Karbala International Journal of Modern Science

Volume 11 | Issue 3 Article 15

Multilingual Cyber Threat Intelligence Feeds Preprocessing for Threat Intelligence Event Extraction: A Systematic Literature Review

Jamal H. Al-Yasiri
School of Computing, Universiti Utara Malaysia, Sintok, Malaysia, jhyondon@gmail.com

Mohamad Fadli bin Zolkipli
School of Computing, Universiti Utara Malaysia, Sintok, Malaysia

Nik Fatinah N. Mohd Farid
School of Computing, Universiti Utara Malaysia, Sintok, Malaysia

Multilingual Cyber Threat Intelligence Feeds Preprocessing for Threat Intelligence Event Extraction: A Systematic Literature Review

Abstract

In cyber threat intelligence (CTI), information security specialists face overwhelming data flows from multiple sources, including hacker forums, dark web markets, and social media. These diverse and multilingual information streams require extensive analysis and processing. However, the current state of CTI faces several challenges, such as reliance on manual annotation and evaluation, as well as limited support for non-English languages, which hinders advanced threat detection and comprehensive analysis. This systematic literature review proposes a conceptual framework designed to overcome existing state-of-the-art limitations. It evaluates recent advancements in CTI methodologies by following PRISMA guidelines and analyzing selected studies from reputable sources, including Scopus, IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar. Our findings emphasize the critical role of artificial intelligence, machine learning, and deep learning in enhancing CTI models by optimizing data collection, preprocessing, and multilingual support for event extraction. Despite significant improvements in event extraction accuracy and processing efficiency through AI-driven techniques, challenges remain in scaling automated systems and expanding language coverage. This review highlights the need for comprehensive, scalable frameworks that minimize manual effort while improving cross-lingual capabilities, ultimately enabling more robust, timely, and accurate threat intelligence extraction in the evolving cyber threat landscape.

Keywords

Cyber threat intelligence; Multilingual support; Data feeds; Event extraction; Text preprocessing

REVIEW ARTICLE

Multilingual Cyber Threat Intelligence Feeds Preprocessing for Threat Intelligence Event Extraction: A Systematic Literature Review

Jamal H. Al-Yasiri [a,b,*], Mohamad F.B. Zolkipli [a], Nik Fatinah N. Mohd Farid [a]

[a] School of Computing, Universiti Utara Malaysia, Sintok, Malaysia
[b] Control and Systems Engineering Department, University of Technology, Baghdad, Iraq

Abstract

In cyber threat intelligence (CTI), information security specialists face overwhelming data flows from multiple sources, including hacker forums, dark web markets, and social media. These diverse and multilingual information streams require extensive analysis and processing. However, the current state of CTI faces several challenges, such as reliance on manual annotation and evaluation, as well as limited support for non-English languages, which hinders advanced threat detection and comprehensive analysis. This systematic literature review proposes a conceptual framework designed to overcome existing state-of-the-art limitations. It evaluates recent advancements in CTI methodologies by following PRISMA guidelines and analyzing selected studies from reputable sources, including Scopus, IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar. Our findings emphasize the critical role of artificial intelligence, machine learning, and deep learning in enhancing CTI models by optimizing data collection, preprocessing, and multilingual support for event extraction. Despite significant improvements in event extraction accuracy and processing efficiency through AI-driven techniques, challenges remain in scaling automated systems and expanding language coverage. This review highlights the need for comprehensive, scalable frameworks that minimize manual effort while improving cross-lingual capabilities, ultimately enabling more robust, timely, and accurate threat intelligence extraction in the evolving cyber threat landscape.

Keywords: Cyber threat intelligence, Multilingual support, Data feeds, Event extraction, Text preprocessing

1. Introduction

Cyber threats have become more advanced and global, making the ability to handle and process CTI feeds increasingly complicated [1]. This is particularly due to the multilingual and heterogeneous nature of input data, which presents challenges in efficiently collecting, preparing, and extracting reliable threat events from these sources [2,3]. CTI enables enterprises to predict and mitigate new risks while functioning as an essential approach within the broad cybersecurity domain [4].

The rapid development of cyber threats in digital infrastructures requires strong CTI systems that can handle extensive, diverse, and multilingual data streams. Despite recent improvements in AI, machine learning, and deep learning methodologies, modern CTI systems experience solid obstacles in preprocessing and extracting critical risk events from multilingual sources [5]. CTI models regularly encounter challenges due to the inherent heterogeneity of unstructured input text, which originates from diverse sources and is presented in various forms, including plain text. Fixed structures and data quality standards impede automated text analysis and event extraction methodologies. Often, these feeds include confusing terms, inconsistent metadata, and irrelevant noise, which can result in misclassification or total neglect of essential threat indicators. The varied sources of these feeds

Received 19 March 2025; revised 24 June 2025; accepted 28 June 2025.
Available online 25 July 2025

* Corresponding author at: School of Computing, Universiti Utara Malaysia, Sintok, Malaysia.
E-mail address: jhyondon@gmail.com (J.H. Al-Yasiri).

https://doi.org/10.33640/2405-609X.3421

complicate the collection phase in CTI models [6]. Recent methodologies relied heavily on manual data labeling and quality assurance procedures, creating scaling challenges that impeded prompt threat identification and action, ultimately slowing the process [7].

Although there has been notable advancement in merging conventional NLP approaches with modern deep learning models, a robust framework that effectively integrates these techniques to address the complexity of CTI data comprehensively remains incomplete. This affects performance in event extraction, where threat narratives and context-specific features are frequently overlooked [8]. Current techniques emphasize cross-lingual input text support; however, some lack details about supported languages, especially low-resource ones. This constraint undermines the efficacy of CTI systems in areas where non-English data predominates [9].

This review systematically analyzes the present situation of multilingual CTI systems, emphasizing three fundamental aspects. These aspects are data collection methods, preprocessing of multilingual textual inputs, and utilizing AI techniques for event extraction.

2. Methods

This review paper adheres to the PRISMA guidelines. It examines the newly published papers since 2020, avoiding outdated research in five databases: Scopus, IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar. For each database, we followed its search rules and regulations, writing the proper search query. We investigated three primary keywords: “Cyber threat intelligence,” “Multilingual,” and “Data feeds.” Alternative keywords were incorporated into the search parameters to broaden the search results and obtain a diverse set of linked papers. In this systematic literature review, several unrelated studies have been excluded because their focus does not align with cyber threat intelligence [10–13]. Furthermore, systematic literature reviews [14–24] and interview papers [25–28] have been excluded. Fig. 1 demonstrates PRISMA steps and results.

3. Results and discussion

3.1. Data sources and datasets

Data have been processed in the relevant literature after being collected from various sources. This systematic review categorizes data based on the sources and the datasets.

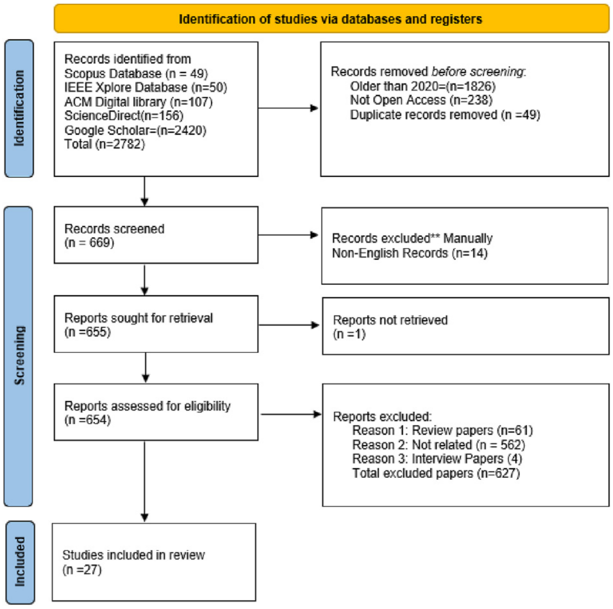
Identification of studies via databases and registers
Identification
Records identified from
Scopus Database (n = 49)
IEEE Xplore Database (n=50)
ACM Digital library (n=107)
ScienceDirect(n=156)
Google Scholar=(n=2420)
Total (n=2782)
Records removed *before screening*:
Older than 2020=(n=1826)
Not Open Access (n=238)
Duplicate records removed (n =49)
Screening
Records screened
(n = 669)
Records excluded** Manually
Non-English Records (n=14)
Reports sought for retrieval
(n =655)
Reports not retrieved
(n =1)
Reports assessed for eligibility
(n =654)
Reports excluded:
Reason 1: Review papers (n=61)
Reason 2: Not related (n = 562)
Reason 3: Interview Papers (4)
Total excluded papers (n=627)
Included
Studies included in review
(n =27)

Fig. 1. PRISMA flow diagram.

