Nr	Author	Year	Title	Brief summary	Data sources: Social Networks	News Articles	Security Reports	Public repositories (GitHub, exploit-DBs, etc)	CTI/Feeds	Underground forums	Darknet marketplace	Darknet websites	Did the authors provide any reason for the selection of OSINT data sources?	Reason	Did the authors consider any alternative data source?	Unsuccessful (e.g., poor)	Super Easy Type: Tested, unsuccessful (e.g., tools that are unfeasible, poor, not yet mature)	Frameworks(Eg. STIX)	Method: No AI	AIML	Details	NSP Frameworks: Name/Identity Recognition(NER); Topic Modeling(TM); Sentiment Analysis (SA); Cybersecurity Named Entity Recognition (Custom); Keyword Extraction (KE)	Type of CTI mentioned(vulnerabilities, attacks, sensor/measures, attack chains, malware, etc)	Type of CTI	unsuccessful CTI	semi-successful CTI	successful CTI (OTH)	Other	Outcome	General Remarks	Link
0	Fabio Giasson	2022	Example - CTI Data Source	Site page where data is extracted CTI data Source with checkpoint collect newest CTI	x				Twitter				No	Twitter Twitter has the cybersecurity space and was connected with CTI practitioners as a platform for data source for enterprises	No					x	MIP		Unknown(Unknown CTI)							
1	Hyojin Shin	2021	IP-Arch: Social Listening for Threat Intelligence	Twitter is used to extract IOC					Twitter				Yes	Twitter Twitter is used to extract malicious robots, timeliness, and diversity in attacks. However, they did not evaluate alternative sources. Twitter is a well-known platform but African Twitter quality is why they sources considered.	No	x				x	NER (In-preprocessing) TC KE(URLs, IP, Hashes etc.)	IOC: mainly malware IOC (IP addresses, URL file hashes); but they assert it could be easily modified to analyze other things (Phishing or spam for example).	Textual			malware IOC	IOC: mainly malware IOC (IP addresses, URL file hashes); but they assert it could be easily modified to analyze other things (Phishing or spam for example).		
2	Victor Adeyemo	2020	Exploiting Open Source Information for Cyber Threat Intelligence	System with 87% accuracy for CTI extraction. On Twitter and Slashdot forums, and analysis of 8500 tweets and 1000 posts of its data profiling for pen-hunters cybersecurity	x	x	x	x	x				Yes	They say Twitter and the darknet are an important sources to gather the relevant information, however they gather tweets that used the same data sources.	No	x	x			x	Logging, random forest classifier, gradient boosting	TC	Classification of raw into CTI relevant or irrelevant	Textual		x	unsuccessful CTI	Classification of raw into CTI relevant or irrelevant		
3	J. Zhao	2020	TIDExtract: Automatically extracting and analyzing unique CTI from Twitter	Development of a CTI extractor. Focus on tagging domain and extracting novel threats	x	x	x	x		x		x	Yes	Twitter, open or less enhanced all possible sources. But in their DRA there are some cons (lag for organizations, but they have for example Intrusion Detection models)	No (but, more or less enhanced all possible sources. But in their DRA there are some cons (lag for organizations, but they have for example Intrusion Detection models)	x	x				x	CTI domain classifier Domain specific CTI generation	TC TG (Tweet indexes created)	Description/Domain/Vulnerability (CVE...)/ Filed/Intended/IP	Textual		x	CVE, IP, unsuccessful CTI	IOCs are combined with their corresponding domain tags to generate the domain-specific CTIs. Domain-specific CTIs can be shared with relevant CTI subscribers, and allow them to quickly identify the security posture of their respective industries. Moreover, Tweet Index is prepared to quantitatively measure the tweet severity caused by different types of attacks in each domain.		
