What Are the Attackers Doing Now? Automating Cyberthreat Intelligence Extraction from Text on Pace with the Changing Threat Landscape: A Survey

MD RAYHANUR RAHMAN, REZVAN MAHDAVI HEZAVEH, and
LAURIE WILLIAMS, North Carolina State University, USA

Cybersecurity researchers have contributed to the automated extraction of CTI from textual sources, such as threat reports and online articles describing cyberattack strategies, procedures, and tools. *The goal of this article is to aid cybersecurity researchers in understanding the current techniques used for cyberthreat intelligence extraction from text through a survey of relevant studies in the literature.* Our work finds 11 types of extraction purposes and 7 types of textual sources for CTI extraction. We observe the technical challenges associated with obtaining available clean and labeled data for replication, validation, and further extension of the studies. We advocate for building upon the current CTI extraction work to help cybersecurity practitioners with proactive decision-making such as in threat prioritization and mitigation strategy formulation to utilize knowledge from past cybersecurity incidents.

CCS Concepts: • **General and reference → Surveys and overviews;**

Additional Key Words and Phrases: Cyberthreat intelligence, CTI extraction, CTI mining, IoC extraction, TTPs extraction, attack pattern extraction, threat reports, tactical threat intelligence, technical threat intelligence

ACM Reference format:
Md Rayhanur Rahman, Rezvan Mahdavi Hezaveh, and Laurie Williams. 2023. What Are the Attackers Doing Now? Automating Cyberthreat Intelligence Extraction from Text on Pace with the Changing Threat Landscape: A Survey. *ACM Comput. Surv.* 55, 12, Article 241 (March 2023), 36 pages.
https://doi.org/10.1145/3571726

1 INTRODUCTION

Defending and preventing cyberattacks have become increasingly difficult as attack tactics and techniques continuously evolve. The attackers' community is now more organized in its operation and is driven by financial motives [146]. Cyberattack trends have shifted from small group efforts to more significant organized crime. For example, in 2020, the University of Utah experienced a ransomware attack where an attacker group stole sensitive student information, and the university suffered a financial loss of $457,000 [40].

To keep pace with attackers' ever-changing ways of launching cyberattacks, **cyberthreat intelligence (CTI)**, also known as *threat intelligence*, can be utilized to help **information technology**

This work is partly supported by the NSA Science of Security award H98230-17-D-0080.
Authors' address: Md R. Rahman, R. M. Hezaveh, and L. Williams, North Carolina State University, Department of Computer Science, 890 Oval Drive, Box 8206 Engineering Building II, Raleigh, NC 27695-8206; emails: {mrahman, rmahdav, lawilli3}@ncsu.edu.

0360-0300/2023/03-ART241 $15.00
https://doi.org/10.1145/3571726

(IT) organizations defend against cyberattacks proactively. According to the definition provided by McMillan et al. [148], "Threat Intelligence is evidence-based knowledge, including context, mechanisms, indicators, implications, and action-oriented advice about an existing or emerging menace or hazard to assets. This intelligence can inform decisions regarding the subject's response to that menace or hazard." Hence, CTI is a set of organized and collected information about cyberthreats that organizations can utilize to predict, prevent, or defend against cyberattacks [91, 125]. CTI can help IT organizations build the necessary tactics and strategies to weaken the attackers' methods and build tools and techniques to thwart malicious attempts. For example, in 2020, Amazon revealed that the company's AWS Shield, developed by the company's threat analysis group and security team, defended a state-sponsored **distributed denial of service (DDoS)** attack with a peak traffic volume of 2.3 Tbps, which is the most substantial DDoS attack to date [151].

CTI can be extracted, aggregated, synthesized, and analyzed from publicly available cyberthreat-related documents, articles, reports, social media, and human intelligence [37]. These information sources can contain how the attackers target the organization, their strategies, which tools and procedures the attackers utilize, and detailed descriptions of how adversaries launch attacks. As the number and types of attacks have grown, so has the volume of textual content focusing on cyberthreat news, attack patterns, tools, and techniques. The attack tactics and techniques are also consistently evolving. Extracting the most critical information has become challenging due to the large volume of data, noise, anomalies, and difficulty in establishing a correlation among the obtained information. Moreover, these data are in textual formats written in natural language (i.e., English). Hence, manually extracting relevant information from the CTI-related documents can be error-prone and inefficient [143].

Cybersecurity researchers and practitioners have focused on the automatic extraction of CTI information, mainly utilizing **natural language processing (NLP)** and **machine learning (ML)** techniques [84, 85, 141]. The scientific literature contains studies on extracting malware indicators, attack patterns, and generating cyberthreat alerts. Systematizing these studies, categorizing the CTI extraction purposes, and identifying associated techniques facilitate the extension and improvement of current work and introduce future research paths in the CTI extraction domain.

The goal of this article is to aid cybersecurity researchers in understanding the current techniques used for cyberthreat intelligence extraction from text through a survey of relevant studies in the literature. To achieve this goal, we collect existing studies on the automatic extraction of CTI from textual descriptions using a keyword search. We use open coding [152] and card sorting [156] techniques to perform a qualitative analysis of these studies. We list our contributions as follows:

(a) A systematic categorization of the CTI extraction purposes (such as extracting malware indicators and attack patterns from the text) performed in this studies[1] (Section 5),
(b) CTI extraction pipeline, where the pipeline abstracts the steps for CTI extraction observed in the studies (Section 6),
(c) A categorization of NLP and ML techniques associated with CTI extraction (Section 6.4),
(d) A set of textual data sources and CTI sharing formats utilized in the studies for CTI extraction and sharing, respectively (Section 6.1 and 6.5),
(e) A compilation of application scenarios of the extracted CTI demonstrated by the researchers in the studies (Section 6.6).
(f) A set of recommendations for cybersecurity researchers in conducting future research in the CTI extraction domain (Section 10).

The rest of the article is organized as follows: In Section 2, we discuss CTI and its role in proactive defense. In Section 3, we discuss the related survey studies. In Section 4, we discuss our

[1]Throughout this article, *studies* refers to the set of studies related to CTI extraction from the text in this survey.

methodology. In Section 5, we discuss the types of CTI extraction purposes observed in the studies. In Section 6, we discuss the CTI extraction pipeline along with each step in the pipeline. In Sections 8, 9, and 10, we discuss several insights on the selected studies, limitations, and further research directions. In Section 11, we conclude the article.

2 CYBERTHREAT INTELLIGENCE

Cybersecurity is a balancing act between attackers and targeted entities/organizations [154]. The attackers constantly probe to exploit security weaknesses in the system, while the organizations constantly monitor and defend their malicious attempts. The modern IT infrastructure can be susceptible to security weaknesses, such as insecure coding practices, zero-day exploits, inconsistent patching, vulnerable third-party libraries, data exposure, human error, and social engineering. These factors provide the attackers with advantages in launching cyberattacks. Thus, handling cyberthreats in a *reactive* manner, such as only responding to security incidents, can make organizations more vulnerable to cyberthreats. To address this issue, organizations need to be *proactive* about defending against attackers who are always on the move in an ever-changing threat landscape.

2.1 Definition

Cyberthreat intelligence (CTI), also known as threat intelligence, can be used as a proactive defense mechanism against cyberattacks. McMillan et al. [148] define CTI as "evidence-based knowledge, including context, mechanisms, indicators, implications, and actionable advice about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding the subject's response to that menace or hazard." Dalziel et al. define CTI as refined, analyzed, or processed data that should be at least potentially relevant to the organization and objectives, specific enough to prompt response, action, or decision and contribute to valuable business outcome [41]. Although we provide several definitions of CTI, there is no universally accepted definition of CTI [45, 120, 125]. Hence, we consider any cybersecurity-relevant information as CTI that aids in predicting, preventing, or defending an attack, shortening the window between compromise and detection, and helping to clarify the risk landscape [91, 125, 148]. Any arbitrary security-relevant information can serve as CTI if the information provides *actionable* intelligence to protect organizations from cyberattacks as defined by Wagner et al. [133]. The author reports two definitions of actionability from Pawlinski et al. [105] and Ponemon Institute [150]. According to the definition, security information becomes actionable CTI if the information fulfills the following criteria: relevance, timeliness, accuracy, completeness, ingestibility, priority, implementation, the trustworthiness of the source, relevance to the industry, clear guidance to resolve the threat, and sufficient context.

2.2 CTI Subdomains

According to Tounsi et al. [125], CTI can be categorized into four subdomains as follows:

(a) **Strategic CTI**, which is the information generally in the form of reports, briefings, or conversations that help security decision-makers understand and identify current and future risks. This information includes financial effects, trends, and historical data on cyberattacks. This kind of CTI is targeted for the nontechnical audience [39].
(b) **Operational CTI**, which is information about specific impending attacks on an organization. This kind of CTI helps organizations understand the relevant factors of specific attacks, such as nature, timing, intent, and threat actors. Security professionals use this CTI while working in security operation centers [39].

(c) **Tactical CTI**, which is often referred to as **Tactics, Techniques, and Procedures (TTPs)** [38, 125]. *Tactic* is the highest-level description of a cyberthreat—the adversary's goal ("why"). *Technique* reflects on "how" adversaries can achieve a tactic. *Procedures* are even lower-level, highly detailed descriptions (containing tools and attack group description) in the context of a technique. Security experts use the tactical CTI to improve their defense strategies for current tactics. Technical press, white papers, and threat reports are sources for tactical CTI.

(d) **Technical CTI** is the set of information usually consumed through computational resources. For example, analytical and visualization tools can consume information (such as IP address and packet contents) collected by firewalls regarding DDoS attacks. Technical CTI includes the collection of **indicators of compromise (IoC)**. IoC refers to the "forensic artifacts of potential intrusions on a host system or network" [26], which includes information of malicious artifacts such as malicious IP addresses, URLs, malware hashes and signatures, anomalous network traffic, and TTPs. Technical CTI usually feeds information towards investigative and monitoring activities inside an organization.

2.3 Uses of CTI

The advantages of utilizing CTI are numerous, which we discuss below.

(a) **Proactive and actionable defense:** In proactive defense, organizations predict future cyberthreat strategies and incorporate these insights into the defense mechanisms of the system [57]. Identifying CTI from previous threats, analyzing the identified CTI information, and deriving actionable insights are helpful keys to preparing a system for proactive defense. Moreover, modern-day cyberattacks can use multiple means to get propagated and can be active in multiple stages. For example, attackers can infiltrate the network first, spread laterally across all devices, and then compromise systems through vertical propagation. These attacks include zero-day exploits, social engineering, and malicious client-side scripts, which help the adversaries evade the traditional cyberattack detection and defense mechanism [125]. The use of CTI can help organizations prevent these attacks.

(b) **Constructing threat profiles:** Organizations can build threat profiles for well-known cyberattack groups from CTI, which can aid organizations in securing their defense mechanisms based on the attack tactics and techniques from the threat profile.

(c) **Information sharing and awareness building:** The collective sharing and exchanging of information and knowledge gained from CTI information can accelerate the learning and awareness among organizations to prevent cyberattacks. For example, if organizations find an intruder in the active phase of an attack, then there are greater chances of defending against the attack through collective sharing [157]. Moreover, CTI sharing is a cost-effective tool for thwarting cybercrime [150]. Organizations can aggregate and synthesize the obtained information to understand the upcoming cyberthreats, trends in cyberattacks, and evolution in the patterns of cyberattacks. Thus, practitioners and researchers could benefit from the shared knowledge to make better decision-making in proactive defense against cyberattacks.

(d) **Cybersecurity research:** CTI-related information, such as attack indicators and TTPs, can be utilized by cybersecurity researchers to develop new insights on cyberthreat domains.

2.4 CTI Extraction

According to the definitions of CTI presented in Section 2.1, CTI is a set of information that organizations can use to prevent and defend against cyberattacks. Hence, any device-to-device

communication data, network/system/application logs, stack traces, malware signatures, news/reports/articles of attacks in the context of a cybersecurity-related incident are sources for CTI. However, the usefulness of collected CTI depends on the extracted information's accuracy, relevance, and soundness. Human operators first orchestrated the CTI extraction process from these sources above, which proved to be error-prone and inefficient, paving the opportunity for automatic extraction of CTI-related information [143]. As the number and means of cyberattacks have grown with time, the textual descriptions and analysis of cyberthreat-related documents, reports, online articles, and online community interactions have also increased. These textual documents of cyberthreat-related activities are a valuable source for collecting CTI-related information. Cybersecurity researchers have already explored automated extracting CTI information from these textual sources. This survey focuses on their CTI extraction goals, sources, approaches, and how practitioners can utilize the information. Threat-related activities such as attackers' TTPs; tools used by the attackers such as malicious scripts; details of malware, zero-day bugs, vulnerabilities, and exploits; cyberthreat-related news, and events-related information can be found in this textual representation. Cybersecurity vendors and practitioners produce this information through monitoring threat activities and collecting raw data from firewalls, network logs, and past security incidents. These data can help researchers and practitioners gain insight into the threat landscape and form strategies to secure the systems from cyberattacks.

2.5 CTI-candidate Text

Cybersecurity vendors publish cyberattack-related reports (also known as threat reports). Cybersecurity researchers and practitioners publish online blogs and articles on cyberattacks and post informative content on social media and forums. The reports, online articles, social media, and hacker forum posts describe in unstructured natural language (i.e., expressed in English) the step-by-step attack procedure, related vulnerabilities, and pertinent aspects of cybersecurity such as attack patterns and mitigation steps. We collectively refer to these reports, articles, and posts as CTI-candidate texts as these documents are sources for CTI extraction.

3 RELATED STUDIES

This section discusses related literature reviews or survey studies on different aspects of CTI.

Trends in CTI research direction and CTI sharing: Tounsi et al. [125] investigated the sources for gathering CTI, how organizations use CTI, and CTI sharing platforms. They identified that obtaining related CTI is challenging for organizations because of the large amount of available CTI, and organizations need to filter the CTI based on relevance. They also evaluated the most popular and available CTI gathering tools and compared their features. However, the authors primarily focused on only technical CTI (i.e., indicators of system and network-level compromise) specific sharing tools, and platforms. The authors emphasize that organizations can obtain CTI from different sources. Thus, to aid in better understanding and management of CTI, they proposed four specific subdomains of CTI, which we discuss in Section 2.2. Sauerwein et al. [120] conducted a study of 22 CTI-sharing platforms that enable automation of the generation, refinement, and examination of security data. Their investigation resulted in eight key findings, such as: "There is no common definition of threat intelligence sharing platforms" and "Most platforms focus on data collection instead of analysis." Wagner et al. [133] explored the current state-of-the-art approaches and technical and non-technical challenges in sharing CTI. They used articles from academic and gray literature. Their investigation covers widely discussed topics in CTI sharing, such as establishing a collaboration between stakeholders and automating parts of the CTI sharing process.

Threat analysis: Tuma et al. [126] performed a systematic literature review on 26 methodologies of cyberthreat analysis. They compared methods based on different aspects such as

applicability, characteristics of the analysis procedure, outcomes, and ease of adoption. They also enlighten the limitations of adopting the existing approaches and discuss the current state of their adoption in software engineering processes. Their observations indicate that the threat analysis procedure is not clearly defined and lacks quality assurance and tool support. Xiong et al. [136] performed a systematic literature review on threat modeling. They reviewed 54 articles and identified three types of articles among these: (i) articles that are contributing (such as introducing a new method) to the field of threat modeling, (ii) articles that are using an existing threat modeling method, and (iii) articles that are presenting work related to the threat modeling process. They observed that most threat modeling work is done manually with limited assurance of their validation.