3.1.1. Data sources

The related papers’ data sources could be classified into social media posts, CTI feeds, and deep web hacker forums and markets. Social media platforms have been utilized as sources of unstructured context for extracting intelligence-related information. For example, Facebook has been identified as a potential data source [29], while Twitter posts have been aggregated in many studies [9,30–36]. With respect to CTI feeds, the China National Vulnerability Database (CNNVD) has been used in studies [37, 38]. The Open Threat Exchange (OTX) platform was conducted by [7]. NIST CVE has been examined by [9,30]]. In 2024, Ji and his team preprocessed cybersecurity incident reports [39]. Security reports from several sources, including Alto Networks, Trend Micro, Fortinet, and Kaspersky, have been critiqued in several research studies [33–36,40]. In contrast, deep web hacker forums and markets have been the focus of other investigators [29,41–43].

3.1.2. Datasets

The researchers employed existing datasets or developed their own in the related studies.

3.1.2.1. Existing datasets. DBPedia developed the DBP15k dataset. Such a dataset contains three multilingual knowledge graphs: Chinese-English, Japanese-English, and French-English, with 15,000 interlingual links (ILLs) [38]. The Exploit Database was developed by Offensive Security (OffSec) as a non-commercial project. It provides open-access and CVE-compliant records, as well as proof-ofconcept findings by security investigators and penetration testers. These records include the publication date, description, and type of exploits, as well as the target platform and network ports [7].

The Global Database of Events, Language, and Tone (GDELT) is a project supported by Google Jigsaw. It monitors international news across various mediums, including broadcast, online, and print, in over 100 languages. It systematically identifies individuals, digital platforms, organizational entities, thematic trends, news outlets, emotional responses, quantitative metrics, citations, visual content, and incidents that collectively influence global society. Thus, it creates a free-access platform for global computing [7].

The Dark Web Markets Data is available at the AZSecure hacker assets portal, where four darknet marketplaces (DNMs) were discovered. It comprises 14,865 threat-related links gathered from Russian, English, and Arabic hacker forums. These links were specifically shared on the three forums with the highest number of attachments: Tuts4you, EXElab, Opensc, and Ashiyane [7].

The Common Vulnerabilities and Exposures (CVE) dataset has documented software and firmware vulnerabilities for 18 years, significantly contributing to network security. It was founded in 1999 by MITRE, a non-profit research entity that manages government-sponsored research and development labs. CVE offers a public-access dictionary to assist businesses in identifying and managing security vulnerabilities. CVE improves security awareness and threat mitigation by standardizing vulnerability information [30,44,45].

MITRE ATT&CK is a public dataset that catalogs attacker strategies and approaches derived from empirical observations. It provides a basis for developing cyber threat models and methods within the business, government, and cybersecurity sectors, including those related to product and service development [44,45].

The Common Attack Pattern Enumeration and Classification (CAPEC) dataset serves as an extensive resource that facilitates the comprehension of hacker behavior, which is essential to effective cybersecurity. CAPEC offers a systematic repository of recognized attack patterns employed by adversaries to exploit vulnerabilities in cyber-enabled systems. It is a vital resource for analysts, developers, testers, and educators, helping improve community awareness and fortify protection tactics [44,45].

The Common Weakness Enumeration (CWE) is a community-driven repository of vulnerabilities, including relevant conditions in software, firmware, hardware, or service components that, under specific circumstances, may lead to risks. CWE systematically identifies and describes these weaknesses to aid in security analysis and mitigation efforts [44,45].

The Spam Hunter dataset was developed by aggregating tweets with SMS-related keywords, performing picture analysis, and extracting URLs associated with phishing. It was used to detect possible threats [31]. The Twitter IOC Hunter dataset contains cybersecurity-related information, such as malicious URLs and IP addresses, from X platform (formerly Twitter). Data were retrieved by leveraging the platform's API over a specified timeframe, facilitating threat intelligence analysis [31].

Table 1 demonstrates the adoption matrix of the datasets for the reviewed papers.

3.1.2.2. Created datasets. Table S2 (https://kijoms.uokerbala.edu.iq/cgi/viewcontent.cgi?filename=0&article=3421&context=home&type=additional) illustrates the details of the datasets created. Reference [41] presented a hacker forum dataset comprising data obtained from eleven hacker forums, with semi-structured data stored in a MySQL database. Then, the author generated a gold-standard dataset containing 5210 manually labeled records derived from his first dataset. Reference [38] utilized a semi-structured knowledge graph dataset (BT4K) comprising 14,866 entities sourced from the CNNVD, accompanied by manual review and confirmation processes.

On the other hand, [7] acquired 18,000 entries from the OTX and saved them in HTML format. Meanwhile, [9] developed three datasets using JSON as a primary storage format. Two datasets were derived from a combination of the NIST CVE and the X platform, while the third dataset was sourced from the X platform. Partial manual labeling was performed with the first two datasets, whereas the third was manually validated. Another study conducted by [42] generated two datasets; the first was compiled from four hacker forums and contained 339,821 unstructured preprocessed records. Subsequently, a gold-standard dataset was derived from the initial one using human labeling. Moreover, [43] scraped data from hacker forums and darknet markets, then created a dataset of 862,715 preprocessed unstructured records.

Later, a gold-standard dataset was manually annotated and derived from the initial dataset. A dataset of 10,000 unstructured raw records was collected from cybersecurity incident reports, generating sequential task seeds, a training dataset, and a test dataset, which was evaluated manually [39]. The BVTED dataset is a manually

Table 1. Datasets adoption matrix for reviewed papers.

Ref.	DBP15K	Exploit Database	GDELT	Dark Web Markets Data	CVE	MITRE ATT&CK	CAPEC	CWE	SpamHunter	Twitter IOC Hunter	Agora	Silk Road	Reddit	Twitter OSINT
[38]	*													
[7]		*	*	*										
[30]					*									
[31]									*	*				
[44]					*	*	*	*						
[45]					*	*	*	*						
[49]											*	*	*	
[53]														*

The asterisk (*) indicates that the above dataset is used in the corresponding reference.

annotated, semi-structured knowledge graph comprising 27,311 records stored in JSON and CoNLL format [37]. Another study developed a multilingual dataset of 1325 tree-structured processed records [38].

A recent paper established two ground-truth datasets in English and Japanese, sourced from the X platform. Thereafter, two additional datasets were randomly selected and manually annotated from the first two, and a fifth dataset (CrowdCanary) was created, comprising text and images collected from Twitter posts [31]. Another paper presented a semi-structured preprocessed STIX2 format dataset, aggregating data from logs, CTI feeds, and social media [46].

In 2023, Siracusano and his team provided a manually annotated, semi-structured STIX2 dataset containing 36,100 records drawn from 204 CTI reports and 62 security organizations, including Alto Networks, Trend Micro, and Fortinet [40].

Two studies utilized multidimensional cyber threat and social media datasets derived from Kaspersky's threat statistics and the X platform, respectively, stored in Microsoft SQL Server and MS Dataverse [33,34]. Furthermore, another study offered a systematically preprocessed dataset collected from a honeypot and stored in STIX format [47]. Lastly, in a conceptual paper, a group of researchers suggested collecting their first unstructured dataset from Facebook groups, hacker forums, and the CERT-In site, storing it in JSON format, deriving a second preprocessed dataset from the initial one, and storing its semi-structured records in Parquet format [29].

3.2. Multilingual support

Multilingual support indicated either input context support or output context support. The input context support refers to handling multilingual data at the preprocessing stage. This includes techniques that enable models to effectively understand the original multilingual data without translating it first. Output context support, on the other hand, involves producing model outputs directly in multiple languages or generating outputs that maintain semantic consistency across languages. Thus, it ensures that extracted information or insights remain accurate and relevant regardless of the language context.