4	P. Kalliosaa	2021	mTIME: A Multi-Lingual Learning-Based Framework for Gathering and Leveraging Web Data to Cyber Threat Intelligence	Automated Data Acquisition, Analysis, Management and sharing framework	x	(x)	(x)		(x)	(x)	(x)	(x)	Yes	Twitter is used to ensure updated scenarios of the threat landscape. There are people which are sharing their data on Twitter. Vulnerabilities are shared and talked about before official announcements, giving more time to exploit.	No	x	x			x	NER, word embedding, AI also used in the results Vulnerability Recognizer, Multilingual Name Recognizer, Domain Cross Classifier, Keyword Sensitivity Classifier, Support Vector Machines Classifier, Sentiment Classifier, as well as CNN-based Sentiment Classifier	NER TC KE (Tweet index) AG	Structured CTI, STIX, Lifter	Operational		x	unsuccessful CTI(STIX)	To create CTI so collect the vulnerability (URL) and weaknesses (CEC) objects predictability MOP. Additionally, we have created extension made objects to encapsulate information that is provided by our sources but is not modelled by the vulnerability and weakness objects. We have created two extension made objects to encapsulate information that is provided by our sources but is not modelled by the vulnerability and weakness objects, namely, Tweet and Post objects respectively. Finally, we every object accross a cosmo dimeribe in the ontologies to be able to distinguish the source of the extracted CTI. - CTI management and sharing via open standards and iterative tools; this is similar to our approach, however we focus on STIX/TAXII and by extending to standardized CTI - translation to spoken, quick and machine-readable graphs.	https://www.sciencedirect.com/science/article/pii/S1389128621000311	
5	Uijin Tahn	2021	Obtaining Cyber Threat Intelligence Data From Twitter With Deep Learning Methods	DL classifier for categorizing tweets as relevant or not and another classifier for type of CTI	x								Yes	Twitter (only English posts)	Yes, but only mentioned in the related works	x					x	LSTM and Deep NN layers were used in the classifiers	TC	Classification of raw into CTI relevant or irrelevant. Classification into CTI type (vulnerability, ransomware, IDoS, General, EmoDog, Hatred and Leak)	Textual	x	x	CVE, threats, unsuccessful CTI	Classification of raw into CTI relevant or irrelevant. Classification into CTI type (vulnerability, ransomware, IDoS, General, EmoDog, Hatred and Leak)	Grammatical errors throughout this paper. Some illustrations are crisp, grainy. The accuracy of short-results is extremely low and in most of the classes they differ from the average.	https://www.hindawi.com/journals/wcmc/2021/6627111/
6	Lima Al-al-Ketanaana	2020	CTI Twitter: Gathering Cyber Threat Intelligence from Twitter using Adversarial Supervised and Unsupervised Learning	CTI extractor from twitter using BERT for binary classification and multi-class classification (what kind of CTI is the tweet). After that semantic clustering is done to remove the resulting CTI.	x								Yes	Twitter is open-free and has a large overhear of cyber-related tweets. Twitter is used because they mention (similar) CTI extractors for twitter, and discuss specific examples of CTI that are shared on twitter. Twitter is reported to remedy before the mainstream media. Focusing on specific accounts rather than all twitter accounts. Twitter is used to detect malware issues. Vulnerabilities are discussed on twitter before hotfixes go to national vulnerability databases.	Yes, Darkweb is mentioned but described as difficult.	x					x	BERT is used for the classifiers. Word2Vec and LDA is used for semantic clustering	TM TC KE	Operations direct by the algorithm: 1) Classification into CTI or not 2) Classification into types (MOV, ATK, CVE, MAL, IOHC) 3) Semantic clustering	Textual		x	unsuccessful CTI, attacks, CVE, malware	Operations direct by the algorithm: 1) Classification into CTI or not 2) Classification into types (MOV, ATK, CVE, MAL, IOHC) 3) Semantic clustering		https://www.hindawi.com/journals/wcmc/2020/8879229/