Cybersecurity information extraction: Bridges et al. [53] evaluated existing methods [51, 77, 78] of accurate and automatic extraction of security entities from text. They used online news and blog articles, websites of **common vulnerabilities and exposures (CVE)**, the **National Vulnerability Database (NVD)**, and Microsoft security bulletins. After comparing the existing approaches, the authors concluded that the existing methods have a low recall, and no large publicly available annotated dataset of security documents is available. Overall, these previous researches focus on CTI from various perspectives such as privacy, sharing, modeling, and performance.

Previous work by authors: In Reference [112], we conducted a systematic literature review on 34 CTI extraction studies. We identified eight data sources for collecting CTI-candidate texts and seven CTI extraction purposes. We identified the natural language processing techniques used for CTI extraction. However, in this study, we expand our previous work by investigating a more significant number of relevant studies, identifying three new CTI extraction purposes, and proposing a CTI extraction pipeline that prospective researchers for CTI extraction can instantiate.

4 METHODOLOGY

This section explains the process of searching, selecting, and analyzing the studies.

4.1 Search Strategy

The first step is to find relevant studies from scholarly databases. We select six scholarly databases to conduct our search: **Institute of Electrical and Electronics Engineers (IEEE)** Xplore,[2] **Association for Computing Machinery (ACM)** Digital Library,[3] ScienceDirect,[4] SpringerLink,[5] Wiley Online Library,[6] and DBLP.[7] We construct a set of search strings to identify relevant studies in the selected scholarly databases. We search each of the six scholarly databases using the search queries below. In total, we find 20,922 publications after removing duplicates.

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

4.2 Selection of Relevant Studies

Our search result includes publications not relevant to extracting CTI from the text. Hence, we establish inclusion and exclusion criteria to filter the irrelevant publications:

[2] https://ieeexplore.ieee.org/Xplore/home.jsp.
[3] https://dl.acm.org/.
[4] https://www.sciencedirect.com/.
[5] https://link.springer.com/.
[6] https://onlinelibrary.wiley.com/.
[7] https://dblp.org/.

Exclusion criteria: (a) Publications that are not peer-reviewed: keynote abstracts, call for papers, and presentations; or (b) publications date before 2000; or (c) publications written in languages except English. **Inclusion criteria:** (a) Publications available for downloading or reading on the web; and (b) title, keywords, and abstract of the study explicitly indicating that the publication: (i) uses unstructured textual documents as CTI source, (ii) automatically extracts cybersecurity-relevant information and knowledge relevant for proactive defense against cyberattack, (iii) uses NLP and ML techniques to extract the information and knowledge from unstructured text.

The first two authors perform the filtering. The first author filtered the search results manually through the inclusion and exclusion criteria. The second author used FAST^2 [137], an intelligent tool for publication selection in literature reviews. Carver et al. [58] show that selecting publications in systematic literature reviews is one of the most challenging and time-consuming tasks. Therefore, using an intelligent tool can make selecting studies more efficient. Using FAST^2, the second author starts picking key publications from the list of papers found in the search results. These key publications are the relevant publications the authors first studied before searching the scholar database for the relevant studies (Section 4.1). Then a model is trained by the tool based on the title and abstract of the studies. The model ranks the rest of the studies based on relevance, showing a list of 10 candidate studies. The second author selects the relevant studies from the list of candidates, and this feedback makes the model more accurate in each iteration. Thus, the tool can predict the number of relevant publications through supervised learning in the list based on user feedback. The tool stops when 95% of all relevant publications are selected.

After the first two authors finish the filtering, we obtain a set of 50 publications. We refer to these 50 studies as the initial set. After a detailed read of these publications, we find irrelevant studies. For example, Iqbal et al. [75] suggested an approach that threat analysts *manually* find CTI from text reports and add the information to a database for STIX[8] generation. With discussions between the first two authors on the relevance of each publication, we select 33 publications from the initial set. We refer to these 33 studies as Study Set A. Then, we apply forward and backward snowballing [130] on Study Set A. We perform forward snowballing by collecting publications that cite Study Set A. Then, we perform backward snowballing by collecting the publications cited by Study Set A. We next apply the exclusion and inclusion criteria on these snowballed publications, which gets us a new set of 51 publications. We refer to these 51 studies as Study Set B. Finally, combining Study Set A and B, we get a total of 84 studies that we select for our survey study.

4.3 Quasi Gold Set

We use three search strings to search relevant papers from the scholarly database. The search results from these search strings might include irrelevant studies and miss relevant studies. We validate the three search strings by applying ***quasi sensitivity metric* (QSM)** proposed by Zhang and Babar [139]. The quasi-sensitivity metric validates whether the chosen search strings are sufficient for finding the relevant studies from the scholarly databases. Calculating the value of the QSM first requires a **quasi-gold set (QGS)** of publication. We construct the QGS by first searching for the relevant studies in the top 10 computer security and cryptography-related conferences and journals listed in the Google Scholar top publications [33] from 2010 to 2021. We find seven studies from the search. Next, we apply forward and backward snowballing to the seven studies and find another 45 relevant studies. Thus, in total, we find 52 studies for the QGS. The QSM is calculated by the following equation: $QSM = \frac{Count_{SS}}{Total_{QGS}}$. $Count_{SS}$ refers to the count of publications from the selected studies in QGS. $Total_{QGS}$ refers to the total number of publications in the QGS.

[8] https://oasis-open.github.io/cti-documentation/stix/intro.html.

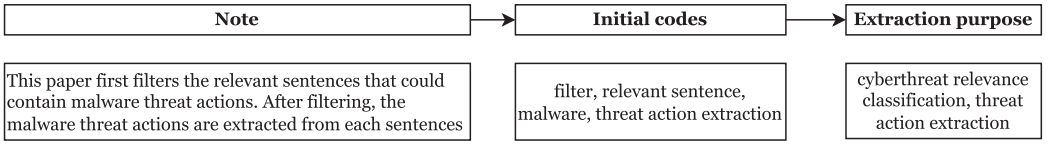

Fig. 1. An example of how the authors apply open coding to identify CTI extraction purposes.

Table 1. CTI Extraction Purposes

Purpose	Studies	Count	Subdomain	Cohen's κ
CRC	[42, 46, 48, 52, 54, 61, 62, 64, 70–72, 76, 81, 83, 84, 87, 90, 92, 96, 99, 102, 108, 113, 122, 124, 128, 135]	27	Strategic	0.73*
ICE	[48, 56, 82, 84, 88, 102, 123, 124, 141–143, 145]	12	Technical	0.95**
TTP	[43, 60, 71, 72, 89, 101, 103, 123, 129, 135, 140, 144]	12	Tactical	0.90**
VIE	[49, 69, 76, 77, 83, 87, 95, 97, 98, 100, 109, 115]	12	Technical	0.95**
CWT	[47, 50, 59, 62, 73, 74, 85, 94, 96, 104, 135]	11	Strategic	0.73*
CTE	[44, 54, 55, 65, 79, 80, 86, 110, 113, 127]	10	Strategic	0.80**
CEE	[64, 68, 93, 107, 108, 111, 132, 134]	8	Technical	0.88**
HRA	[52, 63, 67, 99, 114, 116, 119, 131]	8	Strategic	0.89**
CAG	[63, 97, 100, 113, 118, 121]	6	Operational	0.92**
CTA	[67, 74, 103, 106, 119]	5	Tactical	1.00***
TRC	[43, 44, 56, 138]	4	Tactical	0.79*

*: substantial agreement, **: almost perfect agreement, ***: perfect agreement.

We identify 43 papers in the selected study and the QGS. Hence, $Count_{SS} = 43$, $Total_{QGS} = 52$ and the calculated value of the QSM is $\frac{43}{52} = 0.83$, which is mentioned as *acceptable* in Reference [139].

4.4 Analyzing the Studies

After selecting the studies, the first two authors read the 84 papers. While reading each publication, they take note of the following information from each study: (a) **CTI extraction purpose:** the security information extracted from the text and the information gained from further analysis of the extracted information, (b) **Data source:** the details of the dataset obtained and used in the studies, (c) **Methodology:** the steps followed by the authors in their methodology.

After reading the publications and taking notes, the first two authors apply the open coding technique [152] on the three sets of notes mentioned above. Figure 1 demonstrates how we use open coding to identify the CTI extraction purposes. As the coding process is subjective, the authors resolve the disagreements and agree on the following: (a) 11 types of CTI extraction purposes (Section 5); (b) seven types of data sources to extract CTI (Section 6.1); (c) the pipeline consisting of six steps for CTI extraction process, as shown in Figure 3; and (d) the categorization of techniques and CTI sharing formats (Sections 6.4, and 6.5). We report our findings in the following sections.

5 CTI EXTRACTION PURPOSE

Using the methodology described in Section 4.4, we identified 11 purposes for extracting CTI information. Table 1 reports the CTI extraction purpose, corresponding studies, count, corresponding CTI sub-domain, and the agreement score between the first two authors (Cohen's Kappa [147]). One study can have more than CTI extraction purposes. We discuss these purposes below.

5.1 Identified CTI Extraction Purposes

5.1.1 CTI Relevance Classification (CRC) (n = 27). CTI relevance classification refers to the classification of CTI-candidate texts to (a) CTI relevance and (b) cybersecurity contexts such as cyberthreat topics and types. To extract information from the CTI-candidate texts (textual sources for collecting CTI information), researchers must first determine whether these texts are CTI-related. For example, to extract IoCs from a list of Twitter posts, online blogs, and articles, in References [124] and [84], the IoC-related posts are filtered from the non-relevant posts through text classification. In References [44, 46, 48, 54, 64, 76, 83, 90, 102, 113, 122, 128], the authors classified whether the Twitter posts are CTI-related or not. In References [61, 62, 70, 81, 96, 99], the authors determined whether hacker forum posts are CTI-related. In References [87, 108, 135], the authors determined whether online blogs and articles are CTI-related. In References [71, 72], the authors used text classification to filter TTPs containing sentences from CTI-candidate texts. In Reference [52], the authors classified the forum posts into three types of cyberthreats: vulnerabilities, financial, and others. In Reference [92], the authors classified the online blogs and articles into 12 cyberthreat topics. In Reference [42], the authors classified the hacker forum posts to attack target platform (Windows or Linux) and attack type (local or remote). Classifying CTI-relevant texts is a precursor to other CTI extraction purposes, such as IoC or TTPs extraction. Because extracting IoCs, attack patterns, and cyberthreat events information from text includes the process of CTI-related text classification and the gained insight can serve as strategic CTI.

5.1.2 Indicators of Compromise Extraction (ICE) (n = 12). Indicators of compromise extraction refers to the extraction of indicators and traces of compromise from CTI-candidate texts, which provide evidence of malicious activities or compromises in systems. Extracting these **indicators of compromise (IoC)**, such as malware names, signatures, hashes, IP addresses, and packets, facilitates further research and analysis opportunities for security research and practitioners. In the studies, IoCs have been extracted from social media [48, 102, 124], threat reports [56, 82, 123, 142, 145], and online articles [84, 88, 143, 145]. Moreover, in Reference [141], authors extract specific indicators of malware delivery attempts by **advanced persistent threat (APT)** attack groups from threat reports. These extracted IoCs serve as technical CTI and aid organizations in learning about future attack attempts, analyzing malware behavior, and designing anti-malware tools.

5.1.3 Attack Tactics, Techniques, and Procedures Extraction (TTP) (n = 12). Threat reports contain verbose details on how malicious attacks are being performed. Attack tactics, techniques, and procedure (TTPs)-related information can be extracted from these textual reports. Attack tactics, techniques, and procedures identification refers to (a) identification of threat actions in CTI-candidate text, (b) classification of CTI-candidate texts to MITRE ATT&CK[9] TTPs and Cyber-kill-chain stages. In References [60, 71, 72, 101, 123, 140], the authors extracted threat actions from threat reports. In References [71, 72, 129], the authors classified the extracted threat actions to Cyber-kill-chain stages.[10] In References [43, 89, 103, 135, 144], the authors classified the CTI-candidate texts to MITRE ATT&CK tactics and techniques. Extracted threat actions and TTPs serve as tactical CTI and help the defending organizations plan to prevent themselves from being attacked.

5.1.4 Vulnerability Information Extraction (VIE) (n = 12). Vulnerability information extraction refers to the (a) extraction of software vulnerability-related information from CTI-candidate texts and (b) identification of vulnerability-related texts from CTI-candidate texts. Open-source

[9] https://attack.mitre.org/.
[10] https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html.

repositories of version control systems contain information on vulnerable software packages, insecure development practices, security bugs, issues, and developers' discussions. Organizations can extract software vulnerability-related information to proactively gain insight into securing the systems. For example, in Reference [100], the authors have extracted vulnerable package-related information from Github repositories. In Reference [98], the authors extracted vulnerability trends and patterns from the CVE[11] from the NVD[12] description. In References [69, 77, 87, 95, 97], the authors extracted the vulnerability and related keywords from the vulnerability description in NVD. They extracted their underlying concepts using ontology and online resource descriptors to facilitate cybersecurity practitioners to query and reason over those linked concepts using a graph-based knowledge base. In References [49, 76, 109, 115], the authors classified whether the hacker forum and Twitter posts are vulnerability-related or not. This vulnerability information can serve as technical CTI and can be used to proactively secure software artifacts from vulnerable artifacts.

5.1.5 Cyberthreat Relevant Word and Topic Identification (CWT) (n = 11). Cyberthreat-relevant word and topic identification refers to identifying terms and topics related to cyberthreats from CTI-candidate texts. In References [47, 73, 85, 94, 96], the authors identified emerging and trending cyberthreat-related terms from hacker forum posts and Twitter. In References [50, 59, 62, 74, 104, 135], the authors identified emerging and trending cyberthreat-related topics from hacker forum posts, threat reports, online blogs, and articles. Thus, cyberthreat relates word and topic identification help cybersecurity researchers and practitioners to gain insight into the threat landscape and serves as strategic CTI.

5.1.6 Cyberthreat Event Identification (CTE) (n = 10). Cyberthreat event identification refers to whether CTI-candidate texts describe cyberattack events or incidents. In the studies, Twitter is explored mostly for identifying cyberthreat incident-related posts and extracting the incident-related information in References [54, 55, 65, 79, 80, 86, 110, 113]. In References [44, 127], the authors extracted 9 and 30 types of cyberthreat events, respectively, from threat reports and online articles. These extracted threat-related events serve as strategic CTI, which can help cybersecurity practitioners and researchers to understand past events, monitor future threats, and share the CTI information with other cybersecurity information-consuming organizations.

5.1.7 Cyberthreat Relevant Entities Extraction (CEE) (n = 8). Cyberthreat-relevant entities extraction refers to the extraction of named entities related to cybersecurity concepts and cyberattacks such as malware name and target organization. In References [64, 107, 132, 134], the authors extracted malware-related named entities from online blogs and articles, threat reports, and Twitter. In References [68, 93, 108, 111], the authors extracted named entities on the basis of STIX format from CTI-candidate texts. Thus, extracting named entities can help cybersecurity practitioners to build a knowledge base of cybersecurity concepts, export information to a structured format, and serves as technical CTI.

5.1.8 Hacker Resource Analysis (HRA) (n = 8). Hacker resource analysis refers to the text mining performed on attachments and resources posted in hacker and darknet forums. Hacker and darknet forums are rich in hacking tools, documents, scripts, source code, and online resources regularly used by the attackers' community. Researchers have analyzed these resources in their studies to gain further insight into the attackers' motives and techniques. For example, in References [67, 114, 116, 119, 131], the authors proposed a platform named AZSecure, which facilitates security researchers and practitioners to search and visualize cutting-edge hacking tools, scripts, malware,

[11] https://cve.mitre.org/.
[12] https://nvd.nist.gov/.

and other relevant assets to launch cyberattacks. In References [52, 63, 99], the authors classified the malicious attachments posted in hacker forums as malware types such as keyloggers, worms, spyware, and so on. Extracting CTI from these sources serves as strategic CTI and provides insight into cutting-edge hacking tools and trends on emerging threat-related issues.