This systematic literature review focuses on input context support for linguistic assistance [40,45,46]. The related studies mentioned the supported languages in their work as summarized in Table 2.

Table 2. Supported languages referenced in the relevant literature.

Ref.	English	Russian	Chinese	Arabic	French	Japanese	Danish	Dutch	Finnish	German
[41]	*	*								
[38]	*		*		*	*				
[7]	*	*		*						
[9]	*	*			*		*	*	*	*
[30]	*	*			*	*		*		
[43]	*	*			*					
[42]	*	*			*					
[39]	*		*							
[37]	*		*							
[48]	*							*		*
[31]	*					*				
[32]	*					*				
[47]	*	*								*
[50]	*	*								*

Ref.	Hungarian	Italian	Norwegian	Portuguese	Romanian	Spanish	Swedish	Polish	Greek	Basque	Ukrainian	Other
[41]												
[38]												
[7]												
[9]	*	*	*	*	*	*	*					*
[30]		*			*	*		*	*			*
[43]		*										
[42]												
[39]												
[37]												
[48]				*		*				*		
[31]												
[32]												
[47]												
[50]											*	

The asterisk (*) indicates that the above language is supported in the corresponding reference.

3.3. Data collection tools

The related research utilized various automated and manual methods for data collection, aiming to extract relevant information from online sources. Initially, web crawling techniques were utilized extensively, encompassing conventional crawlers and specialized Tor-based variants. Standard web crawlers systematically fetch, parse, and store web content using HTTP requests, HTML parsing, and URL discovery [7,37, 38, 49]. Meanwhile, Tor-routed techniques utilize the Tor network to protect web surfing by directing HTTP requests through multiple encrypted relays. Therefore, it hides the crawler's identity, location, and IP address, which helps secure and invisible data collection from dark web marketplaces and forums that are generally inaccessible to the traditional browsers [41,42].

Furthermore, obfuscated Breadth-First Search (BFS) was applied as a crawling method. This included randomized delays, node selection, and diverse traversal paths to avoid detection [42]. In addition, a hybrid approach was employed that integrates API extraction with the manual collection reported by [40]. In addition, web scraping was utilized [29,50]. One study enhanced the methodology with heuristic rules, incorporating supplementary scraping techniques [35]. The Twitter API was utilized for collecting posts from the X platform (formerly Twitter) [9,30–32,39]. Conversely, the Facebook Graph API was proposed to collect posts from Facebook groups [29]. Microsoft Power Automate was utilized as a cloud-based automation tool to optimize the data collection task without requiring coding [33,34]. In contrast, advanced data management was accomplished through a Big Data Pipeline [47]. In terms of specialized open-source intelligence tools, the integration of Suricata and Spiderfoot was utilized for collecting data online [48], while the Requests tool was suggested to be used [29]. Furthermore, Telethon was employed to scrape text from Telegram groups [50]. Flashpoint and Palantir Gotham have been used to access and collect data from dark web forums [51]. Several studies did not delineate their data collection tools, underscoring possible deficiencies in methodological transparency [45,46,52,53].

3.4. Data preprocessing and processing

Studies have demonstrated the critical importance of text preprocessing, cleaning, and processing stages to extract targeted information. Fundamental text normalization steps typically include lowercasing, removing URLs and emojis, and enforcing UTF-8 encoding. Furthermore, processing steps involve tokenization, lemmatization, stop-word removal, multilingual embedding generation, and feature extraction using various statistical and transformer-based techniques. Although a similar pattern is observed across researchers in data processing, there is still no standard procedure; in this direction, the reviewed articles varied in the steps chosen.

3.4.1. Text normalization

Text normalization is a preprocessing approach that standardizes textual data into a consistent format (e.g., lowercasing and removing extra whitespace and line breaks). Recent research has explored text normalization using a variety of tools and methodologies. Some methodologies incorporate specialized normalization tools such as Colabeler [37], Graphene [46], NLTK library [44,45], and RegEx as fundamental elements of text standardization processes [29,31,32,45]. In addition, they include Pandas in conjunction with Unicode libraries, for systematic preprocessing [29]; TF-IDF normalization, for the adjustment of word frequencies [47]; and sliding window methodologies, to manage extensive text sequences [48]. Other investigators utilized an iterative summarizing methodology employing LLM-based approaches, such as GPT-3.5, to enhance textual material progressively [40].

Normalization was integrated with translation and language detection processes using AI-based translation frameworks, including Microsoft Power Automate and Microsoft Azure Cognitive Services, as well as the Microsoft Cognitive Services API. These were applied alongside conventional methods, such as Porter Stemming, to improve linguistic consistency [33–36].

Furthermore, normalization techniques have been enhanced by incorporating pre-trained language models, such as BERT, with attribute embedding [38]. A research group examined the process of normalizing context using heuristic rule-based solutions for duplication and quality filtering [39]. Essential preprocessing procedures, such as lowercasing and UTF-8 encoding standardization, have been uniformly implemented as in [7,42,43], whereas other research utilized extensive fundamental preprocessing methods that eliminate emojis, URLs, diacritics, and extraneous Unicode characters [30].

Another paper applied specified procedures, like lowercasing, padding, and the filtration of non-alphanumeric characters [41]. On the other hand, instead of using the traditional normalization process, BERTopic was used, which did not require text processing [49]. A group of researchers did not disclose their text normalization techniques, underscoring a possible gap in the methodology used [9,52].

3.4.2. Tokenization

Tokenization is a crucial preprocessing step in natural language processing, which entails dividing the text into significant pieces. The literature demonstrated various methods, which were adapted to distinct research requirements. Reference [7] applied transformer-based tokenizers to capture contextual subtleties, whilst [48] utilized XLM-RoBERTa for multilingual tokenization. Traditional NLP libraries are essential tools, such as NLTK [44]; SpaCy [29,39]; and Jieba, which was used for Chinese segmentation to tackle language-specific issues [39]. Sentence-BERT (SBERT) was used to tokenize forum posts [49]. ByteLevelBPE tokenizer was utilized as a part of the SecurityBert pipeline [52].

A group of studies conducted tokenization implicitly during their preprocessing phases without specifying a particular tool [9,34–37,40–43,45]. Finally, a set of studies avoided conventional tokenization either by employing sentence embedding methodologies [30], or by skipping the tokenization process totally [31–33,38,46,47].

3.4.3. Lemmatization

Lemmatization is an essential text preprocessing method in natural language processing that converts words to their standard forms, improving semantic analysis and feature extraction. Numerous studies have explicitly utilized specific NLP libraries for lemmatization. For instance, NLTK was employed for its comprehensive linguistic resources [9,44], while SpaCy was adopted due to its rapid processing pipeline [29]. Conversely, certain research highlighted lemmatization in its preprocessing phase without identifying a specific tool [41,44]. A set of studies did not conduct lemmatization as a separate process or name a tool to perform that [7,30–40,42,43,46,47]. The reported methodologies highlighted a dependency on well-known technologies, such as NLTK and SpaCy, and inconsistent documentation standards among various studies.

3.4.4. Stop-word removal

Stop-word removal is a phase in natural language processing that seeks to eliminate common words that generally convey no semantic significance, thereby facilitating subsequent analytical activities. Our study revealed a few papers that explicitly detailed their stop-word removal methodology utilizing specific tools. The Natural Language Toolkit (NLTK) was used to eliminate stop-words, employing its comprehensive, predefined stop-word lists to improve text clarity [29]. The StopwordsISO tool was utilized as a standardized source of stop-words for several languages to eliminate non-informative words effectively [9,30].

However, most of the examined studies did not identify the stop-word removal process or name a tool for that [7,31–37,39–48]. This discrepancy in reporting indicates that, although stop-word removal is typically a fundamental aspect of text preprocessing, its explicit acknowledgment may vary based on the study's emphasis or the presumed knowledge of the audience regarding conventional NLP methodologies.

