web for Cyber Threat Intelligence using Machine Leaning." 2019 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2019.
P23 [61]	Benjamin, Victor, et al. "Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops." 2015 IEEE international conference on intelligence and security informatics (ISI). IEEE, 2015.
P24 [62]	Deliu, Isuf, et al. "Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks." 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017.
P25 [63]	Le, Ba-Dung, et al. "Gathering cyber threat intelligence from Twitter using novelty classification." 2019 International Conference on Cyberworlds (CW). IEEE, 2019.
P26 [64]	Gautam, Apurv Singh, et al. "Hacker Forum Exploit and Classification for Proactive Cyber Threat Intelligence." International Conference on Inventive Computation Technologies. Springer, Cham, 2019.
P27 [65]	Macdonald, Mitch, et al. "Identifying digital threats in a hacker web forum." 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015. 2015.

P28 [66]	Williams, Ryan, et al. "Incremental hacker forum exploit collection and classification for proactive cyber threat intelligence: An exploratory study." 2018 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2018.
P29 [67]	Neil, Lorenzo, et al. "Mining threat intelligence about open-source projects and libraries from code repository issues and bug reports." 2018 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2018.
P30 [68]	Dong, Fangzhou, et al. "New cyber threat discovery from darknet marketplaces." 2018 IEEE Conference on Big Data and Analytics (ICBDA). IEEE, 2018.
P31 [69]	Azevedo, Rui, et al. "PURE: Generating quality threat intelligence by clustering and correlating OSINT." 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, 2019.
P32 [7]	Lee, Kuo-Chan, et al. "Sec-Buzzer: cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation." Soft Computing 21.11 (2017): 2883-2896.
P33 [70]	Li, Ke, et al. "Security OSIF: Toward automatic discovery and analysis of event based cyber threat intelligence." 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation. IEEE, 2018.
P34 [71]	Ramnani, Roshni R., et al. "Semi-automated information extraction from unstructured threat advisories." Innovations in Software Engineering Conference. 2017.
P35 [72]	Kim, Nakhyun, et al. "Study of Natural Language Processing for Collecting Cyber Threat Intelligence Using SyntaxNet." International Symposium of Information and Internet Technology. Springer, Cham, 2018.
P36 [73]	Bo, Tao, et al. "TOM: A Threat Operating Model for Early Warning of Cyber Security Threats." International Conference on Advanced Data Mining and Applications. Springer, Cham, 2019.
P37 [11]	Husari, Ghaith, et al. "Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources." Annual Computer Security Applications Conference. 2017.
P38 [74]	Husari, Ghaith, et al. "Using entropy and mutual information to extract threat actions from cyber threat intelligence." 2018 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2018.

TABLE V: List of CTI Mining Sources

Source Type	**Sources** [1]
Threat Reports	https://www.trendmicro.com/vinfo/us/security/research-and-analysis/threat-reports, https://securitycloud.symantec.com/cc/#/landing, https://securelist.com/, https://www.fireeye.com/current-threats/threat-intelligence-reports.html, https://www.threatminer.org/, https://citizenlab.ca/tag/cybersecurity/, https://krebsonsecurity.com/, https://us-cert.cisa.gov/, https://www.baesystems.com/en/cybersecurity/threat-intelligence-insights, https://labs.bitdefender.com/tag/bitdefender-report/, https://www.checkpoint.com/products-solutions/threat-intelligence/, https://www.cisco.com/c/en_uk/products/security/security-reports.html, https://www.secureworks.com/, https://www.eset.com/us/business/services/threat-intelligence/, https://www.mcafee.com/enterprise/en-us/threat-center/mcafee-labs/reports.html, https://www.microsoft.com/security/blog/microsoft-security-intelligence/, https://www.rsa.com/en-us/company/news/rsa-netwitness-platform-only-solution-to-integrate-threat-intelligence, APT notes (https://github.com/aptnotes/data), Cybermonitor (https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections), https://www.welivesecurity.com/, https://attack.mitre.org/, https://tika.apache.org/
Forums	OpenSC*, Xeksec*, nulled.io*, cracking arena*, http://garage4hackers.com/, http://www.hacksden.com/, https://www.antionline.com/, Crackingzilla*, WebCracking*, http://www.safeskyhacks.com/Forums/forum.php, Open Forum Discussion Crawler*
Blogs	https://cybersecurity.att.com/blogs, https://blog.malwarebytes.com/, https://www.webroot.com/blog/, (https://www.icwsm.org/data.html), Crawled data from https://www.torproject.org/, https://geti2p.net/en/
Darknet Marketplace	Dream Market*, Berlusconi Market*
Version Control Repositories	https://www.github.com

[1]The * mark means the url is not found or no working url found