5.1.9 Cyberthreat Alert Generation (CAG) (n = 6). Cyberthreat alert generation refers to the generation of alerts on emerging attack patterns, tools, malware, and vulnerabilities from CTI-candidate texts. For example, in References [97, 113, 118, 121], the authors have generated threat alerts based on Twitter[13] posts by security experts. Moreover, researchers explore darknet and hacker community discussion threads to warn about future cyberattacks in References [63, 118, 121]. In Reference [100], the authors have generated alerts based on vulnerabilities found in software repositories. Organizations can generate these warnings, which can serve as operational CTI.

5.1.10 Cyberthreat Attribution (CTA) (n = 5). Cyberthreat attribution refers to identifying corresponding malicious actors from their attack techniques, indicators, malware, and so on. CTI-candidate texts often contain information on cyberthreat incidents and associated cyberattack actors, such as their roles, strategies, and procedures. Organizations can use this information to map the attacks to the responsible cyberthreat actors. For example, in References [103, 106], the authors identified the responsible attack groups from the threat reports, where the group(s) are responsible for launching cyberattacks through propagating malware. In References [67, 74, 119], the authors identified top malicious users on their posted malicious scripts, attachments, discussed topics, and user influence. Thus, cyberthreat actor attribution serves as a strategic CTI and can aid in gaining tactical insight into the attackers' strategies and malicious activities.

5.1.11 Threat Report Categorization (TRC) (n = 4). Threat report categorization refers to classifying or clustering the threat reports on attack techniques or types. For example, in References [56, 138], the authors classified and clustered the threat reports based on the malware's name. In Reference [43], the authors classified the threat reports on MITRE ATT&CK TTPs. As multiple cybersecurity vendors publish threat reports, grouping threat reports help cybersecurity researchers and practitioners to identify similar cyberattack threats and attack patterns, which can serve as tactical CTI.

5.2 Association among CTI Extraction Purposes

CTI extraction purposes can have associations with one another. For example, in Section 5.1, we mention that CTI relevance classification serves as a precursor to other CTI extraction purposes, such as IoC or TTPs extraction. We report the identified associations among CTI extraction purposes in Figure 2. In the figure, (a) each rectangle represents CTI extraction purposes, (b) each direction represents the association between the CTI extraction purposes observed in the study, and (c) the box on the right side of the image shows the studies demonstrating the association. For example, in the figure, there is a direction from **CTI relevance classification (CRC)** to **Attack tactics, techniques, and procedures extraction (TTP)**. The direction along with the right side box indicates that: (a) two studies [71, 72] first filter the attack pattern describing sentences from the CTI-candidate texts using text classification, (b) then the two studies extracted threat actions from the CTI-candidate texts. Hence, we observe the dependency between CRC and TTP where authors first performed CRC to perform TTP afterward. We observe several extraction purposes such as CWT, HRA, CEE, VIE, ICE, CTE, and TTP depend on CRC. We identify 18 studies that first

[13] https://twitter.com.

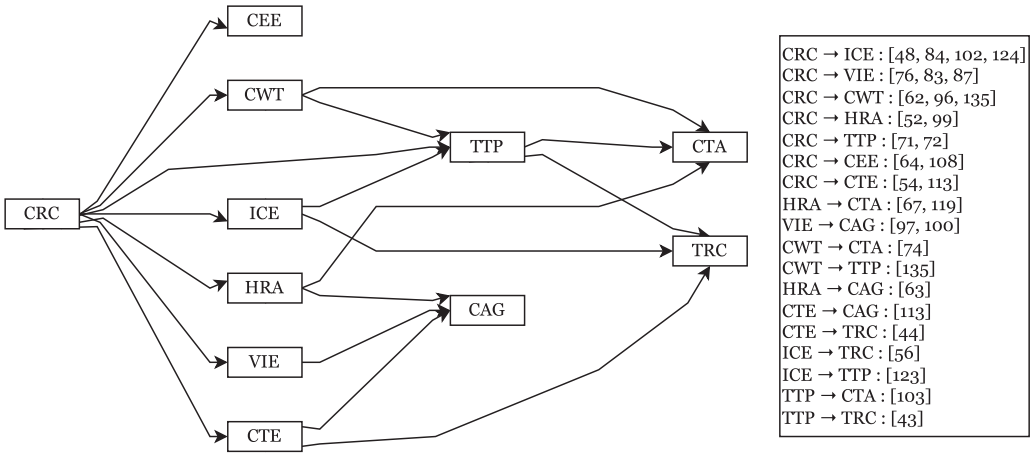

Fig. 2. Association among the CTI extraction purposes.

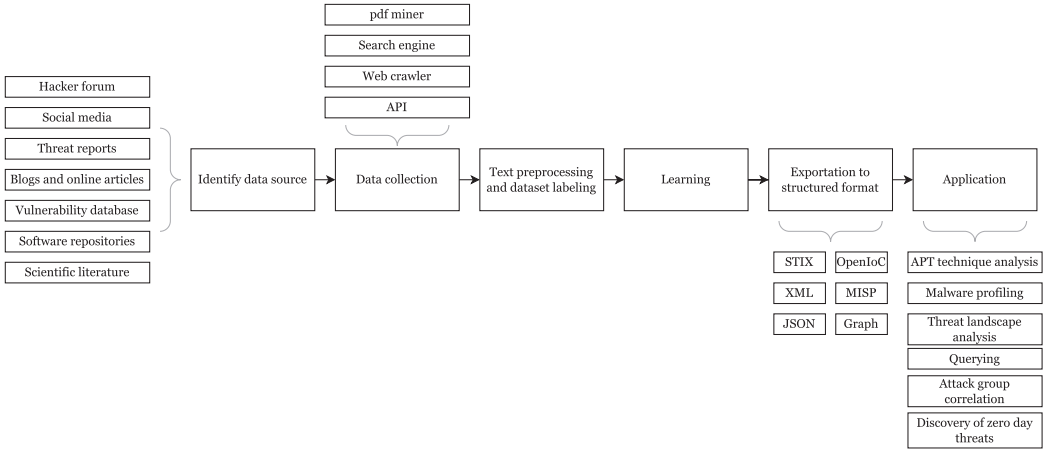

Fig. 3. CTI extraction pipeline.

performed CTI relevance classification and then performed the subsequent CTI extraction, such as an indicator of compromise extraction and hacker resource analysis. We also observe authors perform cyberthreat alert generation after hacker resource analysis, vulnerability information extraction, and cyberthreat event generation. We identify 29 studies associated with more than one CTI extraction purpose and report those studies in Figure 2.

6 CTI EXTRACTION PIPELINE

Extracting CTI-related information from CTI-candidate text requires NLP in combination with ML techniques. Although the extraction techniques depend on the types of data pulled from the text (e.g., the extraction procedure of IoCs can be different than that of TTPs), the abstracted pipeline stays similar. Hence, we propose a CTI extraction pipeline, which abstracts the extraction approaches found in the studies. Thus, the pipeline can be instantiated based on the CTI extraction purposes and provide cybersecurity researchers with possible options to design their own CTI extraction pipeline. Using the methodology described in Section 4.4, we identify six steps, and the agreement score is 1.00. In Figure 3, we show the steps of the pipeline. The pipeline has these six

Table 2. Data Sources for CTI Extraction in the Studies

Study	Purpose	Sources	Type	Size	Dataset
[124]	CRC, ICE	Twitter	SM	22K	[21]
[103]	TTP, CTA	APT Notes	TR	327	[15]
[134]	CEE	*NM*	TR	10K	*NM*
[101]	TTP	APT Notes	TR	445	[15]
[122]	CRC	Twitter	SM	200K	*NM*
[55]	CTE	Twitter	SM	21K	[9]
[68]	CEE	FireEye,[14] Kaspersky Security Lab,[15] Apt Notes	TR	50	[15]
[84]	CRC, ICE	AlienVault,[16] FireEye, MalwareBytes[17]	OB	71K	*NM*
[47]	CWT	APT Notes	TR	147	[19]
[138]	TRC	Threat Expert*	TR	25K	*NM*
[43]	TTP, TRC	Symantec,[18] FireEye, McAfee[19]	TR	18K	*NM*
[132]	CEE	FireEye, Krebs On Security,[20] Securelist,[21] F-secure,[22] Crowdstrike[23]	OB	13K	*NM*
[116]	HRA	OpenSC*	HF	5K	*NM*
[141]	ICE	WeLiveSecurity,[24] TaoSecurity,[25] Malwarebytes, Trend Micro[26]	OB	14K	[3]
[62]	CRC, CWT	nulled.io*	HF	16K	[28]
[88]	ICE	APT Notes	TR	687	[10]
[54]	CRC, CTE	Twitter	SM	21K	[9]
[107]	CEE	FireEye, Kaspersky	TR	474	*NM*
[80]	CTE	Twitter	SM	5.1B	[25, 30]
[44]	CTE, TRC	Recorded Future,[27] FireEye, Security Week,[28] Trend Micro	TR	*NM*[29]	*NM*
[64]	CRC, CEE	Twitter	SM	11K	[18]
[97]	VIE, CAG	Twitter	SM	144K	*NM*
[113]	CRC, CTE, CAG	Twitter	SM	350K	[20]
[99]	CRC, HRA	*NM*	HF	5.4k	*NM*
[46]	CRC	Twitter	SM	195K	*NM*
[65]	CTE	Twitter	SM	5K	[11]
[73]	CWT	CrimeBB	HF	42M	[2]
[121]	CAG	Twitter, web crawler[30]	SM, HF, OB	764	*NM*
[118]	CAG	Twitter, *NM*	SM, HF	-	*NM*
[109]	VIE	Twitter, Cracking Arena*, Dream Market*	SM, HF	148K	[1]
[144]	TTP	MITRE ATT&CK	TR	243	[36]

(Continued)

[14] https://www.fireeye.com/current-threats.html.
[15] https://www.kaspersky.com/enterprise-security/resources/white-papers.
[16] https://cybersecurity.att.com/blogs.
[17] https://blog.malwarebytes.com/.
[18] https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence.
[19] https://www.mcafee.com/enterprise/en-us/resource-library/publications.html.
[20] https://krebsonsecurity.com/.
[21] https://securelist.com/.
[22] https://www.f-secure.com/en/home/articles.
[23] https://www.crowdstrike.com/blog/.
[24] https://www.welivesecurity.com/.
[25] https://taosecurity.blogspot.com/.
[26] https://www.trendmicro.com/vinfo/us/security/news/blogs.
[27] https://www.recordedfuture.com/blog.
[28] https://www.securityweek.com/.
[29] NM stands for "not mentioned" in the corresponding study.
[30] The CTI-candidate texts are collected by a web crawler, however, the name/source of the crawler have not been mentioned.

Table 2. Continued

Study	Purpose	Sources	Type	Size	Dataset
[119]	HRA, CTA	OpenSC*	HF	432K	*NM*
[114]	HRA	Hackfive*, Hackhound*, Icode*	HF	672K	*NM*
[50]	CWT	Cracking Fire*	HF	38K	[1]
[81]	CRC	Sixgill Crawler	HF	3K	[4]
[52]	CRC, HRA	*NM*	HF	1.3M	*NM*
[61]	CRC	nulled.io*	HF	16K	*NM*
[77]	VIE	Microsoft[31] and Adobe[32] security bulletins, National vulnerability database (NVD)[33]	TR, VD, OB	320	[34]
[95]	VIE	NVD, CNET[34]	VD, OB	155	*NM*
[93]	CEE	*NM*	TR, OB	2K	[17]
[89]	TTP	*NM*	TR	*NM*	[22]
[123]	ICE, TTP	APT Notes, Microsoft, Symantec, Threat Encyclopedia,[35] VirusRadar[36]	TR	8K	[16]
[79]	CTE	Twitter	SM	5B	[25, 30]
[140]	TTP	IEEE Symposium of Security and Privacy,[37] IEEE Computer Security Foundation Symposium,[38] USENIX Security Symposium[39]	SL	1K	[6]
[90]	CRC	Twitter, NVD	SM, VD	76K	*NM*
[70]	CRC	Cracking Arena*	HF	45K	[1]
[74]	CWT, CTA	nulled.io*, hackthissite.org*, hiddenanswers*, breachforum*, raid*	HF	316K	*NM*
[96]	CRC, CWT	*NM*	HF	26K	*NM*
[67]	HRA, CTA	Hackhound*, Ashiyane*, VBSpiders*, Zloy*	HF	482K	*NM*
[142]	ICE	FireEye, Kaspersky	TR	400	*NM*
[131]	HRA	OpenSc*, Cracking Zilla*, AntiOnline[40]	HF	1.3M	*NM*
[69]	VIE	NVD	VD	20k	[23]
[83]	CRC, VIE	*NM*, Twitter	HF, SM, OB	*NM*	[32]
[127]	CTE	The Hacker News[41]	OB	*NM*	*NM*
[102]	CRC, ICE	Twitter	SM	2.3k	[12]
[129]	TTP	Fireeye, McAfee, Kaspersky	TR	108	[15]
[100]	VIE, CAG	Github[42]	SR	111K	*NM*
[98]	VIE, CAG	NVD	VD	26K	[29]
[94]	CWT	0x00sec.org	HF	9K	*NM*
[63]	HRA, CAG	Dream Market*, Darkshades market*	HF	48K	*NM*
[106]	CTA	Attack Attribution Dataset	TR	249	[13]
[128]	CRC	Twitter	SM	21K	[9]
[92]	CRC	*NM*	OB	30K	*NM*
[135]	CRC, TTP, CWT	*NM*	OB, HF	207K	[24]
[42]	CRC	*NM*	HF	33K	[14]
[48]	CRC, ICE	Twitter	SM	195K	*NM*

(Continued)

[31] https://docs.microsoft.com/en-us/security-updates/.
[32] https://helpx.adobe.com/security/security-bulletin.html.
[33] https://nvd.nist.gov/.
[34] https://cnet.com.
[35] https://www.trendmicro.com/vinfo/us/threat-encyclopedia/.
[36] https://www.virusradar.com/.
[37] https://ieeexplore.ieee.org/xpl/conhome/9144328/proceeding.
[38] https://www.ieee-security.org/CSFWweb/.
[39] https://www.usenix.org/conferences/byname/108.
[40] https://www.antionline.com/.
[41] https://thehackernews.com/.
[42] https://github.com.