3.4.5. Multilingual embedding generation

The latest developments in multilingual embedding have employed diverse techniques. Transformer-based models, such as BERT, have been utilized to generate contextual representations in cross-lingual text [7,31,32]. Graph-based methodologies have emerged, employing Graph Convolutional Networks (GCN) and enhanced TransE models in combination with pre-trained language models to help understand structural relationships in multilingual datasets [38]. Traditional embedding techniques were exemplified by the use of GloVe [41], while cross-lingual sentence embeddings have been successfully generated using the LASER framework [9,30]. The field has further advanced with transformer-based architectures, such as MBERT and XLM, which have been employed independently [32] or alongside complementary models like ERNIE and Word2Vec [37].

A collection of Microsoft cognitive tools, including Cognitive Service Agent, Cognitive Services, Power Automate, and Text Analytics, was utilized to generate and enhance multilingual embeddings [33–36]. Furthermore, contemporary models have augmented the existing toolset. These include MiniLM [44], multi-qa-mpnet-base-dot-v1 with the MPNET Sentence Transformer [45], Sentence-Transformers available via Hugging Face [39], and OpenAI's text-embedding-ada-002 [40].

Further strategies included traditional NLP-based categorization [46] and NLP-based translation combined with embedding models [47]. Additionally, innovative approaches integrated Long Short-Term Memory (LSTM) networks with Generative Adversarial Networks (GANs) to generate language-invariant embeddings [42,43]]. On the other hand, the Sentence-Transformers library was used to embed text from Russian, English, Ukrainian, and German. [50]. The utilization of XLM-RoBERTa, in its standard and fine-tuned forms, was emphasized in ongoing efforts to achieve effective multilingual representations [29].

3.4.6. Feature extraction

The literature on feature extraction strategies encompassed both conventional statistical methods and advanced neural and transformer-based approaches. Statistical methodologies include TF-IDF [35,44]; Latent Semantic Indexing (LSI) [44]; Porter stemming with n-gram [34–36]; truncated SVD [31,32]; vectorization, and statistical analysis [47]; and t-SNE for dimensionality reduction [48]. Transformer-based models employed XLM-RoBERTa [29], BERT embeddings [31,32,37,44], MPNET Transformer [45], and GPT-4 [39]. Domain-adaptive feature extraction was performed by DTL-EL [41] and Enhanced TransE, which utilized a pre-trained language model [38].

Neural architectures encompassed GCN [38]; MLP [7]; BiLSTM [43]; convolutional feature maps [37]; and user-embedding methodologies, including LASER and User2Vec [9,30]. Generative Adversarial Networks (GANs) were facilitated for feature augmentation [42,43]. In contrast, big data analytics and natural language processing (NLP) were utilized for contextual extraction [46], along with STIX representation [40], expanding the horizons of feature engineering in cybersecurity domains. Finally, SecurityBERT was employed for feature extraction, converting tokenized traffic into contextual embeddings [52].

3.5. Data analysis and event extraction

3.5.1. Named entity recognition (NER)

Named entity recognition (NER) comprises many approaches that address multiple domains and languages. SpaCy, which recognized for its efficient pipeline, has been utilized for tokenization, part-of-speech tagging, and entity recognition [29,41]. Simultaneously, researchers used large language models—such as GPT-4— for sophisticated NER tasks, exploiting their generative capabilities [39]. Other significant methodologies—including BERT-based architectures—were frequently developed by combining diverse embeddings—such as Word2Vec, BERT, ERNIE, or mBERT—with either CRF or BiLSTM + CRF to enhance detection precision across multilingual datasets [37].

The Polyglot and Stanford libraries were employed to extract geographical data from tweets, demonstrating their efficacy in processing concise social media text [9]. The Graphene tool was introduced as a multifaceted platform for entity extraction linked to cybersecurity, facilitating domain-specific modifications [46]. In the domain of temporal entity recognition, frameworks such as LADDER and aCTIon illustrated how specialized models can efficiently identify time-related expressions [40]. In contrast, the NER function was embedded directly within LLM chatbots [53]. Finally, AI-based NER was suggested as a comprehensive solution to address multilingual and cross-domain challenges, emphasizing the need for adaptable and scalable approaches [33].

3.5.2. Word association analysis

Recent research has utilized various techniques and methodologies to analyze word associations. For instance, several studies employed frequency-based methods—such as TF-IDF—and subsequently combined them with n-gram models on numerous occasions [34–36]. Chain-of-Thought (CoT) was employed to enhance interpretability, proving the significance of advanced language modeling [39].

Meanwhile, MBERT and ERINE were integrated for transformer-based embeddings, marking a shift towards contextual comprehension [37]. Probabilistic methodologies were employed as in [31], while PPMI was utilized to highlight the statistical significance of word co-occurrences [32]. Additionally, Cosine Similarity served as a measurement for semantic relationships [44,45], while Graphene explored graph-based representations [46].

Additionally, text-embedding-ada-002 was used to investigate temporal dimensions in word associations [40]. A Class-based C-TF-IDF was applied to quantify word association analysis within each topic cluster [49]. Many studies did not identify any particular tool or method, thus indicating a more generalized or ambiguous approach to word association analysis [7,9,29,30,33,38,41–43,47,48].

3.5.3. Topic modeling

Traditional statistical modeling, such as Latent Dirichlet Allocation (LDA), demonstrated efficacy in theme extraction from multilingual datasets and various cybersecurity contexts [34–36,40,48]. Moreover, the MITRE ATT&CK categorization was incorporated with LDA, resulting in a hybrid technique that connects cybersecurity risks with established adversarial frameworks [40]. An alternate strategy was utilized an NLP-based categorization method to organize cybersecurity issues [46]. BERTopic was applied to extract interpretable topics from the corpus [49,50]. The remaining work failed to provide a subject modeling tool or render the methodology implicitly [7,9,29–32,37–39,41–45,47]. This highlights the need to elucidate and standardize methodological practices in prospective research.

3.5.4. Sequence classification

Many methodologies and tools focus on addressing the complicated processes that come with sequential data, resulting in dramatic development in sequence classification. In this term, Recurrent Neural Network (RNN) based architectures, such as BiGRU [29] and BiLSTM [41–43], demonstrated efficiency in modeling time-based dependencies; furthermore, integrating LSTM with FCNN and CNN showed further enhanced feature extraction [47]. Graph-based approaches, as shown by the GCN-TransE hybrid model [38], GAN-BERT, and ConGAN-BERT [7], provide creative approaches for identifying relationships among data, while feed-forward architectures—such as a 5-layer feed-forward neural network (FFNN) [9] and traditional FFNNs [30]—illustrate the use of deep networks in classification tasks. Simultaneously, advanced embedding techniques—such as SEvenLLM with Llama2 and Qwen [39] and text-embedding-ada-002 [40]—underscore the shift towards robust semantic representation.

Reference [48] employed a multilingual Bi-GRU, whereas [37] augmented the toolbox by incorporating PCNN, BiLSTM + ERNIE, CasRel, and OneRel, offering multiple contextual and relational learning tiers. Alongside these deep learning methodologies, traditional machine learning techniques—such as random forest, neural networks, decision trees, support vector machines, and logistic regression—provided reliable baselines [31,32]]. Meanwhile, approaches like cosine similarity calculation and pattern recognition engines were employed to detect specific sequence patterns [45,46]. Finally, cloud-based analytics and automation technologies—such as the Microsoft Text Analytics API [35] and Microsoft Power Automate [33,34]—showed powerful techniques in sequence categorization, and LLM-based chatbots incorporated embedded sequence classification functionality [53]; however, several studies did not specify the applied methodologies [36,44].

3.5.5. Prediction and event extraction

Various strategies have been utilized for event extraction and prediction. For example, CRF was employed for sequence labeling to enhance predictive accuracy [29]. At the same time, BiLSTM and softmax classifiers were used to handle structured outputs in text classification [41]. Numerous studies have investigated neural network–based methodologies, including the GCN-TransE hybrid model, for blockchain data [38]; GAN-BERT and Con-GAN-BERT, for augmented feature generation [7]; and User2Vec, for individualized predictions [9,30]]. Other investigations presented specialized architectures such as the SEVEN-LLM fine-tuned model [39], the attention-based pointer network with a bi-affine classifier [48], and security intelligence services [44].