7	Ariel Rodriguez	2020	Enhancing data quality in real-time threat intelligence systems using machine learning	Improvements on keyword based filterings to extract CTI. They use word2vec and then do k-means clustering to remove irrelevant clusters	x	x	x						Yes	Twitter, Reddit, Stack Exchange and news articles	Yes, github is also mentioned. They want to focus on the publicly available threats	x					x	Use of dynamic word2vec associated word vectors associated with a keyword filter. Hierarchical learning is used to better data quality. It seems like clustering BERT, Naive, Stochastic gradient descent, Linear regression are used as baselines. They develop the a level of correlation to enrich IoC in between each other.	TM TC	Relevant or irrelevant posts	Textual		x	unsuccessful CTI	Relevant or irrelevant posts	M-ake objective (data positive reduction) by phrase garbage to garbage out is mentioned multiple times)	https://link.springer.com/article/10.1007/s11276-020-02448-z#Table2
8	Roi Asveteli	2019	POSE: Generating Quality Threat Intelligence by Clustering and Correlating OSINT	They take IoCs from different OSINT feeds and correlate them. They also do semantic clustering tools, presents in TIP (Threat intel platforms). They also use TIP (Intel), as a source. The main contribution is the method of development of a similarity metric between IoC.				x					No	AS OSINT feeds from 12 organizations are the source of the data. Twitter and darknet are mentioned which is because of those feeds is.	No	x				x		TM TC (Indicp/Relation) Name (IoC, gc enriched)	Enriched IoC.	Textual		x	unsuccessful CTI	enriched indicators of compromise - CTI	https://www.sciencedirect.com/science/article/pii/S0167404819300756	
9	McGlenon Hosaye	2021	Generating cyber threat intelligence to discover potential security threats using classification and topic modelling	CTI is extracted from hacker forums through supervised text classification and topic modelling				x					Yes	They assert that hacker forum are the most reliable source of information. Hackers are and hacker forums are important information. They point out that hackers share intelligence in these forums.	Yes, they mention online security forums and social media is the alternate but never explain why they ignore them	x					x	Classification into CT relevant/irrelevant Classification into topics Classification is done with multiple different algorithms (naive bayes, logistic regression, support vector machine, SGD, KNN, RF, DT, LDA and non-negative matrix factorization are used for topic modeling	TM TC	Relevant or irrelevant posts Topics of Interest (hacking/exploit/bugs/DDos etc.) Topic of Interest, used to extract the most useful word in each topic	Textual		x	unsuccessful CTI		This is a chapter of thesis https://www.researchgate.net/publication/351120979_DISCOVEL_Organizing_cyber_threat_intelligence_discover_potential_security_threats_using_classification_and_topic_modelling Discover potential security threats using classification and topic modelling. Discover potential threats from various artificial, informal and informal underground cultural fresh. Should I call analyze it?	https://www.researchgate.net/publication/351120979_DISCOVEL_Organizing_cyber_threat_intelligence_discover_potential_security_threats_using_classification_and_topic_modelling https://www.researchgate.net/publication/351120979_DISCOVEL_Organizing_cyber_threat_intelligence_discover_potential_security_threats_using_classification_and_topic_modelling
10	Yongcan Gao	2021	A framework for threat intelligence extraction and fusion	Extraction, correlation and notification of CTI from underground forums and Twitter. They do something similar to a cybersecurity knowledge graph (CKG)			(x)		x	(x)			No	No mention of why twitter for example is not a good source	No	x	x			x	BERT, BiLSTG tested network and ensemble learning model. Bi-LSTM is used for NER. For factor they use an improved Levenshtein distance. The extraction is done as follows: 1) Crawl connections 2) Community Dimension Algorithm 3) Calculate the weight of each relation calculation edge expansion model is used 4) User endorsers created 5) Calculate the weight of each	Name (Name of dangerous is done) KE (exploits are created)	Extraction of triples (subject-verb-object relation, Name, Vulnerability, Exploit, Attack, Vulnerability, Vulnerability, CVE,2009-2018 CVE, CAE, and sometimes similar	Operational		x	Graphical visualization, threats	Knowledge graphs based on APTs, enriched security reports		