Table 2. Continued

Study	Purpose	Sources	Type	Size	Dataset
[145]	ICE	McAfee, Securelist, WeLiveSecurity, Github repository, Trend Micro, Virus Bulletin,[43] Sucuri,[44] Naked Security[45]	TR, OB	3K	*NM*
[59]	CWT	Cracking Arena*	HF	45K	*NM*
[85]	CWT	Twitter	SM	539K	[31]
[87]	CRC, VIE	Freebuf,[46] Easyaq*	OB	610	*NM*
[111]	CEE	Microsoft, Cisco[47] and Adobe Security Bulletins Crowdstrike	TR	*NM*	*NM*
[49]	VIE	Twitter	SM	340K	[27]
[86]	CTE	Twitter	SM	47.8M	*NM*
[82]	ICE	*NM*	TR	190	*NM*
[60]	TTP	Github APTNotes	TR	520	[7]
[143]	ICE	AlienVault, FireEye, Microsoft, and Cisco security bulletins, Kaspersky, Webroot,[48] Hack Forums[49]	SM, OB, HF	15k	*NM*
[56]	ICE, TRC	Threat Expert*	TR	25K	*NM*
[108]	CRC, CEE	China GovCERT,[50] The Hacker News, SonicWall,[51] Securelist, Aus CERT,[52] CBR Online*, CISA US CERT,[53] ZDNet,[54] CNN	OB	920	[7]
[71]	CRC, TTP	Symantec	TR	17K	*NM*
[104]	CWT	Cisco, Symantec, FireEye, Palo Alto[55]	TR	875	*NM*
[72]	CRC, TTP	Symantec	TR	2.2K	*NM*
[115]	VIE	Twitter	SM	1.1B	[5]
[110]	CTE	Twitter	SM	15M	[8]
[76]	CRC, VIE	Twitter	SM	31K	*NM*

Sources marked with an asterisk (*) indicate that the url is not found.

following steps: (i) data source identification, (ii) data collection, (iii) text preprocessing and dataset labeling, (iv) learning, (v) exporting to structured formats, and (vi) application. The fifth and sixth steps do not exist in all studies, such as in References [88, 132]. However, data collection, labeling, and learning steps are common in all studies. In the following subsections, we describe these steps.

6.1 Step 1: Identifying Data Sources

The process for extracting CTI starts with identifying data sources for CTI-candidate texts. Using the methodology described in Section 4.4, we identify seven data source categories, and the agreement score is 1.00. We describe the data source categories below and report them in Table 2.

6.1.1 Social Media (n = 26). In recent years, social media has become an important source for news and updates related to literally every aspect of life and CTI is no exception. Cybersecurity

[43] https://www.virusbulletin.com/.
[44] https://sucuri.net/.
[45] https://nakedsecurity.sophos.com/.
[46] https://www.freebuf.com/.
[47] https://tools.cisco.com/security/center/publicationListing.x.
[48] https://www.webroot.com.
[49] https://hackforums.net/.
[50] https://www.govcert.ch.
[51] https://blog.sonicwall.com/en-us/.
[52] https://auscert.org.au.
[53] https://www.cisa.gov/uscert/.
[54] https://www.zdnet.com/.
[55] https://www.paloaltonetworks.com/blog/.

researchers and vendors post news and updates on threats, attacks, and zero-day exploits. Reputed security experts and organizations regularly post feeds related to CTI and security incidents on Twitter. Social media posts have been used for following purposes: (a) CTI relevance classification [46, 48, 54, 64, 76, 83, 90, 102, 113, 122, 124, 128]; (b) cyberthreat event identification [54, 55, 65, 79, 80, 110, 113]; (c) vulnerability information extraction [49, 76, 83, 97, 109, 115]; (d) IoC extraction [48, 102, 124, 143]; (e) cyberthreat alert generation [97, 113, 118, 121]; (f) cyberthreat-related entities extraction [64]; and (g) cyberthreat-related word and topic identification [85].

6.1.2 Hacker and Darknet Forums (n = 25). Online forums maintained by security vendors, experts, ethical hackers, penetration testers, and malicious users provide a platform for cyberthreat-related knowledge sharing. These forums facilitate hacker communities to share attack tactics, techniques, malicious scripts, and tools, and security best practices. Similarly, hacker forums can be hosted in the darkweb (such as Dream Market), which provides the aforementioned CTI-related information. Hence, CTI information can be extracted from these forum discussion and attachments, such as for (a) CTI relevance classification [42, 52, 61, 62, 70, 81, 83, 96, 99, 116, 135]; (b) cyberthreat-related word and topic identification [50, 59, 62, 73, 74, 94, 96, 135]; (c) hacker resource analysis [52, 63, 67, 99, 114, 119, 131]; (d) cyberthreat alert generation [63, 118, 121]; (e) cyberthreat attribution [67, 74, 119]; (f) vulnerability information extraction [83, 109]; (g) TTPs extraction [135]; and (h) IoC extraction [143].

6.1.3 Threat Reports (n = 25). Reputed cybersecurity vendors regularly prepare and publish threat reports on the current threat landscape, advanced persistent attacks, and corresponding attack groups along with their strategies and TTPs. These reports also cover vulnerabilities and exploits for specific technologies, cyberthreat events such as data breaches and zero-day attacks. Reputed cybersecurity vendors are regularly providing aforementioned analysis through textual reports. Threat reports have been used for: (a) TTPs extraction [43, 60, 71, 72, 89, 101, 103, 123, 129, 144]; (b) IoC extraction [56, 82, 88, 123, 142, 145]; (c) cyberthreat-relevant entities extraction [68, 93, 111, 134]; (d) threat report classification [43, 44, 56, 138]; (e) cyberthreat attribution [103, 106]; (f) cyberthreat-relevant word and topic identification [47, 104]; (g) CTI relevance classification [71, 72]; (h) cyberthreat-relevant entities extraction [88]; (i) cyberthreat event identification [44]; (j) vulnerability information extraction [77].

6.1.4 Online Blogs and Articles (n = 14). With the increase of cyberattacks, the number of cybersecurity-related web articles and blogs are increasing, which discuss cutting-edge attack techniques, malware descriptions, and attack prevention guidelines. They also provide information on threat prevalence and distribution. These blogs and web articles are written by individuals, communities, or organizations having expertise in cybersecurity domains. Online blogs and articles have been used for: (a) CTI relevance classification [83, 84, 87, 92, 108, 135]; (b) IoC extraction [84, 141, 143, 145], (c) vulnerability information extraction [77, 83, 87, 95]; (d) cyberthreat-relevant entities extraction [93, 108, 132]; (e) cyberthreat alert generation [121]; (f) cyberthreat event identification [127]; (g) TTPs extraction [135]; (h) cyberthreat-relevant word and topic identification [135].

6.1.5 Vulnerability Databases (n = 5). The NVD is the U.S. government's repository of known software and hardware component vulnerabilities, where the enlisted vulnerabilities are known as **common vulnerabilities and exposures (CVEs)**. The type of each vulnerability is determined by assigning a **common weakness enumeration (CWE)**[56] number. The CWE is a list of software and hardware weakness types developed by the security community. Researchers extract

[56] https://cwe.mitre.org/index.html.

CTI from the vulnerability descriptions in CVEs. Moreover, several researchers, such as Husari et al. [71], use the ATT&CK Framework and **Common Attack Pattern Enumeration and Classification (CAPEC)**[57] to develop a threat ontology. The ATT&CK Framework is a globally accessible knowledge base of real-world attacks containing tactics and techniques. Each tactic has a list of techniques, and each technique has procedure examples. CAPEC is a comprehensive dictionary of known attack patterns used by adversaries to exploit known vulnerabilities. Vulnerability databases have been utilized to extract CTI such as: (a) vulnerability information extraction [69, 77, 95, 98]; (b) CTI relevance classification [90]; (c) cyberthreat alert generation [98].

6.1.6 Software Repositories (n = 1). Software engineering practitioners use version control repositories to store software development artifacts, such as source code, bugs and issues, wiki, and documentation. Github, and GitLab[58] are examples of popular version control repositories. These repositories contain textual information on security bugs, vulnerable packages, and developers' discussions on pertinent security topics. Software repositories have been utilized to extract CTI such as (a) vulnerability information extraction [100]; and (b) cyberthreat alert generation [100].

6.1.7 Scientific Literature (n = 1). Scientific literature published in cybersecurity conferences and journals contains scientific observation and analysis on cyber threat-related issues, such as malware behaviors, cyberattack approaches, and attack mitigation techniques. These textual documents can thus be the source for extracting CTI-relevant information, such as for TTPs extraction [140].

6.2 Step 2: Data Collection

From the sources mentioned in Table 2, data can be collected in a manual or automated manner. Researchers can manually collect CTI-candidate texts. However, this method is error-prone and inefficient. Hence, researchers used online search engines to find relevant articles and threat reports [142]. Additionally, custom search engines for finding cybersecurity articles have been explored in References [81, 103]. However, using search engines to find relevant articles can be time-consuming, and hence, researchers extensively used web-based crawlers to find online cybersecurity-related topics, especially in hacker forums and darknet web resources. In References [52, 59, 74, 84, 87, 88, 92, 93, 99, 108, 131, 135, 141], researchers used HTML-based web crawlers to identify relevant texts. Moreover, TOR browser-[59] based crawlers are used in References [42, 63, 83, 116, 118, 119, 121] to find relevant topics in the darkweb. Forum web crawlers, such as Sixgill and **Open Discussion Forum Crawler (ODFC)** [81], are used in References [81, 96]. Text from the threat reports in pdf format can be extracted using tools such as PDFMiner.[60] This technique is used to scrape the text from pdf documents in References [71, 107, 140]. Finally, social networking websites and software repositories provide **application programming interfaces (API)** to collect data from their websites as well. The Twitter API[61] and Github API[62] are used for collecting tweets and software development artifacts in References [46, 48, 49, 54, 55, 64, 65, 85, 86, 97, 100, 115, 122, 124]. In References [47, 60, 68, 88, 101, 103, 123], threat reports available on one of the two GitHub repositories [7, 10] are also used.

[57] https://capec.mitre.org/.
[58] https://gitlab.com.
[59] https://www.torproject.org/.
[60] https://pypi.org/project/pdfminer/.
[61] https://developer.twitter.com/en/docs/twitter-api.
[62] https://docs.github.com/en/rest.

6.3 Step 3: Text Preprocessing and Dataset Labeling

- **Substep-3.1 Preprocessing:** Data collected from the mentioned sources in Section 6.1 are in textual format. Before extraction of CTI, these texts need to be preprocessed to discard punctuation marks, URLs, irrelevant symbols, stop words, and incorrect spellings. Then, tokenization and lemmatization are also applied. These preprocessing are needed to prepare the textual data suitable for NLP techniques.
- **Substep-3.2 Labeling:** The data needs to be labeled for building training and test cases for the ML models. For example, in the case of CTI-relevant text classification, text segments are labeled whether they are CTI-relevant or not. In the case of IoC extraction, words in the text are annotated as the type of the indicators. Correct labeling of the dataset is important, as the performance of the ML models depends on the correctness of the labeling done by humans. In the studies, researchers labeled the data themselves or deployed multiple annotators who have expertise in cybersecurity, such as graduate students and cybersecurity practitioners. In References [43, 46, 47, 54, 55, 60, 61, 68, 70, 71, 76, 77, 90, 93, 103, 106, 108, 109, 113, 123, 124, 132, 138, 141, 143] the researchers utilized manual labeling. However, the usage of pre-labeled datasets can omit the need for manual annotation such as in References [79, 80] where they used the pre-labeled dataset as the ground truth named Hackmageddon [25] and PrivacyRights[30].

6.4 Step 4: Learning Approaches

To extract CTI from textual data, researchers have utilized NLP techniques combined with ML models. We identify these techniques using the methodology described in Section 4.4. In the following list, we discuss the techniques used by the researchers for training the model for learning to extract CTI. Note that a study can use more than one technique, and the order of the techniques mentioned in this section is inconsequential. Moreover, the exact implementation may vary from study to study. The summary of the techniques used for each study can be found in Table 1 in the Supplemental Materials.

6.4.1 Supervised Classification. Machine learners are trained with examples of several classes and then applied for the prediction of the class of unseen data. Natural language-based features such as word spelling, **Term Frequency–Inverse Document Frequency (TF-IDF)**, and word embedding are computed from the text segments. Finally, these features are passed to supervised learners to classify the text segments to their corresponding labels. In the selected studies, supervised classification is used to classify (a) whether the text is CTI-relevant or vulnerability or cyberattack event-related (such as References [44, 109, 124]), (b) whether the word is an indicator of compromise, cyberthreat-related word or entity (such as References [47, 88]), (c) TTPs and hacker resource classification, i.e., source code and attachments (such as References [67, 71]). Moreover, supervised classification is also used to classify the threat reports to attack techniques (such as Reference [43]), set of TTPs to their threat actors (such as Reference [124]). Multi-label classification (i.e., each item in training data has more than one label) is also applied in Reference [89] through binary relevance and label propagation techniques. From Table 1 in Supplemental Materials, we observe that SVM is the most used classifier, followed by neural networks (convolutional, recurrent, graph convolutional), logistic regression, random forest, and k-nearest neighbors.

6.4.2 Semi-supervised Classification. Semi-supervised classification refers to learning from a training dataset that has a small amount of labeled data and a large amount of unlabeled data. In Reference [99], label propagation and co-training methods were applied to construct a labeled dataset from a tiny sample of a labeled example. In References [113] and [93], distant supervised

learning and uncertainty sampling were applied to construct a labeled dataset from a tiny piece of a labeled sample. In Reference [110], expectation regularization was used to build a labeled dataset from a small selection of a labeled example.

6.4.3 Reinforcement Learning. Reinforcement learning refers to learning through trial-and-error from user-generated feedback. In Reference [134], reinforcement learning is used to extract the semantic relationships among the cyberthreat entities.

6.4.4 Clustering. Text segments are clustered based on their similarity through unsupervised learners such as K-means. Similarity computing algorithms such as cosine similarity and Jaro Winkler distance algorithms were used as the distance function for computing the clusters. Clustering is utilized to combine similar text segments, such as grouping identical threat reports (such as Reference [56]) or social media posts (such as Reference [85]). From Table 1 in Supplemental Materials, we observe that k-means is the most used clustering technique, followed by hierarchical, DBSCAN, and affinity propagation.

6.4.5 Multitask Learning. Multi-task learning refers to solving more than one learning task simultaneously while utilizing the overlaps and differences in the tasks. In Reference [79], multi-task learning is applied for cyberthreat event detection. In this approach, three different sets of features are used by three different learners. These feature sets are organization-specific features (i.e., organization-specific keywords), threat-specific features (i.e., data breach), and generic features (i.e., a common set of words used in both cybersecurity and non-cybersecurity domains). Then multi-task learners such as LASSO is used to learn an optimal model for identifying cyberthreat events.

6.4.6 Regular Expression. Cyberthreat indicators, such as malware hash, IP addresses, and software version names, contain specific spelling patterns that can be captured with the use of regular expressions. Based on these spelling patterns (such as OpenIoC[63] patterns), indicators can be extracted. In References [84, 88, 102, 141–143], the authors utilized this technique.

6.4.7 Topic Modeling. Topic modeling technique is applied to discover the abstract topics in the text. This technique is taken for generating topic words (such as data breach, denial of service attack) for finding the trending cybersecurity topics in Twitter (such as Reference [65]), hacker forum posts (such as Reference [116]). Topic models, such as **Latent Dirichlet Allocation (LDA)** and **Latent Semantic Analysis (LSA)**, are applied to the text segments to generate topics from the text segments.

6.4.8 Named Entity Recognition. **Named entity recognition (NER)** refers to the classification task of entity types mentioned in texts. NER can be applied to identify named entities in the cybersecurity domain, such as malware names, IP addresses, malware hashes, and so on. NER technique is applied in identifying indicators (such as Reference [84]), identifying vulnerability (such as Reference [97]), and STIX vocabulary information (such as Reference [134]) from CTI-candidate texts.

6.4.9 Parts of Speech Tagging. **Part of speech (POS)** tagging refers to the mapping of words in a corpus to the corresponding POS, such as nouns and verbs. This technique can be used to extract cybersecurity-related noun keywords from CTI-relevant text such as in References [92, 96, 132].

6.4.10 Dependency Parsing. Dependency parsing refers to the parsing of grammatical structure of a sentence. The pattern of dependencies among the parts of speech can be learned, such as subject-verb-object combination. The dependency parsing technique is used to extract the malicious actions (verb) and victims (object) to identify the attack patterns from the CTI-candidate

[63] https://www.mandiant.com/resources/openioc-basics.

text (such as Reference [71]) and to identify the semantic relationships among the named entities (such as Reference [111]).