Furthermore, the Peracton MAARS security analytics engine was implemented for cybersecurity forecasting [46], exponential smoothing was incorporated for time-series analysis [34,35], and Microsoft Power BI facilitated data visualization [33]. Deep learning architectures—including FCNN, CNN, and LSTM—integrated with SIEM, were conducted in this domain [47]. Finally, SecurityBERT performed prediction by labeling each network flow as either normal or belonging to one of the anomalies or attack classes [52].

3.5.6. Evaluation metrics

Evaluation sections of pertinent experimental studies demonstrated a significant reliance on classification metrics to validate proposed models and benchmark them against state-of-the-art approaches. Results varied notably depending on the datasets employed and methodologies applied, as illustrated in Table 3, which presents a comparative summary of achieved values for evaluation metrics across selected studies. Despite these variations, a clear and consistent pattern has emerged, showing that accuracy, F1-score, precision, and recall are the primary evaluation metrics used across CTI studies. These metrics provide comprehensive insights into model performance and enable robust comparisons and reliable validation in multilingual Cyber Threat Intelligence (CTI) event extraction research.

3.6. Limitations in current CTI approaches

A comparative overview of text-processing and analysis steps, as well as the tools adopted across CTI studies, was conducted, as detailed in Subsections 3.3, 3.4, and 3.5, and summarized in Table 4.

Significant limitations emerged regarding the clarity and comprehensiveness of descriptions provided by existing methodologies. Studies have been categorized based on their approach to describing processes and tools as follows:

1. Explicitly described the process and identified the tool used.
2. Partially or implicitly referenced the tool, which perform the task, without providing a detailed description.

Table 3. Comparative evaluation metrics: Achieved values for selected studies.

Ref.	Classification metrics										Ranking metrics			NLP/sentiment analysis metrics				
	Accuracy (ACC)	Area under the curve (AUC)	F1-Score	Precision (PPV)	Recall (TPR)	Specificity (SPC/TNR)	False negative rate (FNR)	False positive rate (FPR)	Matthew's correlation coefficient (MCC)	Negative predictive value (NPV)	Hits@1	Hits@10	MRR	Avg. positive sentiment	Avg. neutral sentiment	Avg. negative sentiment	Jaccard similarity	Rouge-L
[41]	0.77		0.80	0.82	0.77													
[38]											0.74	0.86	0.81					
[7]	0.85		0.84	0.85														
[9]	0.87	0.87	0.86	0.86	0.87				0.73									
[30]	0.87	0.87	0.86	0.86	0.87				0.73									
[43]	0.97	0.99	0.96															
[42]	0.76	0.77	0.70	0.68	0.76													
[37]			0.95	0.95	0.95													
[48]			0.89															
[31]	0.96		0.96	0.96	0.95	0.96												
[32]	0.96		0.96	0.96	0.95	0.96												
[45]	0.61		0.85	0.86	0.83												0.40	
[40]			0.80	0.78	0.84													
[33]														0.21	0.43	0.36		
[34]	0.83		0.88	0.91	0.85	0.78	0.15	0.22		0.65				0.22	0.42	0.36		
[35]														0.22	0.42	0.36		
[36]														0.22	0.42	0.36		
[53]			0.95	0.97	0.93													

Table 4. Comparative overview of text processing steps and tool adoption in CTI studies.

Ref.	Data collection	Text normalization	Tokenization	Lemmatization	Stop-word removal	Multilingual embedding generation	Feature extraction	Named entity recognition	Word association analysis	Topic modelling	Sequence classification	Prediction	Anomaly detection	Used manual review	Used translation
[48]	D	G	E			E	E			E	E	E		Y	Y
[29]	E	E	E	E	E	E	E	E			E	E		N	Y
[30]	E	D			E	E	E		D		E	G		Y	N
[7]	G	G	C				E				E	E		Y	N
[9]	E	D	D	D	E	E	E	E			E	E		Y	N
[31]	E	D	C			E	E	E	G		E	E		Y	N
[32]	E	E	C			E	E		E		E	E		Y	N
[33]	E	E						E			E	E	E	N	Y
[34]	E	E	D				G		G	G	E	G	G	N	Y
[36]	E	D	D		D		E		E	E				N	Y
[35]	G		D				G		G	E	E	E	E	N	Y
[37]	D		C			E	E	E			E	E		Y	N
[38]	C		C	C		E	E					G		Y	N
[47]	E	E	D	D	D	D	G	D			E	E		Y	Y
[39]	G	G	E			E	C	E	C		E	E		Y	N
[46]	E	E	E	E	E		C	E	E		C	C	E	Y	Y
[40]	D	E	D			E	E	E	D		G			Y	N
[41]	D	G	D	G		E	G	E			E	E		Y	N
[42]	E	D	D		D	E	E				E	D		Y	Y
[43]	E	G	D			E	E				E	G		Y	N
[44]	D	E	E	E	E	E	G		G			G		N	N
[45]	D	D	D	D	D	E	E		G		G			Y	N
[49]	D	D	C		D		D		E	E				N	N
[50]	E		C			E	C			E				Y	Y
[51]	E						G					C	G	Y	Y
[52]	D	D	E			G	E				E	E	E	N	Y
[53]	D	D						E			E			Y	N

E = Paper explicitly names a specific tool.
C = Paper cites a tool that performs part of the process or undertakes the process implicitly.
D = Paper describes the process but does not name any tool.
G = Paper mentions only a generic technique without citing a specific tool.
− = Process/tool not mentioned.
Y = Acknowledged.
N = Not acknowledged.

3. Provided a detailed explanation of the procedure without naming an associated tool.
4. Referenced the technique generally, without specifying a particular tool.
5. Omitted both the process and the used tool entirely.

This classification highlights substantial deficiencies—in documenting procedural fundamentals clearly and identifying utilized software explicitly—that complicating reproducibility and comparative evaluation.

Although, most of the related papers specified the supported languages; however, some studies addressed the multilingual context support without clearly specifying which languages they accommodate [40,45,46]]. Reference [29] broadly classified languages as simply: English and non-English, without further specification. Several researcher merely indicated the number of the supported languages without providing detailed descriptions [33–36,44]]. This imprecise specification of supported languages introduces an additional limitation, risking misinterpretation and undermining reliable benchmarking.

On the other hand, several experiments have shown a frequent reliance on manual evaluation, as indicated in Table 3; and manual labeling, as elaborated in Subsection 3.1.2.2 and summarized in Table S2 (https://kijoms.uokerbala.edu.iq/cgi/viewcontent.cgi?filename=0&article=3421&context=home&type=additional). This approach demands considerable and high-cost professional human resources, consumes more time, and restricts scalability. Moreover, it significantly limits the efficiency and applicability of CTI methodologies, practically considering the large and continuously expanding datasets in cybersecurity.

Furthermore, the use of translation in CTI processes identifies extra limitations. Table S2 (https://kijoms.uokerbala.edu.iq/cgi/viewcontent.cgi?filename=0&article=3421&context=home&type=additional) illustrates the frequent translation processes during dataset creation, while Table 4 highlights instances of translation directly integrated into text-processing workflows. These practices introduce semantic inaccuracies and inconsistencies, potentially undermining the precision and reliability of subsequent analyses and derived threat intelligence insights.

Table 2 identifies another critical limitation—inadequate linguistic processing support for under-resourced languages, particularly Arabic. Despite Arabic's geopolitical and cybersecurity relevance, its specific processing requirements are largely overlooked in current methodologies, restricting the effectiveness and global reach of multilingual CTI systems.

Addressing these limitations is essential to enhance the robustness, scalability, and international applicability of future CTI systems.

3.7. Summary of findings

The research underscores the critical importance of high-quality, diverse, and up-to-date data resources for effective CTI systems, demonstrating that precision directly depends on data quality, variety, and timeliness, as detailed in Subsections 3.1.1 and 3.1.2. Consequently, a substantial demand persists in developing robust data pipelines that can continuously ingest and validate new CTI datasets.

Regarding advancements in AI-driven preprocessing and event extraction, integrating multiple AI techniques—for instance, transformer-based embeddings were combined with RNN-based and traditional machine learning methods [34–36]— has shown high performance enhancement, as illustrated in Table 4 and Table S6 (https://kijoms.uokerbala.edu.iq/cgi/viewcontent.cgi?filename=0&article=3421&context=home&type=additional).

