CTI Dataset Construction from Telegram

Dincy R. Arikkat[1,†], Sneha B. T.[1,†], Serena Nicolazzo[2,*,†], Antonino Nocera[3,†], Vinod P.[1,†], Rafidha Rehiman K. A.[1,†] and Karthika R.[1,†]

[1]Department of Computer Applications, Cochin University of Science and Technology, Kerala, India
[2]Department of Science, Technology and Innovation, University of Eastern Piedmont, Alessandria, Italy
[3]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy

Abstract

Cyber Threat Intelligence (CTI) enables organizations to anticipate, detect, and mitigate evolving cyber threats. Its effectiveness depends on high-quality datasets, which support model development, training, evaluation, and benchmarking. Building such datasets is crucial, as attack vectors and adversary tactics continually evolve. Recently, Telegram has gained prominence as a valuable CTI source, offering timely and diverse threat-related information that can help address these challenges. In this work, we address these challenges by presenting an end-to-end automated pipeline that systematically collects and filters threat-related content from Telegram. The pipeline identifies relevant Telegram channels and scrapes 145, 349 messages from 12 curated channels out of 150 identified sources. To accurately filter threat intelligence messages from generic content, we employ a BERT-based classifier, achieving an accuracy of 96.64%. From the filtered messages, we compile a dataset of 86, 509 malicious Indicators of Compromises, including domains, IPs, URLs, hashes, and CVEs. This approach not only produces a large-scale, high-fidelity CTI dataset but also establishes a foundation for future research and operational applications in cyber threat detection.

Keywords

CTI, OSN, Social Network, Telegram, CTI dataset, Deep Learning, CTI Dataset, Cyber Threat Intelligence

1. Introduction

Cyber Threat Intelligence (CTI) has become indispensable for security analysts, enabling them to identify, collect, manage, and disseminate information on vulnerabilities and attacks, and to respond proactively to emerging threats [1]. Within the CTI lifecycle, data collection encompassing sources such as security alerts and threat intelligence reports from the web represents a critical foundational stage [2].

In this context, one challenge is that not all threat intelligence is published in standard CTI databases or integrated into commercial security platforms. Valuable CTI is often disseminated through unstructured channels such as blogs, social media posts, or reports from security companies and independent experts. To capture these dispersed insights, multiple online sources can be leveraged as early signals of emerging cyber threats. Information gathering thus becomes the first and most critical step, enabling the collection of relevant data on newly discovered vulnerabilities, active exploits, security alerts, threat intelligence reports, and security tool configurations. Curating CTI datasets requires addressing key challenges, including data sourcing from heterogeneous streams, ensuring data reliability, preserving privacy, and mitigating bias. A well-designed CTI dataset not only accelerates the advancement of automated threat intelligence systems but also strengthens global cyber defense capabilities through knowledge sharing and standardized evaluation frameworks. While platforms like Twitter [3] have been widely explored for their CTI potential, other communication ecosystems remain underexamined. Among them, messaging applications, particularly Telegram[1], have experienced exponential growth, evolving into key venues

Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09-13, 2026, Cagliari, IT
*Corresponding author.
†These authors contributed equally.
dincyrarikkat@cusat.ac.in (D. R. Arikkat); snehabtsneha@pg.cusat.ac.in (S. B. T.); serena.nicolazzo@uniupo.it (S. Nicolazzo); antonino.nocera@unipv.it (A. Nocera); vinod.p@cusat.ac.in (V. P.); rafidharehimanka@cusat.ac.in (R. R. K. A.); karthikar@pg.cusat.ac.in (K. R.)

[1]https://web.telegram.org

not only for general interaction but also for niche communities engaged in cybersecurity discourse, tool dissemination, and, in some cases, illicit activities. Telegram is a cloud-based messaging platform recognized for speed, privacy, and scalability in communication. It hosts public channels and groups covering a range of general and specialized topics, including cybersecurity and threat activity. Thanks to its openness and extensive global reach, Telegram has emerged as a significant source of Open Source Intelligence (OSINT), especially for tracking and analyzing emerging cyber threats, positioning it as a potentially valuable yet challenging source for CTI [4].

Building on these considerations, this study presents a comprehensive dataset comprising $145,349$ messages collected from 12 Telegram channels between January 2023 and February 2025[2]. Since messages from identified CTI sources may include content unrelated to threat intelligence, we developed a filtering mechanism to identify relevant CTI messages based on transformer models. Such filtering is a critical preprocessing step, as it eliminates generic, non-security-related content and ensures that downstream models are trained exclusively on high-fidelity CTI data. Following the identification of relevant messages, we construct an IoC dataset from Telegram content for CTI analysis.

In summary, the key contributions of this work are as follows:

- We systematically identified 12 high-value Telegram channels (from 150 candidates) as reliable CTI sources and implemented a custom crawler to continuously collect intelligence-rich content.
- We compiled a large-scale dataset of $145,349$ messages spanning two years, providing a substantial and timely resource for advancing CTI research.
- We designed and thoroughly evaluated a BERT-based automated filtering model that achieved high accuracy in identifying cybersecurity-relevant intelligence, thereby ensuring the dataset's reliability for downstream CTI applications.
- We compiled Indicators of Compromise dataset that can support both research and operational cyber threat detection.

The rest of the article is organized as follows: Section 2 reviews related work, Section 3 details our CTI dataset compilation, Section 4 presents experiments and evaluation, and Section 5 concludes the paper.