6.4.11 Ontology. The relationships among the cybersecurity-related named entities are computed using an ontology such as UCO, STIX. In Reference [97], the ontology is used to apply reasoning over event description to generate alerts. Ontology is also used in References [77, 95, 100], where the contextual information of the entities is retrieved from external ontologies such as DbPedia,[64] MITRE ATT&CK taxonomy, and NVD database.

6.4.12 Feature Selection and Ranking. Textual and contextual features from CTI-candidate texts are filtered and ranked using mutual information [124] and information gain [72].

6.4.13 Graph Mining. Graph mining is a special case of structured data mining where information is extracted from structured or semi-structured data. The dependency relation (grammar, POS) of words can be represented as a graph, and then the constructed graph can be mined to identify the relations of IoC and context words [84] or to identify social media posts of similar topics [80].

6.4.14 Bias Correction. Bias correction is applied to mitigate the class imbalance issue in a dataset where examples for all the target classes are unevenly distributed. Bias correction techniques such as KMM, KLIEP, aruLSIF are used in Reference [43].

6.4.15 Text Similarity Computation. The selected studies mention the use of similarity score calculation for finding similar texts (such as Reference [89]) or keywords (such as Reference [80]) and for clustering texts (such as Reference [113]). From Table 1 in Supplemental Materials, cosine similarity is the most used metric, followed by BM25, and Wordnet.

6.4.16 Textual and Contextual Features. CTI-candidate texts are represented as different textual features from which machine learners can perform classification and clustering. From Table 1 in Supplemental Materials, we observe that TFIDF, word2vec, and n-gram are the most common feature used across different studies. POS, bag of words, doc2vec, BERT, GloVe, graph embeddings are also used as textual features. Several studies also used different contextual features for learners, such as co-occurring words, word density, word verbosity, word count, casing, and presence of special characters.

6.4.17 Subject-verb-object Tuple Extraction. Any English sentence contains subject, verb, and object that are extracted as **subject-verb-object (SVO)** tuple. SVO tuple is extracted to represent the adversary (subject), malicious action (verb), and target (object). TTPs extraction studies such as References [71, 72] used this technique.

6.4.18 Miscellaneous. Reference [80] used a dynamically typed query expansion technique to find similar cyberattack event posts from social media posts. Reference [123] used co-reference resolution, passive to active voice conversion, text summarization, and word homologation techniques to convert complex CTI-candidate sentences to simple sentence(s). Reference [96] uses sentiment analysis to find threatening forum posts to users. References [100, 111] use rule and pattern finders to learn the grammatical structure of CTI-candidate texts. Predefined keywords are used to find similar posts from CTI relevance in Reference [124]. TextRank and PageRank techniques are applied to weight and rank the identified cybersecurity keywords in References [55, 74].

[64] https://wiki.dbpedia.org/.

Table 3. Structured CTI Export Format Used in Studies

Format	Studies used the format
Knowledge graph	[77, 95, 97, 100, 107]
STIX	[68, 71, 111, 135]
MISP	[46, 48, 83]
OpenIoC	[84, 143]
Domain agnostic	[56, 87]

6.5 Step 5: Exporting to Structured Format

After extraction, CTI could be presented and shared in a structured format. We identify these sharing formats using the methodology described in Section 4.4, and the agreement score is 1.00. We provide details on the sharing formats used in studies in Table 3 and the following subsections.

6.5.1 Cybersecurity Knowledge Graph (n = 5). A knowledge graph is a set of entity pairs and their relationship. **Cybersecurity Knowledge Graph (CKG)** represents CTI as a knowledge base [107]. To present CTI in the knowledge graph, researchers use **resource description framework (RDF)** and **unified cybersecurity ontology (UCO)**. RDF[65] is a standard format linked to the data representation for data interchange on the Web. **Unified cybersecurity ontology (UCO)** [117] is an ontology-based on STIX. Piplai et al. [107] described a system to extract information from the security report and represent that in a CKG. Neil et al. [100] extracted information on vulnerable packages and dependencies from open source projects and libraries from code repository issues and bug reports. Their extracted CTI is represented in RDF format as a security knowledge graph. Mittal et al. [97] discover CTI from Twitter and represent the gathered CTI using RDF format. Joshi et al. [77] and Mulwad et al. [95] also used RDF and created knowledge bases of extracted CTIs.

6.5.2 STIX (n = 4). STIX[66] is one of the most commonly used structured language and serialization formats to share CTI in enterprise organizations [155]. STIX information is human and machine-readable in JSON and contains domain and relation objects. We find four studies exporting extracted attack tactics, techniques, and procedures into STIX format [68, 71, 111, 135].

6.5.3 MISP (n = 3). MISP[67] is an open-source CTI platform for gathering, storing, and sharing CTI. Using MISP, CTI can be stored in a structured format and exported in formats such as STIX, OpenIoC, XML, or CSV. MISP also provides a CTI sharing format based on JSON. Alves et al. [46, 48] used the MISP format to generate IoCs extracted from the Twitter posts. In Reference [83], the authors implemented a MISP-compatible module to share the extracted vulnerability-related information.

6.5.4 OpenIoC (n = 2). OpenIoC[68] is a standardized open-source framework for sharing CTI. OpenIoC format is based on XML and is machine-readable. By using the OpenIoC framework, organizations can access the latest IoCs shared by other organizations and can communicate with each other [35]. We observe two studies, References [84, 143], exported the extracted IoCs to OpenIoC format.

[65] https://www.w3.org/RDF/.
[66] https://oasis-open.github.io/cti-documentation/stix/intro.html.
[67] https://www.misp-project.org/.
[68] https://github.com/fireeye/OpenIOC_1.1.

6.5.5 Domain-agnostic Structured Format (n = 2). Not all researchers used CTI sharing standard formats to represent and share the extracted CTI. For example, Bo et al. [56] proposed a threat operating model that captures the information of cyberthreats gathered from publicly available threat reports and presented them in an XML format. Li et al. [87] defined a CTI template to represent CTI in security articles. Their CTI template has two parts, including CTI-related entities and summarizations of the article. The entities are CVEs, victimized devices, device manufacturers, and impacted locations.

6.6 Step 6: Applications

In the studies, the authors extracted CTI from the CTI-candidate text and then demonstrated how they utilized the CTI in application scenarios. In this section, we discuss these applications.

6.6.1 Threat Landscape (n = 6). The extracted CTI from the text can give security experts insights into the threat landscape. For example, Liao et al. [84] observed the largest number of extracted IoCs are with the type "PortItem/remoteIP," which shows the popularity of download-driven phishing and other web-based attacks in the landscape. Macdonald et al. [96] identified potential threats to critical infrastructures. For example, they found a strong relationship between "DDoS" and "bank" that shows the popularity of financial institutions being targets for DDoS attacks. Husari et al. [72] developed *ActionMiner* to extract threat actions and showed that their results could help to understand the threat landscape. For example, they found "process," "DLL," "code," "library," and "Chrome," which are the five most related objects to the action "inject." This information can also help security experts to plan for mitigating injection attacks.

Zhao et al. [143] extracted CTI with domain tags, such as **industrial control system (ICS)** and **internet-of-things (IoT)**. After clustering CTIs based on their domains and analysis of the clusters, they identified insights on different attack types in various domains. For example, they found that the implementation of DDoS attacks varies across multiple domains, and the complexity of the phishing attack depends on the value of the target domain. They also found that IoT-related threats have developed rapidly because of the growing number of IoT devices in recent years. Based on their proposed metrics to quantitatively measure the threat severity from the perspective of security-related social opinion, ICS and governments have experienced threats with higher severity impact than those of other domains.

Samtani et al. [42] focused on two types of **denial of service (DoS)** and web application threats that target PHP technology at intervals of three months. Their results showed that the DoS threat landscape is growing more rapidly than the web applications. They found that although new threat types are emerging in both threat categories, the core functions of the threats remain the same over time. This information can guide cybersecurity experts in prioritizing activities to mitigate threats. Nagai et al. [104] collected IoCs and showed that their approach could help security experts understand attack methods and threat trends in the IoT industry and financial institutions. For example, their results showed that in the IoT industry, the attack methods are being focused on firmware. The authors found a relationship between Mirai malware and IoT devices such as routers and printers in 2017, which confirms the attack by this malware on IoT devices worldwide.

Extracting CTI from a corpus of articles may also show connections between threats that were never known before. For instance, Liao et al. [84] clustered the articles in their dataset into 527 clusters based on having at least one IP, email, or domain in common. After analyzing the clusters, they found that the authors of these articles did not realize that the attacks they were documenting were related to other attacks. They also observed that the IoCs reported by many articles disappear quickly. For example, 92% of IoCs are mentioned on average with 68 articles per month during a 0- to 1-month time window before they are stopped.

6.6.2 Querying CTI (n = 4). The extracted CTI from the unstructured text can be queried to find information if presented in a query-able platform. For instance, Samtani et al. [116, 119] collected malicious attachments and source code from hacker communities and enabled searching, sorting, and browsing of those data through a portal. They also provide a dashboard to show the hacker resource trend over time, key hackers that use those resources, and a list of sources. The information in the dashboard can be filtered by time, resources, and by a hacker. Neil et al. [100] extracted CTI from open source projects and libraries. They presented software dependencies in a security knowledge graph. Before using a project or a library, developers can query the security knowledge graph and find known vulnerabilities. Piplai et al. [107] built a **cybersecurity knowledge graph (CKG)** for malware that allows querying the entities of the CKG. For instance, one possible question is what "Tool" a particular malware uses.

6.6.3 Dataset Generation (n = 4). Extracted CTI from the text can contribute to a dataset for use by security researchers and practitioners. In Reference [127], the authors developed a new dataset for cyberthreat event identification that is manually annotated and compatible with word embedding-based deep learners. Moreover, in References [77, 95, 107], the authors stored their extracted CTI in a graph-based data structure that security researchers and other CTI platforms can consume.

6.6.4 Establishing Correlation with Attack Groups and Key Hackers (n = 3). CTI-candidate texts often contain information on cybersecurity incidents and associated cyberthreat actors, such as their roles, strategies, and procedures. This information can be used to map the attacks to the responsible cyberthreat actors. For example, the authors in Reference [103] selected 36 cyberthreat actors and an average of 9 CTI documents for each. After extracting TTPs from the documents, they mapped the attack patterns to the responsible cyberthreat actors. Then for an unseen CTI-candidate text, their system can predict the cyberthreat actors. Moreover, in References [67, 119], the authors identified the social networks of cyberthreat actors from CTI-candidate texts. Each publication listed identified key cyberthreat actors such as "KriPpLer" and "mjrod5" in Reference [119] and "LinX64" and "AsAs" in Reference [67].

6.6.5 Malware Protection Improvement (n = 3). Security protection organizations can use the analysis from the extracted CTI to respond quickly. For example, Liao et al. [84] estimated the time intervals between the first appearance of the IoCs and their adoption by anti-malware tools and web scanners. They observed that 47% of the IoCs were detected by anti-virus scanners or IP/URL scanners before the technical blogs reported them. For the remaining IoCs, the duration between the first IoC being released and uploaded for a scan is between 0–2 days to more than 12 days. For IPs and domains, the whole process often took more than 12 days. However, malware hashes were often quickly added to anti-virus scanners for scanning, in most cases within two days.

Zhu et al. [141] showed that detection systems focus more on blocking network intrusion and removing malicious programs. They are not capable of detecting attacks that use social engineering to download payloads. Hence, a campaign attack can continue for more than one year, even after its discovery. This information helps anti-virus vendors to know the weakness of their tools. Williams et al. [131] collect exploit information from hacker forums. They visualized data based on posted exploits and author activity. Data collection was done incrementally, so the information about recent exploits can be helpful for security experts to find new threats. Considering author activity is also valuable to find the most active hacker communities and the exploits they share.

6.6.6 Increase Awareness of Cyberattacks (n = 3). Extracting CTI from online resources can help security experts to be aware of possible future attacks and to predict and prevent them more

effectively. DISCOVER [121] is an early cyberthreat warning system that uses Twitter, cybersecurity blogs, and dark web forums to generate warnings based on novel terms in these data sources that co-occur with context terms. For example, the NotPetya malware attack went public on June 27, 2017, but DISCOVER generated the first warning related to the malware in February 2017 based on security blogs and in March 2017 based on Twitter data. Neil et al. [100] extracted CTI from open source projects and libraries and presented software dependencies in a security knowledge graph. An alert generation system can use this security knowledge graph and generate alerts if a developer can link a library to known vulnerabilities or if a client installs a vulnerable application. Dionísio et al. [64] showed that their model could identify security-relevant tweets with labeled NER from tweets published from 1 to 148 days before the NVD disclosure of a vulnerability. The CVSS[69] severity of tweets ranges from 4.9 to 9.9 show the importance of alert generation of the model.

6.6.7 Discovery of Zero-day Exploits (n = 1). One of the applications of extracting CTI is to discover zero-day exploits. Detection of these exploits at an earlier stage can help organizations protect their system or minimize the damage caused by the attack [99]. Nunes et al. [99] detected 16 zero-day exploits from darknet marketplace data in a four-week period.

6.6.8 Cross-site Connection between Multiple CTI Sources (n = 1). The extracted data can reveal cross-site connections between sources if CTI extraction is done on more than one source. For example, Nunes et al. [99] used darknet marketplaces and hacker forums to collect CTI. They created a connected graph using the "usernames" used in two domains. They found individuals selling products related to malicious hacking in marketplaces and hacking forums simultaneously. This information is helpful to determine the social groups of the domain.

6.6.9 CTI Relationship Analysis (n = 1). From the extracted CTI, relationship analysis can be performed among them such as the association between threats, techniques, tools, and mitigation. For example, Piplai et al. [107] built a cybersecurity knowledge graph for malware where they included details such as malware's campaign, used tools, and targeted software in the knowledge graph. The graph can be used to compare malware and cluster similar malware.

6.6.10 APT Technique Analysis (n = 1). Analyzing **advanced persistent threats (APT)**-related technique trends can result in valuable insights, such as the trends of the attack techniques used. For example, Niakanlahiji et al. [101], analyze the trend for the 14 most-mentioned techniques in their data. They found that from 2013 to 2016, exploiting browsers and malicious scripts were the most used techniques. In 2017, using PowerShell became one of the top techniques; using malicious scripts remained one of the top techniques; however, exploiting browsers became insignificant. This type of information can inform security experts to prioritize mitigation strategies. In the same study, researchers analyzed the relationship between co-occurring techniques and calculated the strength of these relationships. For instance, they observe a strong relationship between obfuscation and using scripts in their APT reports and show that APTs commonly use obfuscation techniques.

6.6.11 Malware Profiling (n = 1). Malware can be profiled based on CTI-candidate texts. For example, Bo et al. [56] extracted the malware information from multiple threat reports and constructed the malware profile in an XML format. The malware profile contains static attributes, such as operating system, platform, and CVE; and dynamic attributes, such as hotness (the degree to which the malware is attacking lately) and hack interest (the degree to which the malware is

[69] https://nvd.nist.gov/vuln-metrics/cvss.

frequently attacking). For example, in the profile of the malware W32.Kwbot.Worm operating system is "Windows CE;", CVE is "CVE-2012-0158", hotness is 149, and hack interest is 9,913. This information can help organizations to check their IT environment and generate early warnings.

7 CTI EXTRACTION PIPELINE INSTANTIATION

We propose a NLP and ML-based pipeline in Section 6 and Figure 3 for CTI extraction purposes. We report the identified learning approaches in Section 6.4 used in the studies. We find that the instantiated pipeline for each of the CTI extraction purposes has common steps such as data source identification, data collection, preprocessing, and exporting to a structured format. However, the learning steps vary across the extraction purposes. We discuss how the pipeline can be instantiated for each of the CTI extraction purposes below.