The literature emphasized that no single AI method can alone adequately capture the complexity of CTI; therefore, future solutions must incorporate diverse AI paradigms in multilingual support.

The use of multilingual embeddings—for instance, XLM-RoBERTa [9], BERT [7], LASER [42,30]—has improved the handling of linguistic subtleties, but these methods still face limitations with less frequently represented languages.

3.8. Future work recommendations

Future research—in Cyber Threat Intelligence (CTI)—should prioritize enhancing multilingual capabilities—particularly for underrepresented languages, such as Arabic—and developing real-time detection frameworks responsive to emerging threats. Automated data annotation methods need to be standardized to reduce manual processes. Integrating advanced AI and large language models (LLMs) is recommended to improve event extraction accuracy. Finally, establishing robust quantitative metrics and privacy-preserving techniques ensures the scalability and adaptability of CTI systems.

3.9. Proposed conceptual framework

This systematic literature review informs the design of a conceptual framework, which is proposed to overcome existing state-of-the-art limitations—including dependency on manual annotation, reliance on translating original texts into English before processing, and the lack of models customized for processing Arabic text. Fig. 2 demonstrates the proposed conceptual framework's block diagram. The framework integrates cross-lingual processing and contextual understanding, addressing the inherent complexities of diverse languages and threat formats. It suggests employing a combination of (XLM-RoBERTa + Bi-GRU + CRF), aiming to create scalable and adaptive Cyber Threat Intelligence (CTI) systems that are capable of efficiently extracting relevant cyber threat events from multilingual data sources. Data collection involves reading data from diverse sources (Facebook Groups, Hack Forums, Saudi-CERT) and storing the collected raw data in a JSON storage. Language detection utilizes Polyglot for automated language categorization. Subsequently, text normalization applies Unicode normalization, whitespace normalization, and line break normalization. Additionally, custom language cleaning using Regex—which identifies indicators of compromise (IoC) words—preserves important cybersecurity terms, which are then directly tokenized using SentencePiece tokenizer from XLM-RoBERTa. Next, preprocessing steps include stop-word removal using the Stopwords ISO library and language-specific lemmatization tailored to each language. The lemmatization process bypasses recognized CTI terms to maintain the contextual integrity.

Furthermore, the customized NER generates contextual embeddings using XLM-RoBERTa and encodes sequential context employing a BiGRU layer. Token label scores are then predicted using a linear layer combined with a Conditional Random Field (CRF). The optimal label sequence is then decoded via CRF decoding. Subsequently, entity spans are post-processed and grouped, preparing for structured event extraction.

The CTI event extraction process begins by identifying event triggers using a fine-tuned XLM-R classifier, then extracting contextual arguments by applying Bi-GRU, and finally, role labeling and template matching employing the CRF classifier. The extracted structured events are stored in the final data storage.

Model training and fine-tuning involve preparing a multilingual dataset, which is automatically annotated with event-role BIO tags. The process includes fine-tuning the XLM-RoBERTa model, for token classification; training a BiGRU encoder, to capture sequential contextual embeddings; and optimizing a CRF decoder, to ensure coherent tagging sequences. Performance evaluation is conducted, employing accuracy, precision, recall, and F1 scores to benchmark the framework's

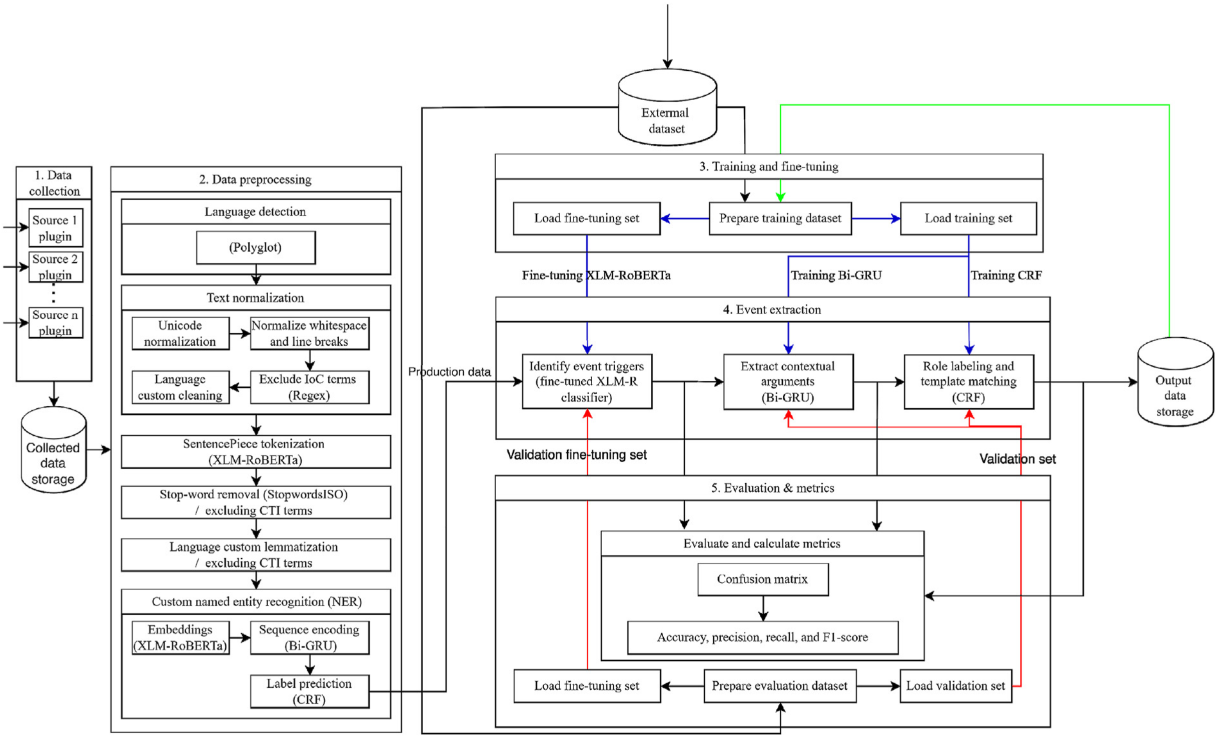
External dataset
1. Data collection
Source 1 plugin
Source 2 plugin
Source n plugin
Collected data storage
2. Data preprocessing
Language detection
(Polyglot)
Text normalization
Unicode normalization
Normalize whitespace and line breaks
Language custom cleaning
Exclude IoC terms (Regex)
SentencePiece tokenization (XLM-RoBERTa)
Stop-word removal (StopwordsISO) / excluding CTI terms
Language custom lemmatization / excluding CTI terms
Custom named entity recognition (NER)
Embeddings (XLM-RoBERTa)
Sequence encoding (Bi-GRU)
Label prediction (CRF)
Production data
3. Training and fine-tuning
Load fine-tuning set
Prepare training dataset
Load training set
Fine-tuning XLM-RoBERTa
Training Bi-GRU
Training CRF
4. Event extraction
Identify event triggers (fine-tuned XLM-R classifier)
Extract contextual arguments (Bi-GRU)
Role labeling and template matching (CRF)
Output data storage
Validation fine-tuning set
Validation set
5. Evaluation & metrics
Evaluate and calculate metrics
Confusion matrix
Accuracy, precision, recall, and F1-score
Load fine-tuning set
Prepare evaluation dataset
Load validation set

Fig. 2. Proposed framework block diagram.

effectiveness against existing state-of-the-art methods and validating improvements in multilingual CTI event extraction.

In the authors' future endeavors, a practical prototype will be implemented and assessed, substantiating the proposed conceptual framework. This prototype will implement each key component—including data collection, multilingual preprocessing, and event extraction—to evaluate practical feasibility, performance, and scalability in a real-world CTI environment.

4. Conclusion

This extensive literature review indicated that the development of data collection, preprocessing, and event extraction approaches has dramatically influenced the evolution of cyber threat intelligence (CTI).