11	Avishek Bose	2021	Tracing Relevant Twitter Accounts Active in Cyber Threat Intelligence Domain by Exploiting Content and Structure of Twitter Networks	Twitter accounts are ranked according to contextual cyber security vocabulary information in regard to CTI relevance	x								No	NA	No	x	x			x	User ranking in regards to CTI relevancy is extracted	TM TG (User ranking)	User ranking in regards to CTI relevancy is extracted	Textual		x		CTI extractor ranking		https://www.sciencedirect.com/science/article/pii/S1877050921011534
12	Manasik Karligoda	2020	Deep Self-Supervised Clustering of the Dark Web for Cyber Threat Intelligence	DL classifier for categorizing darkweb forum post as relevant or not (for taking tweets or messages)				x					Yes	Darkweb forums. In related work darkweb pages classified and other forum posts analysis, are mentioned	No	x				x	Classification into CT relevant/irrelevant Clustering of the relevant posts Classification is done with multiple different algorithms (naive bayes, logistic regression, support vector machine, SGD, KNN, RF, DT, LDA and non-negative matrix factorization are used for topic modeling	TM	Forum posts containing information that needs preparation measures are extracted	Textual		x	unsuccessful CTI	classified set of forum posts into critical and non-critical	Dataset composed of 850 malware offers and 850 weaponized malware samples (Kaggle) (https://www.cs.fortiangel.com). Final training accuracy 57%, while unsure data 89%, so a medium-term, as relevant	https://www.researchgate.net/publication/348416000_Deep_Self-Supervised_Clustering_of_the_Dark_Web_for_Cyber_Threat_Intelligence
13	Yinghai Zhou	2022	CTI Slice: APT Threat Intelligence Analysis System	Automatic system that groups diverse CTI into one forum		x	x						No	NA	In related works twitter is mentioned.	x	x				NER (used for comparison against BERT) KE(STIX objects)	Balanced CTI with IOC, TTPS and other metadata (Attacker,tools,origin,Industry,campaign)	Balanced CTI with IOC, TTPS and other metadata (Attacker,tools,origin,Industry,campaign)	Textual		x	unsuccessful CTI, IOC, TTPS, attackers info, tools, origin, industry	Balanced CTI with IOC, TTPS and other metadata (Attacker,tools,origin,Industry,campaign)	Dataset might be small. (1200 reports)	https://www.hindawi.com/journals/wcmc/2022/7401386/
14	Georgios Sakellariou	2023	RECIFAN: A Cyber Threat Intelligence System for Discussion Forums Evaluation	A system to use discussion forums (DF) for the extraction of CTI. There's a differential analysis of DF					(x)				No	NA	In related works: underground forums, social networks, chat logs, cerberus a cybersecurity service, mentioned. In related work: Twitter, Reddit, Stack Exchange and news articles are mentioned for CTI research in general	x	x				x	Formal concepts were used, which might empty. Sl. Processing layers are explained in detail. Formal concepts are used in algorithms or their family are only mentioned (LDA, correlation, modelled classification	TM SA KE(STIX objects)	STIX objects (a specific CTI form)	Textual		x	unsuccessful CTI(STIX)	STIX objects	Strong focus on modeling how a CTI system should be modeled. Also a focus is a differential analysis to determine which DF is more valuable to be scanned. no experimental results are presented in this paper	https://www.sciencedirect.com/science/article/pii/S1389128623000311
15	Nolan Arnold	2019	Dark Net Ecosystem Cyber Threat Intelligence (CTI) Tool	Dark net monitoring tool that uses social networks to extract CTI					x		x		Yes	Twitter and darknet forums are considered as the data sources. The data is provided by the List of Arizona AI lab	No	x				x	Tools and strategies for branches are exchanged on the darknet. The data is provided by the List of Arizona AI lab	TM KE VA (visualization) AG	Visualization of the darknet ecosystem. Which assets are connected to which entity, which forum, my targeted ransomware	Strategic		x	Graphical visualization, threats	Visualization of the darknet ecosystem. Which assets are connected to which entity, which forum, my targeted ransomware.	This paper and tools of discoverable malicious hacker assets is the defined in section. The exact posing/operation of the algos and tools in the network is unclear, is very unclear.	https://www.researchgate.net/publication/342174164