- **CTI relevance classification:** The CTI-candidate text can first be represented as textual features such as word vectors, word embedding, and n-grams. Moreover, contextual features such as cybersecurity-specific keyword count and word spelling can also be considered features. For many features, feature selection techniques such as mutual information and information gain can be applied. Finally, the supervised learners can be trained with these features to classify the texts to their corresponding labels.
- **Indicators of compromise identification:** The CTI-candidate text first should be filtered to identify the IoC containing sentences (CTI relevance classification). Then regular expression and named entity recognition technique can be applied to identify the words representing indicators of compromises.
- **Attack tactics, techniques, and procedures extraction:** The CTI-candidate text should be filtered to identify the attack patterns and techniques describing sentences (CTI relevance classification). Then, POS tagging and dependency parsing can be applied to extract the SVO tuples from the sentences. Next, the similarity score of these tuples can be computed to a relevant framework such as MITRE ATT&CK technique descriptions or Cyber-kill-chain descriptions. Another approach would be to apply classification directly to the filtered sentences to learn their corresponding labels, such as technique name and kill chain phases.
- **Vulnerability information extraction:** The CTI-candidate text first should be filtered to identify the vulnerability describing sentences (CTI relevance classification). Then regular expression and named entity recognition can be applied to determine the words representing vulnerability information such as CVE identifier, CVSS score, application name, version, and so on.
- **Cyberthreat-relevant word and topic identification:** The CTI-candidate text should be filtered to identify the cybersecurity-relevant sentences (CTI relevance classification). Then parts of speech tagging can be applied to find the noun and verbs that could be cyberthreat-relevant. Moreover, topic modeling techniques such as LDA can also be applied to identify cyberthreat-relevant topics.
- **Cyberthreat event identification:** The CTI-candidate text first should be filtered to identify the cyberattack describing sentences (CTI relevance classification). Then the text can be represented to features and fed to machine learners to classify the event categories. Moreover, clustering can also be applied to aggregate the texts describing similar types of events. Researchers can also find events of specific types (such as data breaches) with topic modeling, keyword weighting, and so on.
- **Cyberthreat-relevant entities extraction:** The CTI-candidate text first should be filtered to identify the cybersecurity-relevant sentences (CTI relevance classification). Then regular expression and named entity recognition technique can be applied to identify the words

representing various concepts regarding cyberthreats such as malware names, IP address, attack means, organization, and so on.

- **Hacker resource analysis:** The CTI-candidate text first should be filtered to identify the cybersecurity-relevant sentences and then classify the sentences to corresponding labels such as attachment types, script language, hacker comments, and so on. A graph data structure can be instantiated to represent the relation between the hackers and their resources to identify key hackers, hacker influence, and so on.
- **Cyberthreat alert generation:** The alert generation process is preceded by the identification of cyberthreat-relevant information such as indicators of compromise, vulnerability, or events. Hence, after extracting information on indicators, vulnerability, or cyberattack events, alerts can be generated by first aggregating the information, clustering the information to filter and duplicate, and then issuing the alert on user-defined rules and threshold.
- **Cyberthreat attribution:** Attribution to attackers can be represented as a classification problem where CTI-candidate text can be represented as features and attackers are used as labels. Another approach is to (a) obtain the set of indicators, TTPs, and named entities, (b) represent them as features, and (c) train the machine learners for supervised classification.
- **Threat report categorization:** First, the texts inside the threat reports can be represented as features or transformed to a set of TTPs. Then, the features can be fed towards classifiers or clustering algorithms to group the reports.

8 EVALUATION AND DISCUSSION

We evaluate the selected studies based on the following criteria: (a) data availability, (b) code availability, (c) whether the study compared their proposed approach to other approaches from the literature, and (d) whether the extracted CTI is exported for sharing. The first two authors individually took notes of the mentioned criteria for each study and then resolved the disagreements. The agreement score is 1.00. The evaluated criteria for all the 84 selected studies is reported in Table 2 in Supplemental Materials. We list our observations below based on our evaluation and CTI extraction pipeline.

- **Availability of replication package:** We observe that 49% of the studies made their dataset available. For the rest of the papers, only sources are mentioned for collecting textual data. However, the collected and labeled dataset are not shared for future researchers. We observe that the source code of the studies is made available in only 13% of the studies, implying the majority of the studies are not replicable. We also observe only 10 studies are built upon prior studies (Table 3 in Supplemental Materials), which also reflects the lack of replication package.
- **Comparison to other CTI extraction approaches:** We observe 17% of the studies compared their proposed approaches to other proposed CTI extraction approaches in the literature. Rerunning the studies on the latest textual dataset and choosing a better-performing CTI extraction pipeline can be difficult for future researchers, given the lack of replication package and comparison with other proposed approaches.
- **Low classification performance:** We observe that the classification performance of 43% of the studies demonstrates less than 0.80 scores in case of accuracy, precision, or recall. Moreover, the classification performance of the proposed model is not reported in 11% of the studies. We also observe that 11 studies evaluated their proposed model with a dataset with a relatively small size (i.e., the count of instances in the dataset is less than 500). We observe that roughly half of the proposed CTI extraction studies might be susceptible to a high false-positive rate. The observation emphasizes the need for further human verification of the actionability criteria on the extracted CTI.

- **Extracted CTI should be exported to enable CTI sharing:** We observe that only 19% of the studies exported their extracted CTI to structured format. We observe four CTI extraction purpose-related papers exported CTI to structured format: IoC extraction, TTPs extraction, cyberthreat-relevant entities extraction, and vulnerability information extraction. However, we observe several factors that contribute to the generic lack of exported CTI to structured format. First, there is a lack of well-defined structured format for sharing CTI for several CTI extraction purposes: such as studies performing classification tasks on whether the text or word is cyberthreat-related or not. Second, information-related to cyberthreats and attacks can generally be considered as CTI; however, as mentioned in Section 2.1, lack of a clear definition of CTI leads to a lack of consensus on what information can be considered as CTI, what kind of information should be extracted, and how to represent the extracted information that is useful for the practitioners. Finally, the evolution of the nature and concepts of the threat landscape and cyberattack trends provides a further challenge to the industry to construct well-defined ontology and semantic relation between concepts related to cyberattacks and cybersecurity-specific terms. Cybersecurity research organizations such as MITRE and OWASP have made progress on conceptualizing the attack patterns and threat models of cyberattacks such as MITRE ATT&CK, OWASP Threat Ontology; however, organizations also need to design and adapt these structured representations to accommodate the extracted CTI in universally shareable formats.
- **Actionability of extracted CTI should be ensured:** As mentioned in Section 2.1, the actionability of the extracted CTI depends on the following criteria: accuracy, relevance, ingestibility, trustworthiness, timeliness, and so on. We observe that, in many of the selected studies, the actionability aspects of the extracted CTI are unexplored, as 40% of the studies did not reflect on any practical observation from the extracted CTI. Moreover, the validation of how practically useful the extracted CTI is is not investigated in any of the studies and should be investigated thoroughly in future research in the domain.
- **Class imbalance issue:** A challenge researchers face is having imbalanced classes because of the nature of the collected data. For instance, from 1 million news articles, just 500 articles may be CTI-related. Imbalance classes can affect the quality of the training of the model and the accuracy of testing the model. We observe several studies taking steps to mitigate the negative effect of class imbalance issues, such as in References [43, 54, 66, 90, 109], however, the issue is in general overlooked in the majority of the studies.
- **Lack of clean, labeled, and published dataset:** Collecting and cleaning data are the first steps done by researchers and can be time-consuming and complex. If cleaned and labeled data is available, then other researchers can use the dataset without further effort. In addition, labeled data is needed in supervised tasks, such as classification. Even if the raw data is available, researchers have to perform labeling, and the process takes significant time and effort. However, in the selected studies, we observe authors construct their own dataset due to the lack of labeled and published dataset. One possible reason could be the changing nature of attack patterns. Collected data in a period may not be valuable in the future for CTI extraction. However, creating and publishing datasets, especially labeled datasets, should be considered to ease the potential of extending the work by other cybersecurity researchers in the future for training data and to enable comparison between analysis techniques. We advocate future researchers and practitioners to put focus on (a) constructing a shared repository and framework for gathering, analyzing, and publishing label datasets; (b) launching coordinated projects to share, analyze, and validate the published dataset; (c) continuously increment the dataset and labels to adapt with the evolution of threat landscape.

9 LIMITATIONS

The search process of finding the relevant publications may not be comprehensive, as we use six scholarly databases as sources. However, other scholarly databases might contain more studies. Moreover, the search terms we used may also not return all the relevant CTI extraction papers. We choose these search terms to be generic (such as "Threat Intelligence" or "Hacker forums") to find out all the relevant papers that extract CTI from textual sources. However, we did not use specific CTI search terms such as "indicator of compromise" or "vulnerabilities," which could have returned relevant studies from the search. We observe that almost half of the papers came from snowballing. From our search string, as we used generic search terms, our search results in many irrelevant papers, and the titles of the papers contained the search terms. This is why we found 33 papers from the search after filtering and 51 papers from snowballing, which were missed in the search. Thus, the chosen search terms may have limitations in returning the relevant studies. Hence, we applied snowballing and computed the QGS score to measure the effectiveness of the search strategy in this article. The process of searching, applying filtering criteria, and coding are subjective. Hence, the first two authors of the article individually applied open coding techniques and then resolved the disagreements. We evaluated the studies based on four criteria. However, specific evaluation based on each CTI extraction purpose is not performed—such as the classification performance of various types of IoCs extracted from the text. We also did not evaluate what specific design choice of CTI extraction pipeline could yield better results. Finally, as stated in Section 2.1, CTI is yet to have a universally agreed-upon definition. We consider any cybersecurity or cyberthreat-relevant information as CTI if the extracted information could be useful for preventing, detecting, and mitigating cyberthreats. Consequently, we attribute cybersecurity-relevant sentence classification studies to CTI-relevant papers as well (based on our CTI extraction purpose derivation process at Section 4.4); however, these studies do not extract any specific TTPs or indicators apart from only performing the classification tasks. Hence, the CTI extraction purpose categories and findings in our article may be limited from the aspect of generalizability.

10 FUTURE RESEARCH DIRECTION

In this section, we discuss potential future research directions.

(a) **Extraction of higher levels of threat intelligence based upon the existing extraction models:** Existing CTI extraction models can act as building blocks for extracting higher levels of CTI. For example, existing extraction models for cybersecurity entities, IoCs, and TTPs can aid researchers and practitioners in extracting information on what attack techniques are more prevalent for exploiting certain types of vulnerabilities, what set of identified IoCs can lead to the identification of certain attack techniques, and so forth.

(b) **Introduction of standard taxonomies and templates:** We observe automatic extraction of cybersecurity concepts first requires manual annotation, which incurs significant human effort. The annotation effort can be reduced if various concepts in cybersecurity have standard taxonomies so researchers and practitioners can extract the concepts through parsing with defined rules. Such taxonomies exist for vulnerabilities (such as CVE identifier) and IoCs (such as OpenIoC standard). However, standard taxonomies should be introduced for other concepts such as cybersecurity events, damage caused by a cyberattack, means of cyberattack, and tools used in cyberattacks. Moreover, CTI-candidate texts are written in plain language and do not have any specific formats, which also increases the manual annotation effort. Hence, a semi-structured format for cybersecurity-related articles may also help parse the documents automatically without significant human intervention.

(c) **Threat mitigation and hunting:** CTI-candidate texts provide details on both how attacks are being performed and how the cyberattack attempts can be mitigated through prevention

and detection measures. Thus, NLP and ML models can be constructed to automatically suggest security requirements and mitigation steps for a given course of actions by the attackers, which could provide organizations with a starting point to protect them from past cyberattack incidents. In addition, extracted attack techniques from CTI-candidate texts may aid cybersecurity red and blue teams in threat hunting through identifying correlation among techniques, sequence of adversarial actions, and patterns of occurrence in adversarial techniques. Extracted CTI over a long period of time can provide a great source of information for performing longitudinal analysis over cyberthreat landscape. The information can also be utilized to perform probabilistic analysis on various cybersecurity aspects, such as what attack technique is most likely given the architecture of a system. The probabilistic analysis also aids practitioners in designing defensive strategies based on the predicted cyberthreats.

(d) **CTI extraction to improved cybersecurity practices:** The studies we investigated put their focus primarily on extracting CTI-relevant information such as attack patterns, IoCs, attack events, and responsible actors. Based upon this extracted information, cybersecurity practitioners and researchers could deploy the CTI extraction pipeline, aggregate the obtained information to better understand how to mitigate risks, optimize security practice, and establish correlation among the obtained CTI.

(e) **Mitigating duplicated efforts:** Section 6.1 demonstrates that practitioners and researchers can obtain CTI from a plethora of data sources. Moreover, multiple sources can discuss the same cyberthreat topic (i.e., a given cyberattack is analyzed in threat reports from different vendors). Hence, practitioners and researchers need to focus on filtering the duplicate information from obtained CTI from multiple sources, grouping the obtained CTI based on CTI subdomain. With the evolution of threat landscape, collected CTI can be aggregated to find patterns in cyberthreat landscape. Moreover, with the passage of time, extracted CTI may become obsolete or irrelevant in the future. Hence, the focus should be given on how researchers can utilize the insight gained from already-extracted CTI to establish correlations between attack and defense strategies.

(f) **Extracting CTI from large, multiple datasets:** In an academic environment, researchers usually work on a single dataset and report the observation [153], which results in duplicated efforts from researchers and weak correlation among the extracted CTI. Hence, the focus should be given on how cybersecurity researchers can aggregate the information gained on extracted CTI from large and multiple datasets, find the relationship among this information, and instantiate actionable knowledge for relevant organizations from this collection of information. Moreover, extracted CTI should be checked for quality issues such as false alarms and consistency.

(g) **Prioritization and automated decision making:** CTI-candidate texts provide raw data for extracting potential CTI. However, the focus should be given on how cybersecurity practitioners can prioritize proactive actions to defend from attacks. Moreover, researchers can explore how extracted CTI can be followed by automated decision-making.

(h) **Cybersecurity language model:** Cybersecurity-specific NER model and word embeddings (i.e., sec2vec,[70] Harvard NER corpus[71]) can also be used instead of stock pre-trained models (such as word2vec [149], glove[72]) to gain better performance in machine learning models.

(i) **Incremental learning:** We observe the authors prepare or procure static datasets and propose machine learning models to extract CTI. However, practical usage requires

[70] https://github.com/0xyd/sec2vec.
[71] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/1TCFII.
[72] https://nlp.stanford.edu/projects/glove/.

online/incremental machine learning models that can continuously extract CTI from real-time streams of CTI-candidate texts. Hence, we advocate for CTI extraction models that can learn to extract CTI from a continuous feed of data with minimal human interventions.

11 CONCLUSION

CTI can be extracted from unstructured texts on cyberthreat-related topics by cybersecurity researchers, organizations, and vendors. In this survey, we identify 84 relevant studies from six scholarly databases. We categorize the CTI extraction purposes, propose a CTI extraction pipeline, and identify the data sources, techniques, and CTI sharing formats utilized in the context of the proposed pipeline. Our work finds 11 types of extraction purposes where CTI text classification, IoCs, and TTPs extractions have been given greater attention. We also identify seven types of textual sources for CTI extraction where hacker forums, threat reports, social media posts, and online news articles are primarily utilized. We also observe that natural language processing and machine learning-based techniques, such as supervised classification, named entity recognition, topic modeling, and dependency parsing, are the primary techniques used for CTI extraction. Finally, we conclude with a set of technical challenges observed in the studies and future research directions in the CTI extraction domain. Prospective cybersecurity researchers can benefit from our work, such as instantiating their own CTI extraction pipeline based on their extraction purposes, identifying relevant data sources, and selecting appropriate techniques for mining CTI. Overall, the work provides researchers with options for making design decisions for their own CTI extraction method from natural language artifacts.