The diversity and quantity of data sources, particularly those utilizing transformer-based models and hybrid architectures—have significantly enhanced the accuracy and effectiveness of CTI event extraction. Challenges remain in achieving comprehensive multilingual support and reducing reliance on human annotation. To address these limitations, this review proposes a conceptual framework aimed at overcoming current state-of-the-art constraints in CTI methodologies. This comprehensive literature review indicated that future research should focus on expanding the linguistic range of CTI systems by incorporating tailored modules for underrepresented languages. This review suggested that the integration of cutting-edge AI approaches can substantially enhance the standardization of data pipelines. The automation of annotation processes will be crucial for maintaining data quality and timeliness in an ever-evolving threat landscape. These advancements are essential for creating more resilient, flexible, and proactive CTI systems that can tackle the complexities of modern cybersecurity issues.

Ethical information

This systematic literature review study did not conduct any human or animal experimentation; therefore, ethical approval is unnecessary.

Funding

No financial support was received.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgments

The authors thank Mr. Mouhammed Jessam Lafta and highly appreciate his valuable effort in revising and proofreading.

References

[1] Y. Keim, A.K. Mohapatra, Cyber threat intelligence framework using advanced malware forensics, Int. J. Inf. Technol. 14 (2022) 521–530, https://doi.org/10.1007/s41870-019-00280-3.

[2] R. Gruetzemacher, D. Paradice, Deep transfer learning & beyond: transformer language models in information systems research, ACM Comput. Surv. 54 (2022) 1–24, https://doi.org/10.1145/3505245.

[3] E.M. Abdelzaher, Lexicon-based detection of violence on social media, Cogn. Semant. 5 (2019) 32–69, https://doi.org/10.1163/23526416-00501002.

[4] H. Griffioen, T. Booij, C. Doerr, Quality evaluation of cyber threat intelligence feeds, in: M. Conti, J. Zhou, E. Casalicchio, A. Spognardi, eds., Applied Cryptography and Network Security (ACNS), Springer, Cham. 2020, pp. 277–296, https://doi.org/10.1007/978-3-030-57878-7_14.

[5] J. Thom, Y. Shah, S. Sengupta, Correlation of cyber threat intelligence data across global honeypots, in: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, NV. 2021, pp. 766–772, https://doi.org/10.1109/CCWC51732.2021.9376038.

[6] X. Wang, R. Chen, B. Song, J. An, J. Jiang, J. Wang, P. Yang, Learning cyber threat intelligence knowledge graph embedding with heterogeneous relation networks based on multi-head relational graph attention, in: 2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta), IEEE, Haikou. 2022, pp. 1796–1803, https://doi.org/10.1109/SmartWorld-UIC-ATC-ScalCom-DigitalTwin-PriComp-Metaverse56740.2022.00256.

[7] O. Cherqi, Y. Moukafih, M. Ghogho, H. Benbrahim, Enhancing cyber threat identification in open-source intelligence feeds through an improved semi-supervised generative adversarial learning approach with contrastive learning, IEEE Access 11 (2023) 84440–84452, https://doi.org/10.1109/ACCESS.2023.3299604.

[8] R. Alguliyev, B. Nabiyev, K. Dashdamirova, CTI challenges and perspectives as a comprehensive approach to cyber resilience, in: 2023 5th International Conference on Problems of Cybernetics and Informatics (PCI), IEEE, Baku. 2023, pp. 1–5, https://doi.org/10.1109/PCI60110.2023.10325971.

[9] A. Cotov, Improving Cybersecurity Awareness: Tweet Classification Using Multilingual Sentence Embeddings and Contextual Features, The Polytechnic University of Milan. 2022. Master's thesis.

[10] B. Ampel, C. Yang, J. Hu, H. Chen, Large language models for conducting advanced text analytics information systems research, ACM Trans. Manag. Inf. Syst. 16 (2024) 1–27, https://doi.org/10.1145/3682069.

[11] C.T. Duong, D.P. David, L. Dolamic, A. Mermoud, V. Lenders, K. Aberer, From scattered sources to comprehensive technology landscape: a recommendation-based retrieval approach, World Pat. Inf. 73 (2023) 1–10, https://doi.org/10.1016/j.wpi.2023.102198.

[12] H.T. Tran, H.H.P. Vo, S.T. Luu, Predicting job titles from job descriptions with multi-label text classification, in: 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), IEEE, Hanoi. 2021, pp. 513–518, https://doi.org/10.1109/NICS54270.2021.9701541.

[13] I. Machado, W. Assuncao, R. Souza, Combinatorial Interaction Testing Tools for Software Product Lines Engineering: A