16	Andrea Tundis	2022	A Fuzzy-driven Method for Assessing the Assessment of OSINT Cyber Threat Sources	A model to automate the assessment of cyber threat intelligence sources and predict a relevance score for each asset (so order is proposed)	x								Yes	Hacker post information about ongoing attacks.	The related works section is very extensive.	x					x	Based on structured features and acting as a kind of function of non-intelligence published, which is validated with the Open Cyber Exchange as external dataset. IOC and reputation analysis are used to rate a source. Reputation is based on the number of the hacktivist tags and their expertise. Cases are Process, Reputation and Linear SVM	NA (relevance score of a source is computed)	A relevance score of a source (usually a vector score with respect to a CTI).	Operational		x	unsuccessful CTI	CTI relevance score including utility, score, timeliness score, intelligence count	This paper is basically the same as the one above handed in 2 months prior. It only includes security reports in the data, otherwise the 2 paper are basically identical.	https://www.sciencedirect.com/science/article/pii/S0167404822000359 https://www.sciencedirect.com/science/article/pii/S0167404822000359
17	Cogney Yewel	2021	Detecting Cyber Behavior in Cyber Threat Intelligence Development of Honey pot System	Prediction of hacker's expertise and personality based on their behavior. Data is collected from crawl by mining models (a honeypot sensor, a CTI and a cyber expertise sys.									Yes	The prediction of non-behavior is the focus. Twitter is used because it is a social media and platform hacker behaviors could be advantageous for defenders.	Yes, for related work other sources for psychological information, cyber expertise and profiling hacker behaviors could be advantageous for defenders.	x				x	BIG, and reputation analysis are used to rate a source. Reputation is based on the number of the hacktivist tags and their expertise. Cases are Process, Reputation and Linear SVM	NA	Chart? Or reputation Expertise of the hacker Reputation of the hacker	Textual		x	attackers info	The CTI extracted is: Reputation of the hacker Expertise of the hacker	This say that one of the DM participants only 30 detail the hacker behavior. Interesting.	https://www.hindawi.com/journals/wcmc/2021/7929266/
18	Jeonghoon Oh	2020	Blockchain-Based Cyber Threat Intelligence System Architecture for Sustainable Computing	TIP design for reliability, privacy, reliability and sustainability. The main objective of the paper is to provide a blueprint for a CTI sharing platform platform with the purpose of sustaining CTI feeds.				x					No	NA	In related work hacker forums, SIEM, CTI feeds and Twitter are mentioned.	x	x				x	The architecture is based on blockchain technology	Name	Relevant CTI.	Textual		x	unsuccessful CTI	unprovoked CTI	This is a chapter of a thesis https://www.researchgate.net/publication/351120979_DISCOVEL_Organizing_cyber_threat_intelligence_discover_potential_security_threats_using_classification_and_topic_modelling Discover potential security threats using classification and topic modelling.	https://www.researchgate.net/publication/351120979_DISCOVEL_Organizing_cyber_threat_intelligence_discover_potential_security_threats_using_classification_and_topic_modelling
19	Benjamin Ampel	2020	Labeling Hacker Exploits for Proactive Cyber Threat Intelligence: A Deep Transfer Learning Approach	Framework for an automated collection and categorization of hacker forum exploit source code				x		x	x	x	Yes	Hacker Forums and Darknet Marketplaces. Its a source in the paper. Subflog, EnglishDB, NulledForum, Mesopotic, Vulnerlab, Zerocodeez	No	x	x			x	Data was scraped from the forums and source code/binaryr than 100 characters was collected. Logistic regression, Naive Bayes, SVM, LSTM, Bi-LSTM, Bi-LSTM-CRF, BERT, and pretrained through CBOW, pooling BiLSTM are tested. Bi-LSTM-CRF is used for the classification.	TC TG	Source code is classified into classes of exploits. (Web applications, Dos, Remote Issue, ...)	Textual		x	exploits	model for identifying exploits	https://www.ieee.org/publications/tcsi/index.html	https://www.ieee.org/publications/tcsi/index.html