ACKNOWLEDGMENT

Any findings and opinions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] AZSecure Portal. Retrieved from www.azsecure-data.org.
[2] Cambridge crime dataset. Retrieved from www.cambridgecybercrime.uk/.
[3] Chainsmith. Retrieved from https://ioc-chainsmith,org.
[4] Cybersixgill. Retrieved from https://www.cybersixgill.com/.
[5] Exploit Database. Retrieved from https://www.exploit-db.com/.
[6] Featuresmith. Retrieved from http://featuresmith.org.
[7] Github aptnotes. Retrieved from https://github.com/aptnotes/data.
[8] Github aritter. Retrieved from https://github.com/aritter/twitter_nlp.
[9] Github behzadanksu. Retrieved from https://github.com/behzadanksu/cybertweets.
[10] Github Cybermonitor. Retrieved from https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections.
[11] Github das-lab. Retrieved from https://github.com/das-lab/Cyberthreat-Detection.
[12] Github DissectMalware. Retrieved from https://github.com/DissectMalware/IoCMiner.
[13] Github eyalmazuz. Retrieved from https://github.com/eyalmazuz/AttackAttributionDataset.
[14] Github HongyiZhu. Retrieved from https://github.com/HongyiZhu.
[15] Github kbandla. Retrieved from https://github.com/kbandla/APTnotes.
[16] Github ksatvat. Retrieved from https://github.com/ksatvat/Extractor.
[17] Github luoluoluoyl. Retrieved from https://github.com/luoluoluoyl/relation_extract_dataset.
[18] Github ndionysus. Retrieved from https://github.com/ndionysus/twitter-cyberthreat-detection.
[19] Github nicholasprayogo. Retrieved from https://github.com/nicholasprayogo/CyberATE.
[20] Github PEASEC. Retrieved from https://github.com/PEASEC/CySecAlert.
[21] Github Samsung. Retrieved from https://github.com/Samsung/Twiti.
[22] Github scu-igroup. Retrieved from https://github.com/scu-igroup/Attack-Technique-Dataset.
[23] Github stucco. Retrieved from https://github.com/stucco/auto-labeled-corpus.
[24] Github yimingwu510. Retrieved from https://github.com/yimingwu510/TAG.

[25] Hackmageddon. Retrieved from https://www.hackmageddon.com/.
[26] Indicator of compromise - CSRC - NIST Glossary. Retrieved from https://csrc.nist.gov/glossary/term/indicator_of_compromise. [accessed 15-June-2022].
[27] IOT vulnerability data. Retrieved from https://www.kaggle.com/salevizo/tweets-related-unrelated-to-iot-vulnerabilities.
[28] Leak forum. Retrieved from http://leakforums.net/thread-719337.
[29] NVD full listing. Retrieved from https://nvd.nist.gov/vuln/full-listing.
[30] Privacy rights clearinghouse. Retrieved from https://privacyrights.org/data-breaches.
[31] Secbuzzer. Retrieved from http://secbuzzer.iii.org.tw/.
[32] Stackexchange achieve. Retrieved from https://archive.org/details/stackexchange.
[33] Top Publications. Retrieved from: https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computersecuritycryptography.
[34] UMBC Ebiquity. Retrieved from http://ebiquity.umbc.edu/r/355.
[35] What is Open Indicators of Compromise (OpenIOC) Framework? Retrieved from https://cyware.com/educational-guides/cyber-threat-intelligence/what-is-open-indicators-of-compromise-openioc-framework-ed9d.
[36] Working with ATTACK. Retrieved from https://attack.mitre.org/docs/enterprise-attack-v11.2/enterprise-attack-v11.2-datasources.xlsx.
[37] Bank of England. 2016. *CBEST Intelligence-Led Testing-Understanding Cyber Threat Intelligence Operations*. Bank of England, Technical Report.
[38] Staff Contributor. 2020. What is Threat Intelligence? Retrieved from https://www.dnsstuff.com/what-is-threat-intelligence.
[39] Kurt Baker. 2022. What is cyber threat intelligence. Retrieved from https://www.crowdstrike.com/epp-101/threat-intelligence/.
[40] Catalin Cimpanu. 2020. University of Utah pays USD 457,000 to ransomware gang. Retrieved from https://www.zdnet.com/article/university-of-utah-pays-457000-to-ransomware-gang/.
[41] Henry Dalziel. 2014. Introduction. In *How to Define and Build an Effective Cyber Threat Intelligence Capability*. Syngress.
[42] Sagar Samtani, Hongyi Zhu, and Hsinchun Chen. 2020. Proactively identifying emerging hacker threats from the dark web: A diachronic graph embedding framework (D-GEF). *ACM Trans. Privac. Secur.* 23, 4 (2020), 1–33.
[43] Gbadebo Ayoade, Swarup Chandra, Latifur Khan, Kevin Hamlen, and Bhavani Thuraisingham. 2018. Automated threat report classification over multi-source data. In *IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 236–245.
[44] Mohamad Syahir Abdullah, Anazida Zainal, Mohd Aizaini Maarof, and Mohamad Nizam Kassim. 2018. Cyber-attack features for detecting cyber threat incidents from online news. In *Cyber Resilience Conference (CRC)*. IEEE, 1–4.
[45] Md Sahrom Abu, Siti Rahayu Selamat, Aswami Ariffin, and Robiah Yusof. 2018. Cyber threat intelligence—Issue and challenges. *Indon. J. Electric. Eng. Comput. Sci.* 10, 1 (2018), 371.
[46] Fernando Alves, Pedro Miguel Ferreira, and Alysson Bessani. 2019. Design of a classification model for a Twitter-based streaming threat monitor. In *49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. IEEE, 9–14.
[47] Ehsan Amjadian, Nicholas Prayogo, Serena McDonnell, Cathal Smyth, and Muhammad Rizwan Abid. 2021. Attended-over distributed specificity for information extraction in cybersecurity. In *IEEE Aerospace Conference*. IEEE, 1–12.
[48] Fernando Alves, Aurélien Bettini, Pedro M. Ferreira, and Alysson Bessani. 2021. Processing tweets for cybersecurity threat awareness. *Inf. Syst.* 95 (2021), 101586.
[49] Sofia Alevizopoulou, Paris Koloveas, Christos Tryfonopoulos, and Paraskevi Raftopoulou. 2021. Social media monitoring for IoT cyber-threats. In *IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 436–441.
[50] Mohammad Al-Ramahi, Izzat Alsmadi, and Joshua Davenport. 2020. Exploring hackers assets: Topics of interest as indicators of compromise. In *7th Symposium on Hot Topics in the Science of Security*. ACM, 1–4.
[51] Robert A. Bridges, Corinne L. Jones, Michael D. Iannacone, Kelly M. Testa, and John R. Goodall. 2014. Automatic Labeling for Entity Extraction in Cyber Security. Retrieved from http://arxiv.org/abs/1308.4941.
[52] Victor Benjamin, Weifeng Li, Thomas Holt, and Hsinchun Chen. 2015. Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 85–90.
[53] Robert A. Bridges, Kelly M. T. Huffer, Corinne L. Jones, Michael D. Iannacone, and John R. Goodall. 2017. Cybersecurity automated information extraction techniques: Drawbacks of current methods, and enhanced extractors. In *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 437–442.
[54] Vahid Behzadan, Carlos Aguirre, Avishek Bose, and William Hsu. 2018. Corpus and deep learning classifier for collection of cyber threat indicators in Twitter stream. In *IEEE International Conference on Big Data (Big Data)*. IEEE, 5002–5007.

[55] Avishek Bose, Vahid Behzadan, Carlos Aguirre, and William H. Hsu. 2019. A novel approach for detection and ranking of trendy and emerging cyber threat events in Twitter streams. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 871–878.

[56] Tao Bo, Yue Chen, Can Wang, Yunwei Zhao, Kwok-Yan Lam, Chi-Hung Chi, and Hui Tian. 2019. TOM: A threat operating model for early warning of cyber security threats. In *Advanced Data Mining and Applications*. Vol. 11888. Springer International Publishing, 696–711.

[57] Richard Colbaugh and Kristin Glass. 2011. Proactive defense for evolving cyber threats. In *IEEE International Conference on Intelligence and Security Informatics*. IEEE, 125–130.

[58] Jeffrey C. Carver, Edgar Hassler, Elis Hernandes, and Nicholas A. Kraft. 2013. Identifying barriers to the systematic literature review process. In *ACM / IEEE International Symposium on Empirical Software Engineering and Measurement*. IEEE, 203–212.

[59] Chia-Mei Chen, Dan-Wei Wen, Ya-Hui Ou, Wei-Chih Chao, and Zheng-Xun Cai. 2021. Retrieving potential cybersecurity information from hacker forums. *Int. J. Netw. Secur.* 23, 6 (2021), 1126–1138.

[60] Chia-Mei Chen, Jing-Yun Kan, Ya-Hui Ou, Zheng-Xun Cai, and Albert Guan. 2021. Threat action extraction using information retrieval. In *Computer Science & Information Technology (CS & IT)*. AIRCC Publishing Corporation, 13–19.

[61] Isuf Deliu, Carl Leichter, and Katrin Franke. 2017. Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks. In *IEEE International Conference on Big Data (Big Data)*. IEEE, 3648–3656.

[62] Isuf Deliu, Carl Leichter, and Katrin Franke. 2018. Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent Dirichlet allocation. In *IEEE International Conference on Big Data (Big Data)*. IEEE, 5008–5013.

[63] Fangzhou Dong, Shaoxian Yuan, Haoran Ou, and Liang Liu. 2018. New cyber threat discovery from darknet marketplaces. In *IEEE Conference on Big Data and Analytics (ICBDA)*. IEEE, 62–67.

[64] Nuno Dionísio, Fernando Alves, Pedro M. Ferreira, and Alysson Bessani. 2019. Cyberthreat detection from Twitter using deep neural networks. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[65] Yong Fang, Jian Gao, Zhonglin Liu, and Cheng Huang. 2020. Detecting cyber threat event from Twitter using IDCNN and BiLSTM. *Appl. Sci.* 10, 17 (2020), 5922.

[66] Paolo Frasconi, Daniele Baracchi, Betti Giusti, Ada Kura, Gaia Spaziani, Antonella Cherubini, Silvia Favilli, Andrea Di Lenarda, Guglielmina Pepe, and Stefano Nistri. 2021. Two-dimensional aortic size normalcy: A novelty detection approach. *Diagnostics* 11, 2 (2021), 220.

[67] John Grisham, Sagar Samtani, Mark Patton, and Hsinchun Chen. 2017. Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 13–18.

[68] Yumna Ghazi, Zahid Anwar, Rafia Mumtaz, Shahzad Saleem, and Ali Tahir. 2018. A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources. In *International Conference on Frontiers of Information Technology (FIT)*. IEEE, 129–134.

[69] Houssem Gasmi, Jannik Laval, and Abdelaziz Bouras. 2019. Information extraction of cybersecurity concepts: An LSTM approach. *Appl. Sci.* 9, 19 (2019), 3945.

[70] Apurv Singh Gautam, Yamini Gahlot, and Pooja Kamat. 2020. Hacker forum exploit and classification for proactive cyber threat intelligence. In *Inventive Computation Technologies*. Vol. 98. Springer International Publishing, 279–285.

[71] Ghaith Husari, Ehab Al-Shaer, Mohiuddin Ahmed, Bill Chu, and Xi Niu. 2017. TTPDrill: Automatic and accurate extraction of threat actions from unstructured text of CTI sources. In *33rd Annual Computer Security Applications Conference*. ACM, 103–115.

[72] Ghaith Husari, Xi Niu, Bill Chu, and Ehab Al-Shaer. 2018. Using entropy and mutual information to extract threat actions from cyber threat intelligence. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 1–6.

[73] Jack Hughes, Seth Aycock, Andrew Caines, Paula Buttery, and Alice Hutchings. 2020. Detecting trending terms in cybersecurity forum discussions. In *6th Workshop on Noisy User-generated Text (W-NUT'20)*. Association for Computational Linguistics, 107–115.

[74] Cheng Huang, Yongyan Guo, Wenbo Guo, and Ying Li. 2021. HackerRank: Identifying key hackers in underground forums. *Int. J. Distrib. Sensor Netw.* 17, 5 (2021), 155014772110151.

[75] Zafar Iqbal, Zahid Anwar, and Rafia Mumtaz. 2018. STIXGEN - A novel framework for automatic generation of structured cyber threat information. In *International Conference on Frontiers of Information Technology (FIT)*. IEEE, 241–246.

[76] Denis Iorga, Dragos-Georgian Corlatescu, Octavian Grigorescu, Cristian Sandescu, Mihai Dascalu, and Razvan Rughinis. 2021. Yggdrasil - Early detection of cybernetic vulnerabilities from Twitter. In *23rd International Conference on Control Systems and Computer Science (CSCS)*. IEEE, 463–468.

[77] Arnav Joshi, Ravendar Lal, Tim Finin, and Anupam Joshi. 2013. Extracting cybersecurity related linked data from text. In *IEEE 7th International Conference on Semantic Computing*. IEEE, 252–259.

[78] Corinne L. Jones, Robert A. Bridges, Kelly M. T. Huffer, and John R. Goodall. 2015. Towards a relation extraction framework for cyber-security concepts. In *10th Annual Cyber and Information Security Research Conference*. ACM, 1–4.

[79] Taoran Ji, Xuchao Zhang, Nathan Self, Kaiqun Fu, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Feature driven learning framework for cybersecurity event detection. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 196–203.

[80] Rupinder Paul Khandpur, Taoran Ji, Steve Jan, Gang Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2017. Crowdsourcing cybersecurity: Cyber attack detection using social media. In *ACM Conference on Information and Knowledge Management*. ACM, 1049–1057.

[81] Masashi Kadoguchi, Shota Hayashi, Masaki Hashimoto, and Akira Otsuka. 2019. Exploring the dark web for cyber threat intelligence using machine leaning. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 200–202.

[82] Nakhyun Kim, Minseok Kim, Seulgi Lee, Hyeisun Cho, Byung-ik Kim, Jun-hyung Park, and MoonSeog Jun. 2019. Study of natural language processing for collecting cyber threat intelligence using SyntaxNet. In *3rd International Symposium of Information and Internet Technology (SYMINTECH'18)*. Springer International Publishing, 10–18.

[83] Paris Koloveas, Thanasis Chantzios, Sofia Alevizopoulou, Spiros Skiadopoulos, and Christos Tryfonopoulos. 2021. inTIME: A machine learning-based framework for gathering and leveraging web data to cyber-threat intelligence. *Electronics* 10, 7 (2021), 818.

[84] Xiaojing Liao, Kan Yuan, XiaoFeng Wang, Zhou Li, Luyi Xing, and Raheem Beyah. 2016. Acing the IOC game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In *ACM SIGSAC Conference on Computer and Communications Security*. ACM, 755–766.

[85] Kuo-Chan Lee, Chih-Hung Hsieh, Li-Jia Wei, Ching-Hao Mao, Jyun-Han Dai, and Yu-Ting Kuang. 2017. Sec-Buzzer: Cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation. *Soft Comput.* 21, 11 (2017), 2883–2896.

[86] Quentin Le Sceller, ElMouatez Billah Karbab, Mourad Debbabi, and Farkhund Iqbal. 2017. SONAR: Automatic detection of cyber security events over the Twitter stream. In *12th International Conference on Availability, Reliability and Security*. ACM, 1–11.