Comparative Analysis, Master's Thesis, Federal University of Bahia. 2020.
[14] C. Bratsas, E.K. Anastasiadis, A.K. Angelidis, L. Ioannidis, R. Kotsakis, S. Ougiaroglou, Knowledge graphs and semantic web tools in cyber threat intelligence: a systematic literature review, J. Cybersecur. Priv. 4 (2024) 518—545, https://doi.org/10.3390/jcp4030025.
[15] S. Samtani, M. Abate, V. Benjamin, W. Li, Cybersecurity as an industry: a cyber threat intelligence perspective, in: The Palgrave Handbook of International Cybercrime and Cyberdeviance, Springer International Publishing, Cham. 2019, pp. 1—20, https://doi.org/10.1007/978-3-319-90307-1_8-1.
[16] F. Tazi, S. Shrestha, J. De La Cruz, S. Das, SoK: an evaluation of the secure end user experience on the dark net through systematic literature review, J. Cybersecur. Priv. 2 (2022) 329—357, https://doi.org/10.3390/jcp2020018.
[17] S. Madan, S. Sofat, D. Bansal, Tools and techniques for collection and analysis of internet-of-things malware: a systematic state-of-art review, J. King Saud Univ. Sci. 34 (2022) 9867—9888, https://doi.org/10.1016/j.jksuci.2021.12.016.
[18] Z.T. Sworna, Z. Mousavi, M.A. Babar, NLP methods in host-based intrusion detection systems: a systematic review and future directions, J Netw. Comput. Appl. 220 (2023) 1—29, https://doi.org/10.1016/j.jnca.2023.103761.
[19] T.O. Browne, M. Abedin, M.J.M. Chowdhury, A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications, Int. J. Inf. Secur. 23 (2024) 2911—2938, https://doi.org/10.1007/s10207-024-00868-2.
[20] A. Mahboubi, K. Luong, H. Aboutorab, H.T. Bui, G. Jarrad, M. Bahutair, S. Camtepe, G. Pogrebna, E. Ahmed, B. Barry, H. Gately, Evolving techniques in cyber threat hunting: a systematic review, J. Netw. Comput. Appl. 232 (2024) 1—34, https://doi.org/10.1016/j.jnca.2024.104004.
[21] G. Cascavilla, D.A. Tamburri, F. Leotta, M. Mecella, W.J. Van Den Heuvel, Counter-terrorism in cyber-physical spaces: best practices and technologies from the state of the art, Inf. Softw. Technol. 161 (2023) 1—20, https://doi.org/10.1016/J.INFSOF.2023.107260.
[22] D. Javaheri, M. Fahmideh, H. Chizari, P. Lalbakhsh, J. Hur, Cybersecurity threats in FinTech: a systematic review, Expert Syst. Appl. 241 (2024) 1—31, https://doi.org/10.1016/j.eswa.2023.122697.
[23] W.S. Admass, Y.Y. Munaye, A.A. Diro, Cyber security: state of the art, challenges and future directions, Cyber Secur. Appl. 2 (2024) 1—9, https://doi.org/10.1016/j.csa.2023.100031.
[24] G. Cascavilla, D.A. Tamburri, W. Van, D. Heuvel, Cybercrime threat intelligence: a systematic multi-vocal literature review, Comput Secur 105 (2021) 1—29, https://doi.org/10.1016/j.cose.2021.102258.
[25] L.S.Y. Teagle, Community, Word and Wonder: Discerning Key Elements in the Faith Inquiry of Chinese International Students, Doctoral Dissertation, Middlesex University London. 2023.
[26] P.R.J. Trim, Y.I. Lee, Combining sociocultural intelligence with artificial intelligence to increase organizational cyber security provision through enhanced resilience, Big Data Cogn. Comput. 6 (2022) 1—20, https://doi.org/10.3390/bdcc6040110.
[27] A. Bass, N. Demm, T. Jigme, A. Snyder, Advocating for Education Equity: Strategies and Opportunities for Great Expectations, Master's Thesis, University of Minnesota. 2021.
[28] E. Evans, S. Bardhan, Adult third culture kids and sojourner intercultural communication: exploring belonging through a multilevel approach, Int. J. Intercult. Relat. 96 (2023) 1—11, https://doi.org/10.1016/j.ijintrel.2023.101844.
[29] J.H. Al-Yasiri, M.F. bin Zolkipli, N.F.N. Mohd Farid, M. Alsamman, Z.A. Mohammed, A threat intelligence event extraction conceptual model for cyber threat intelligence feeds, in: 2024 7th International Conference on Internet Applications, Protocols, and Services (NETAPPS), IEEE, Kuala Lumpur. 2024, pp. 1—8, https://doi.org/10.1109/NETAPPS63333.2024.10823639.
[30] A. Cotov, C. Bono, C. Cappiello, B. Pernici, Improving cybersecurity awareness: tweet classification using multilingual sentence embeddings and contextual features, in: 2023 IEEE International Conference on Big Data (BigData), IEEE, Sorrento. 2023, pp. 3600—3607, https://doi.org/10.1109/BigData59044.2023.10386480.
[31] H. Nakano, D. Chiba, T. Koide, N. Fukushi, T. Yagi, T. Hariu, K. Yoshioka, T. Matsumoto, Canary in twitter mine: collecting phishing reports from experts and non-experts, in: Proceedings of the 18th International Conference on Availability, Reliability and Security (ARES ‘23), ACM, New York. 2023, pp. 1—7, https://doi.org/10.1145/3600160.3600163.
[32] H. Nakano, D. Chiba, T. Koide, N. Fukushi, T. Yagi, T. Hariu, K. Yoshioka, T. Matsumoto, Understanding characteristics of phishing reports from experts and non-experts on twitter, IEICE Trans. Inf. Syst. E107.D (2024) 807—824, https://doi.org/10.1587/transinf.2023EDP7221.
[33] F. Sufi, A global cyber-threat intelligence system with artificial intelligence and convolutional neural network, Decis. Anal. J. 9 (2023) 1—12, https://doi.org/10.1016/j.dajour.2023.100364.
[34] F. Sufi, A new AI-based semantic cyber intelligence agent, Future Internet 15 (2023) 1—27, https://doi.org/10.3390/fi15070231.
[35] F. Sufi, Novel application of open-source cyber intelligence, Electronics 12 (2023) 1—25, https://doi.org/10.3390/electronics12173610.
[36] F. Sufi, Social media analytics on Russia-Ukraine cyber war with natural language processing: perspectives and challenges, Information 14 (2023) 1—28, https://doi.org/10.3390/info14090485.
[37] K. Liu, Y. Wang, Z. Ding, A. Li, W. Zhang, BVTED: a specialized bilingual (Chinese—English) dataset for vulnerability triple extraction tasks, Appl. Sci. 14 (2024) 1—19, https://doi.org/10.3390/app14167310.
[38] X. Chang, Y. Liu, L. Huang, J. Li, Y. Liang, S. Li, Y. Sun, Blockchain threat intelligence knowledge graph alignment via graph convolutional networks, in: Proceedings of the 2023 International Conference on Information Education and Artificial Intelligence (ICIEAI '23), ACM, New York. 2023, pp. 421—430, https://doi.org/10.1145/3660043.3660119.
[39] H. Ji, J. Yang, L. Chai, C. Wei, L. Yang, Y. Duan, Y. Wang, T. Sun, H. Guo, T. Li, C. Ren, Z. Li, SEvenLLM: benchmarking, eliciting, and enhancing abilities of large language models in cyber threat intelligence, arXiv preprint arXiv.2405.03446 (2024) 1—11, https://doi.org/10.48550/arXiv.2405.03446.
[40] G. Siracusano, D. Sanvito, R. Gonzalez, M. Srinivasan, S. Kamatchi, W. Takahashi, M. Kawakita, T. Kakumaru, R. Bifulco, Time for action: automated analysis of cyber threat intelligence in the wild, arXiv preprint arXiv.2307.10214 (2023) 1—16, https://doi.org/10.48550/arXiv.2307.10214.
[41] B. Ampel, Predicting organizational cybersecurity risk: a deep learning approach, arXiv preprint arXiv.2012.14425 (2020) 1—5, https://doi.org/10.48550/arXiv.2012.14425.
[42] M. Ebrahimi, S. Samtani, Y. Chai, H. Chen, Detecting cyber threats in non-english hacker forums: an adversarial cross-lingual knowledge transfer approach, in: 2020 IEEE Security and Privacy Workshops (SPW), IEEE, San Francisco. 2020, pp. 20—26, https://doi.org/10.1109/SPW50608.2020.00021.
[43] M. Ebrahimi, Y. Chai, S. Samtani, H. Chen, Cross-lingual cybersecurity analytics in the international dark web with adversarial deep representation learning, Manag. Inf. Syst. Q. 46 (2022) 1209—1226, https://doi.org/10.25300/MISQ/2022/16618.
[44] R.T. Othman, Vulnerability detection for software-intensive system, in: Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering (EASE ‘24), ACM, Salerno. 2024, pp. 510—515, https://doi.org/10.1145/3661167.3661170.
[45] R. Othman, B. Rossi, B. Russo, Cybersecurity defenses: exploration of CVE types through attack descriptions, arXiv

preprint arXiv.2407.06759 (2024) 1—4, https://doi.org/10.48550/arXiv.2407.06759.

[46] T. Schaberreiter, J. Röning, G. Quirchmayr, V. Kupfersberger, C. Wills, M. Bregonzio, A. Koumpis, J.E. Sales, L. Vasiliu, K. Gammelgaard, A. Papanikolaou, K. Rantos, A. Spyros, A cybersecurity situational awareness and information-sharing solution for local public administrations based on advanced big data analysis: the CS-AWARE project, in: J. Bernal Bernabe, A. Skarmeta, eds., Challenges in Cybersecurity and Privacy: the European Research Landscape, 1st ed., River Publishers, Gistrup. 2020, pp. 149—180, https://doi.org/10.1201/9781003337492-8.

[47] R. Vast, S. Sawant, A. Thorbole, V. Badgujar, Artificial intelligence based security orchestration, automation and response system, in: 2021 6th International Conference for Convergence in Technology (I2CT), IEEE, Maharashtra. 2021, pp. 1—5, https://doi.org/10.1109/I2CT51068.2021.9418109.

[48] Z. Liu, K. Shi, N.F. Chen, Multilingual neural RST discourse parsing, in: Proceedings of the 28th International Conference on Computational Linguistics, ICCL, Barcelona. 2020, pp. 6730—6738, https://doi.org/10.18653/v1/2020.coling-main.591.

[49] G. Shin, D. Kim, S. Park, A. Park, Y. Kim, M. Han, Identifying similar users between dark web and surface web using BERTopic and authorship attribution, Electronics 14 (2025) 1—17, https://doi.org/10.18653/v1/2020.coling-main.591.

[50] K. Sprenkamp, M. Dolata, G. Schwabe, L. Zavolokina, Data-driven intelligence in crisis: the case of Ukrainian refugee management, Gov. Inf. Q. 42 (2025) 1—15, https://doi.org/10.1016/j.giq.2024.101978.

[51] D. Cohen, A. Elalouf, D. Citrinowicz, Uncovering Salafi jihadist terror activity through advanced technological tools, J. Polic. Intell. Count. Terror. 50 (2025) 1—17, https://doi.org/10.1080/18335330.2025.2478553.

[52] W. Kasri, Y. Himeur, H.A. Alkhazaleh, S. Tarapiah, S. Atalla, W. Mansoor, H. Al-Ahmad, From vulnerability to defense: the role of large language models in enhancing cybersecurity, Computation 13 (2025) 1—59, https://doi.org/10.3390/computation13020030.

[53] S. Shafee, A. Bessani, P.M. Ferreira, Evaluation of LLM chatbots for OSINT-based cyber threat awareness, Expert Syst. Appl. 261 (2025) 1—16, https://doi.org/10.1016/j.eswa.2024.125509.