20	Mouneika Das Prelia	2023	Extracting Actionable Cyber Threat Intelligence from Twitter Stream	Present a method to extract textual manifestations (e.g. modification of a particular registry key) from tweets and map them to the ATT&CK framework. The textual manifestations, which is the information on observable manifestations of an attack	x			x					No	NA	In related works CTI reports are mentioned as source	x	x				x	Twoer filtration for most/tweet scan, automated semantic labels and the extracted score phrases. Tweets are ranked by the TF-IDF score. The same procedure is done to match distances of a tweet to entries in the MITRE ATT&CK. Tweets are ranked by the TF-IDF. The proposed pipeline ahs field with a GPT-trained embedding for the analysis.	TC Semantic Role Labeling	Textbook manifestations (information on observable manifestations of an attack or the linking of IoCs with CTI)	Textual		x	threats	classified tweets if they are threat or not	https://www.sciencedirect.com/science/article/pii/S1877050923000403	
21	Tianliang Sun	2021	An Automatic Generation Approach of the Cyber Threat Intelligence Records Based on Multi-Source Information Fusion	Design of a method to automatically generate CTI records based on TIPs			x						Yes	They point out weaknesses to current TIP and want to connect them into machine-readable (such as STIX) formats.	Darkweb marketplaces mentioned in related works.	x (Step specifically design a HTML-reader)					x	Classification into threat/no threat articles is done by extracting features from the related articles and then using a MLP. Relevant information is then to be extracted with a method they developed called HRRG.	TC Name (STIX objects)	STIX objects (a specific CTI form) and a CTI database	Textual		x	unsuccessful CTI(STIX)	STIX	They often mention articles and OSINT(opensource TIP) and they connect one hack an aquisition which I feel confusing.	https://www.sciencedirect.com/science/article/pii/S1389128621000244
22	Shin-YingHuang	2020	Monitoring Social Media for Vulnerability-Threat Prediction and Topic Analysis	Model to predict how likely CVE events are to be exploited by analyzing twitter. Other meaningful information is also extracted.	x	x	x					x	Yes	Twitter, open public repository data	Yes	x	x			x	Prediction if CVE will be exploited is done with random forest LDA is used for topic analysis Classification CT relevant/irrelevant done with LSTM.	TC TG (Indp.based of exploits)	Twitter sentiment analysis. Tweets get classified into CT relevant/irrelevant, there is a prediction of how likely the exploit using the CVE is.	Operational		x	CVEs	rated the dates for high-risk vulnerabilities	The details of which event (they analyzed, or how they collected those events remains unclear to me. They collected 500 tweets and manually label them as CyBersec related or not.	https://www.ieee.org/publications/tcsi/index.html
23	A. Zenebe	2019	Cyber Threat Discovery from Dark Web	CTI is extracted from darknet blackhat forums, exploits (cyber) are predicted from the posts					x				Yes	Hackers share information and learning from each other. Hackers are active in the darknet, so darknet is more promotion	No	x	x			x	Classification into exploit (cyber) or not (Random Forest) Classification into Topics (Random Forest and Random Forest)	TC	Top forum hackers and exploit: were extracted. Exploit (cyber) was extracted from the post	Textual		x	CTI intelligence ranking, exploits	authors numbers per forum, number of posts per exploit, exploit types over time	Literary review (LR) contains Analysis of CTI. Discussion of Dark Web Forums. Dark Web Forums. They use the WEKA software tool which was mentioned in other sources too.	https://www.researchgate.net/publication/337406254_Cyber_Threat_Discovery_from_Dark_Web https://www.researchgate.net/publication/337406254_Cyber_Threat_Discovery_from_Dark_Web https://www.researchgate.net/publication/337406254_Cyber_Threat_Discovery_from_Dark_Web