[87] Ke Li, Hui Wen, Hong Li, Hongsong Zhu, and Limin Sun. 2018. Security OSIF: Toward automatic discovery and analysis of event based cyber threat intelligence. In *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)*. IEEE, 741–747.

[88] Zi Long, Lianzhi Tan, Shengping Zhou, Chaoyang He, and Xin Liu. 2019. Collecting indicators of compromise from unstructured text of cybersecurity articles using neural-based sequence labelling. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[89] Mengming Li, Rongfeng Zheng, Liang Liu, and Pin Yang. 2019. Extraction of threat actions from threat-related articles using multi-label machine learning classification method. In *2nd International Conference on Safety Produce Informatization (IICSPI)*. IEEE, 428–431.

[90] Ba Dung Le, Guanhua Wang, Mehwish Nasim, and Ali Babar. 2019. Gathering cyber threat intelligence from Twitter using novelty classification. In *International Conference on Cyberworlds (CW)*. IEEE, 316–323.

[91] Valentine Solange Marine Legoy. 2019. *Retrieving ATT&CK tactics and techniques in cyber threat reports*. Master's thesis. University of Twente. Retrieved from http://essay.utwente.nl/80012/.

[92] Dong Li, Xiao Zhou, and Ao Xue. 2020. Open source threat intelligence discovery based on topic detection. In *29th International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 1–4.

[93] Yali Luo, Shengqin Ao, Ning Luo, Changxin Su, Peian Yang, and Zhengwei Jiang. 2021. Extracting threat intelligence relations using distant supervision and neural networks. In *Advances in Digital Forensics XVII*. Vol. 612. Springer International Publishing, 193–211.

[94] Ying Li, Jiaxing Cheng, Cheng Huang, Zhouguo Chen, and Weina Niu. 2021. NEDetector: Automatically extracting cybersecurity neologisms from hacker forums. *J. Inf. Secur. Applic.* 58 (2021), 102784.

[95] Varish Mulwad, Wenjia Li, Anupam Joshi, Tim Finin, and Krishnamurthy Viswanathan. 2011. Extracting information about security vulnerabilities from web text. In *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE, 257–260.

[96] Mitch Macdonald, Richard Frank, Joseph Mei, and Bryan Monk. 2015. Identifying digital threats in a hacker web forum. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 926–933.

[97] Sudip Mittal, Prajit Kumar Das, Varish Mulwad, Anupam Joshi, and Tim Finin. 2016. CyberTwitter: Using Twitter to generate alerts for cybersecurity threats and vulnerabilities. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 860–867.

[98] Syed Shariyar Murtaza, Wael Khreich, Abdelwahab Hamou-Lhadj, and Ayse Basar Bener. 2016. Mining trends and patterns of software vulnerabilities. *J. Syst. Softw.* 117 (2016), 218–228.

[99] Eric Nunes, Ahmad Diab, Andrew Gunn, Ericsson Marin, Vineet Mishra, and Vivin Paliath, John Robertson, Jana Shakarian, Amanda Thart, and Paulo Shakarian. 2016. Darknet and deepnet mining for proactive cybersecurity threat intelligence. In *IEEE Conference on Intelligence and Security Informatics (ISI)*. IEEE, 7–12.

[100] Lorenzo Neil, Sudip Mittal, and Anupam Joshi. 2018. Mining threat intelligence about open-source projects and libraries from code repository issues and bug reports. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 7–12.

[101] Amirreza Niakanlahiji, Jinpeng Wei, and Bei-Tseng Chu. 2018. A natural language processing based trend analysis of advanced persistent threat techniques. In *IEEE International Conference on Big Data (Big Data)*. IEEE, 2995–3000.

[102] Amirreza Niakanlahiji, Lida Safarnejad, Reginald Harper, and Bei-Tseng Chu. 2019. IoCMiner: Automatic extraction of indicators of compromise from Twitter. In *IEEE International Conference on Big Data (Big Data)*. IEEE, 4747–4754.

[103] Umara Noor, Zahid Anwar, Tehmina Amjad, and Kim-Kwang Raymond Choo. 2019. A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise. *Fut. Gen. Comput. Syst.* 96 (2019), 227–242.

[104] Tatsuya Nagai, Makoto Takita, Keisuke Furumoto, Yoshiaki Shiraishi, Kelin Xia, Yasuhiro Takano, Masami Mohri, and Masakatu Morii. 2019. Understanding attack trends from security blog posts using guided-topic model. *J. Inf. Process.* 27 (2019), 802–809.

[105] Paweł Pawlinski, Przemylaw Jaroszewski, Piotr Kijewski, Lukasz Siewierski, Pawel Jacewicz, Przemyslaw Zielony, and Radoslaw Zuber. 2014. *Actionable Information for Security Incident Response*. Technical Report. European Union Agency for Network and Information Security.

[106] Lior Perry, Bracha Shapira, and Rami Puzis. 2019. NO-DOUBT: Attack attribution based on threat intelligence reports. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 80–85.

[107] Aritran Piplai, Sudip Mittal, Anupam Joshi, Tim Finin, James Holt, and Richard Zak. 2020. Creating cybersecurity knowledge graphs from malware after action reports. *IEEE Access* 8 (2020), 211691–211703.

[108] Panos Panagiotou, Christos Iliou, Konstantinos Apostolou, Theodora Tsikrika, Stefanos Vrochidis, Periklis Chatzimisios, and Ioannis Kompatsiaris. 2021. Towards selecting informative content for cyber threat intelligence. In *IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 354–359.

[109] Andrei Lima Queiroz, Susan Mckeever, and Brian Keegan. 2019. Eavesdropping hackers: Detecting software vulnerability communication on social media using text mining. In *4th International Conference on Cyber-technologies and Cyber-systems*. 41–48.

[110] Alan Ritter, Evan Wright, William Casey, and Tom Mitchell. 2015. Weakly supervised extraction of computer security events from Twitter. In *24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 896–905.

[111] Roshni R. Ramnani, Karthik Shivaram, and Shubhashis Sengupta. 2017. Semi-automated Information extraction from unstructured threat advisories. In *10th Innovations in Software Engineering Conference*. ACM, 181–187.

[112] Md Rayhanur Rahman, Rezvan Mahdavi-Hezaveh, and Laurie Williams. 2020. A literature review on mining cyberthreat intelligence from unstructured texts. In *International Conference on Data Mining Workshops (ICDMW)*. IEEE, 516–525.

[113] Thea Riebe, Tristan Wirth, Markus Bayer, Philipp Kühn, Marc-André Kaufhold, Volker Knauthe, Stefan Guthe, and Christian Reuter. 2021. CySecAlert: An alert generation system for cyber security events using open source intelligence data. In *Information and Communications Security*. Vol. 12918. Springer International Publishing, 429–446.

[114] Sagar Samtani, Ryan Chinn, and Hsinchun Chen. 2015. Exploring hacker assets in underground forums. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 31–36.

[115] Carl Sabottke, Octavian Suciu, and Tudor Dumitraş. 2015. Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real-world exploits. In *24th USENIX Security Symposium (USENIX'15)*. USENIX, 1041–1056.

[116] Sagar Samtani, Kory Chinn, Cathy Larson, and Hsinchun Chen. 2016. AZSecure hacker assets portal: Cyber threat intelligence and malware analysis. In *IEEE Conference on Intelligence and Security Informatics (ISI)*. IEEE, 19–24.

[117] Zareen Syed, Ankur Padia, Timothy W. Finin, Lisa Mathews, and Anupam Joshi. 2016. UCO: A unified cybersecurity ontology. In *Proceeding of the AAAI Workshop: Artificial Intelligence for Cyber Security*.

[118] Anna Sapienza, Alessandro Bessi, Saranya Damodaran, Paulo Shakarian, Kristina Lerman, and Emilio Ferrara. 2017. Early warnings of cyber threats in online discussions. In *IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 667–674.

[119] Sagar Samtani, Ryan Chinn, Hsinchun Chen, and Jay F. Nunamaker Jr. 2017. Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. *J. Manag. Inf. Syst.* 34, 4 (2017), 1023–1053.

[120] Clemens Sauerwein, Christian Sillaber, Andrea Mussmann, and Ruth Breu. 2017. Threat intelligence sharing platforms: An exploratory study of software vendors and research perspectives. In *13th International Conference on Wirtschaftsinformatik (WI'17)*. 837–851.

[121] Anna Sapienza, Sindhu Kiranmai Ernala, Alessandro Bessi, Kristina Lerman, and Emilio Ferrara. 2018. DISCOVER: Mining online chatter for emerging cyber threats. In *the Web Conference*. ACM Press, 983–990.
[122] Han-Sub Shin, Hyuk-Yoon Kwon, and Seung-Jin Ryu. 2020. A new text classification model based on contrastive word embedding for detecting cybersecurity intelligence in Twitter. *Electronics* 9, 9 (2020), 1527.
[123] Kiavash Satvat, Rigel Gjomemo, and V. N. Venkatakrishnan. 2021. Extractor: Extracting attack behavior from threat reports. In *IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 598–615.
[124] Hyejin Shin, WooChul Shim, Saebom Kim, Sol Lee, Yong Goo Kang, and Yong Ho Hwang. 2021. #Twiti: Social listening for threat intelligence. In *the Web Conference*. ACM, 92–104.
[125] Wiem Tounsi and Helmi Rais. 2018. A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Comput. Secur.* 72 (2018), 212–233.
[126] Katja Tuma, Gül Calikli, and Riccardo Scandariato. 2018. Threat analysis of software systems: A systematic literature review. *J. Syst. Softw.* 144 (2018), 275–294.
[127] Hieu Man Duc Trong, Duc-Trong Le, Amir Pouran Ben Veyseh, Thuât Nguyên, and Thien Huu Nguyen. 2020. Introducing a new dataset for event detection in cybersecurity texts. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 5381–5390.
[128] Uğur Tekin and Ercan Nurcan Yilmaz. 2021. Obtaining cyber threat intelligence data from Twitter with deep learning methods. In *5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE, 82–86.
[129] Aaruni Upadhyay, Samira Eisaloo Gharghasheh, and Sanaz Nakhodchi. Mapping CKC model through NLP modelling for APT groups reports. In *Handbook of Big Data Analytics and Forensics*. Springer International Publishing.
[130] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering*. Springer Berlin.
[131] Ryan Williams, Sagar Samtani, Mark Patton, and Hsinchun Chen. 2018. Incremental hacker forum exploit collection and classification for proactive cyber threat intelligence: An exploratory study. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 94–99.
[132] Tianyi Wang and Kam Pui Chow. 2019. Automatic tagging of cyber threat intelligence unstructured data using semantics extraction. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 197–199.
[133] Thomas D. Wagner, Khaled Mahbub, Esther Palomar, and Ali E. Abdallah. 2019. Cyber threat intelligence sharing: Survey and research directions. *Comput. Secur.* 87 (2019), 101589.
[134] Xuren Wang, Rong Chen, Binghua Song, Jie Yang, Zhengwei Jiang, Xiaoqing Zhang, Xiaomeng Li, and Shengqin Ao. 2021. A method for extracting unstructured threat intelligence based on dictionary template and reinforcement learning. In *IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 262–267.
[135] Yiming Wu, Qianjun Liu, Xiaojing Liao, Shouling Ji, Peng Wang, Xiaofeng Wang, Chunming Wu, and Zhao Li. 2021. Price TAG: Towards semi-automatically discovery tactics, techniques and procedures of e-Commerce cyber threat intelligence. *IEEE Trans. Depend. Secure Comput.* (2021), 1–1.
[136] Wenjun Xiong and Robert Lagerström. 2019. Threat modeling—A systematic literature review. *Comput. Secur.* 84 (2019), 53–69.
[137] Zhe Yu and Tim Menzies. 2019. FAST2: An intelligent assistant for finding relevant papers. *Expert Syst. Applic.* 120 (2019).
[138] Wenzhuo Yang and Kwok-Yan Lam. 2020. Automated cyber threat intelligence reports classification for early warning of cyber attacks in next generation SOC. In *Information and Communications Security*. Vol. 11999. Springer International Publishing, 145–164.
[139] He Zhang, Muhammad Ali Babar, and Paolo Tell. 2011. Identifying relevant studies in software engineering. *Inf. Softw. Technol.* 53, 6 (2011), 625–637.
[140] Ziyun Zhu and Tudor Dumitraş. 2016. FeatureSmith: Automatically engineering features for malware detection by mining the security literature. In *ACM SIGSAC Conference on Computer and Communications Security*. ACM, 767–778.
[141] Ziyun Zhu and Tudor Dumitras. 2018. ChainSmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports. In *IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 458–472.
[142] Panpan Zhang, Jing Ya, Tingwen Liu, Quangang Li, Jinqiao Shi, and Zhaojun Gu. 2019. iMCircle: Automatic mining of indicators of compromise from the web. In *IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 1–6.
[143] Jun Zhao, Qiben Yan, Jianxin Li, Minglai Shao, Zuti He, and Bo Li. 2020. TIMiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data. *Comput. Secur.* 95 (2020), 101867.
[144] Huixia Zhang, Guowei Shen, Chun Guo, Yunhe Cui, and Chaohui Jiang. 2021. EX-action: Automatically extracting threat actions from cyber threat intelligence report based on multimodal learning. *Secur. Commun. Netw.* 121 (2021), 1–12.

[145] Wenli Zeng, Zhi Liu, Yaru Yang, Gen Yang, and Qin Luo. 2021. QBC inconsistency-based threat intelligence IOC recognition. *IEEE Access* 9 (2021), 153102–153107.
[146] Swati Khandelwal. 2019. New Group of Hackers Targeting Businesses with Financially Motivated Cyber Attacks. Retrieved from https://thehackernews.com/2019/11/financial-cyberattacks.html.
[147] Marry L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochem. Medica* 22, 3 (2012), 276–282.
[148] Rob McMillan. 2013. Definition: Threat intelligence. Retrieved from https://www.gartner.com/en/documents/2487216.
[149] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. DOI: https://doi.org/10.48550/arXiv.1301.3781
[150] Larry Ponemon. 2014. *Exchanging Cyber Threat Intelligence: There Has to Be a Better Way*. Technical Report. Ponemon Institute Research Report, Ponemon Institute LLC.
[151] Jon Porter. 2020. Amazon says it mitigated the largest DDoS attack ever recorded. Retrieved from https://www.theverge.com/2020/6/18/21295337/amazon-aws-biggest-ddos-attack-ever-2-3-tbps-shield-github-netscout-arbor.
[152] Johnny Saldaña. 2015. *The Coding Manual for Qualitative Researchers*. Sage.
[153] Sagar Samtani, Murat Kantarcioglu, and Hsinchun Chen. 2020. Trailblazing the artificial intelligence for cybersecurity discipline: A multi-disciplinary research roadmap. *ACM Trans. Manag. Inf. Syst.* 11, 4 (2020), 1–19.
[154] Bruce Schneier. 1998. Security pitfalls in cryptography. In *Proceeding of the EDI FORUM-OAK PARK-*, Vol. 11, THE EDI GROUP, LTD., 65–69.
[155] Dave Shackleford. 2015. *Who's Using Cyberthreat Intelligence and How?* Technical Report. SANS Institute.
[156] Donna Spencer. 2009. *Card Sorting: Designing Usable Categories*. Rosenfeld Media.
[157] K. Zurkus. 2015. Threat intelligence needs to grow up. Retrieved from https://www.csoonline.com/article/2969275/threat-intelligence-needs-to-grow-up.html.

Received 4 June 2021; revised 12 July 2022; accepted 17 October 2022