24	Fabian Sabi	2023	A Very Social Media-Driven Cyber Threat Intelligence	ML to analyze centennial information from social media posts to generate a single index or a score level.	x								Yes	Data hacks of social medias. OSINT source are analyzed. In the conclusion it is pointed out that the analysis of social media information is a cyber security, or cyber-crime (hackers could falsify results)	Yes, closed network (Jamboards and snapchat) opposed to open networks (twitter instagram public profiles) are excluded	x					x	CNN based anomaly detection was used because the auto says that RNN are not supported on low code multi-platform	NA	Country, scale time series index on cyber-crime	Strategic		x	Threats	Country, scale time series index on cyberthreats	Literary review (LR) contains Analysis of CTI. Discussion of Dark Web Forums. Dark Web Forums. They use the WEKA software tool which was mentioned in other sources too.	https://www.sciencedirect.com/science/article/pii/S1877050923001463
25	Fabian Sabi	2023	A global cyber-threat intelligence system with artificial intelligence and convolutional neural network	ML to give CT country/cyber indices including data from routers and cybersecurity provides	x			x					No	There is no analysis as in the related paper above	No	x	x			x	CNN based anomaly detection was used because the auto says that RNN are not supported on low code multi-platform	NA	Country, scale time series index on cyber-crime	Strategic		x	Threats	Country, scale time series index on cyberthreats	This paper is basically the same as the one above handed in 2 months prior. It only includes security reports in the data, otherwise the 2 paper are basically identical.	https://www.sciencedirect.com/science/article/pii/S1877050923001463 https://www.sciencedirect.com/science/article/pii/S1877050923001463
26	Kolade Fiyomero	2023	Detecting Threat Intelligence Related IoT Botnet Posts Latest Dark Web Data Collection	Development of a crawler to scrape the darkweb forums. Data is collected from darkweb forums. Data is extracted further into top topics extract CTI from this database	x					x		x	No	Darkweb forums (mentioned) and darkweb forums are the sources.	No	x	x			x	Qualitative assessment are extracted such as Topic Modeling, Sentiment Analysis, Keyword Extraction. Quantitative analysis is done with data analysis (number of usernames and vendors etc. They also do a cross-analysis with Cybersecurity Vocabulary.	(Statistical qualitative results)	Statistical analysis of the data provided by the crawler can be done.	Textual		x	unsuccessful CTI	security reports including statistical analysis		https://www.sciencedirect.com/science/article/pii/S2352711023001693
27	Arpan Singh Gautam	2019	Darker Forum Digital and Classification for Proactive Cyber Threat Intelligence	ML model to classify hacker forum data and visualize it.					x				Yes	They say they focus on hacker forums which are publicly available. Association from 42 forum discussion boards (Hackers Arise, Hackers Hotel) They analyzed cyber data in Darkbot		x					x	RNN LSTM and RNN GRU used to classify posts into CT relevant or irrelevant	TC	Classification of raw into CTI relevant or irrelevant	Textual		x	unsuccessful CTI	Jaw comparison of two NN approaches, discovery of threat indicators		https://link.springer.com/chapter/10.1007/978-3-030-22868-2_18
28	Panos Panagiotou	2021	Towards Selecting Information Content for Cyber Threat Intelligence	With a focus on a good classification into CT relevant/irrelevant (CTI that were classification for cyberattacks vs. CTI that)	x								Yes	The Simple and high quality of data provided by Twitter and Darknet are considered. For classification SVM and Random Forest were used NER with validation from the Malware/FastNlk	Yes, Twitter and darkweb reddit and crack exchanges are mentioned	x (times are longer than poorly)	x (overlapped)			x	For text processing different approaches were used (Word2Vec) For classification SVM and Random Forest were used NER with validation from the Malware/FastNlk	NER TC	CT-relevant/irrelevant classification	Textual		x	unsuccessful CTI	CT-relevant/irrelevant classification	There are explanations on why the relevant/irrelevant classifications are important and what exactly or what would happen without this actual filtering	https://www.sciencedirect.com/science/article/pii/S1877050921011736