2. Related Work

A wide range of online platforms, including security blogs, forum posts, and Online Social Networks (OSNs), are frequently leveraged by both cybersecurity vendors and malicious actors to disseminate CTI in highly unstructured formats. These early disclosures often precede formal reporting and integration into authoritative and standardized repositories such as the Common Vulnerabilities and Exposures (CVE[3]) database or the National Vulnerability Database (NVD[4]) [5]. While CVE and NVD provide timely and potentially critical insights but focus exclusively on known vulnerabilities. Hence, several researchers and practitioners have started to collect OSINT data through custom crawlers [6]. Crawling for CTI is not confined to Clear Web resources (the publicly accessible portion of the Internet) but also extends to the Dark Web, Deep Web, and OSNs [2].

Several domain-specific crawlers target the Clear Web. For example, inTime [7] and MalCrawler [8] are optimized to identify relevant pages before initiating the crawl, allowing them to filter out benign content and improve efficiency. Few contributions propose the construction of a CTI-relevant dataset [9, 10], but specifically deal with NER and RE tasks within the CTI domain [11]. In the context of OSN-based crawling, several works [12, 13, 14, 15, 3] have focused on leveraging Twitter as a primary CTI source. These approaches are capable of detecting, geolocating, categorizing, and tracking cybersecurity-related events in real time by monitoring the Twitter stream. Typically, they rely on a curated list of

[2]The dataset is publicly available at: https://github.com/OPTIMA-CTI/Telegram-CTI
[3]https://cve.mitre.org
[4]https://nvd.nist.gov

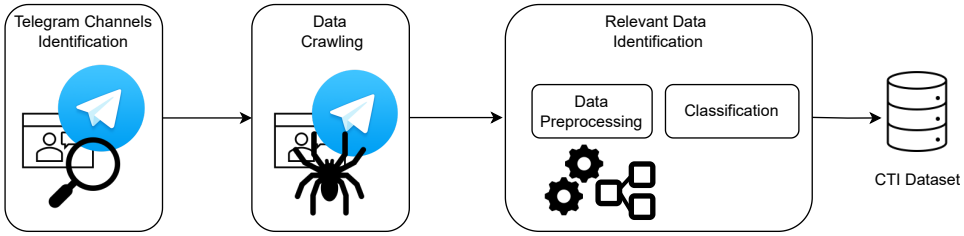

Figure 1: Architecture for the construction of the CTI dataset from Telegram

seed keywords, often provided by domain experts, that serve as input to the streaming API, allowing the collection, detection, and classification of cyber threat indicators from tweets. However, other platforms such as Reddit, Pastebin, and GitHub have also been explored for CTI extraction [16, 17]. Although Telegram presents significant potential as a CTI source, systematic approaches for extracting intelligence from the platform remain limited [4]. Several obstacles complicate this process, including the overwhelming volume of messages, the frequent presence of non-English content, the unstructured and conversational style of discussions, and technical constraints such as API limitations that hinder large-scale data collection [18, 19]. Overcoming these challenges requires automated methods capable of efficiently processing high-throughput message streams while filtering actionable intelligence from background noise. Initial efforts have begun to bridge this gap, combining AI-based models with human annotation to analyze Telegram-derived threats [4], alongside broader research on automated CTI extraction from social data streams that can be adapted to this domain [20]. In addition, specialized tools such as TelegramScrap have emerged to address platform-specific scraping limitations, further supporting data acquisition.

3. Proposed Approach

The workflow begins with the *selection of suitable Telegram channels* for data collection (Section 3.1). Next, the *gathering phase* is conducted through the platform's standard interface (Section 3.2). The collected messages then undergo *text cleaning and preprocessing*, after which multiple *BERT-based models* are trained to automatically filter valuable CTI content from irrelevant material (Section 3.3). Finally, we compile an IoC dataset to support in-depth CTI analysis (Section 3.4). An overview of the framework architecture is provided in Figure 1.

3.1. Telegram Channels Identification

In this phase, our objective is to identify Telegram channels that actively discuss attacks, threats, vulnerabilities, and share indicators of compromise (IoCs). To achieve this, we surveyed approximately 150 public channels, drawing from both prior research and open-source repositories. In particular, we referenced DarkGram [21], as well as publicly available Telegram channel lists such as BreachSense's catalog of threat actor channels[5] and one GitHub repostory [6]. In addition, we performed manual exploration within Telegram using cyber-related keywords (e.g., IoCs, CVE, DDoS, cyber attack, malware, ransomware, etc), thereby ensuring comprehensive coverage of channels relevant to CTI. Subsequently, we evaluated each channel against five criteria: *(i)* demonstrated relevance to threat intelligence [20], *(ii)* depth of technical discussion and frequency of activity, *(iii)* primary language (English), *(iv)* evidence of direct IoC sharing, and *(v)* whether the channel was active and accessible. Following this multi-criteria assessment, we selected 12 channels deemed most suitable for in-depth data collection.

[5] https://www.breachsense.com/threat-actor-channels/
[6] https://github.com/ghostwond3r/telegram_channel

Table 1
Cybersecurity Telegram Channels Statistics

Number	Channel Name	Messages	Subscribers	AMD*	Top 5 words
C1	DLM - CVE Monitor	19,142	884	84	cve, vulnerability, affected, link, products
C2	Cybersecurity & Privacy - News	27,254	23,542	38	vulnerability, cve, database, cibsecurity, security
C3	Pro-Palestine Hackers Movement	967	5,295	6	company, data, website, israeli, attacked
C4	Z-BL4CK-H4T	250	4,359	17	israel, website, rippersec, investigation, undergroundnet
C5	RipperSec	4,510	5,151	26	rippersec, sedihcrew, zenimous, target, team
C6	Dark Web Informer - Cyber Threat Intelligence - CVE Alerts	6,021	157	118	cve, threat, intelligence, cyber, vulnerability
C7	BleepingComputer	2,489	8,544	6	data, windows, security, ransomware, microsoft
C8	The Hacker News	3,472	144,734	5	security, malware, data, cve, critical
C9	CVE Notify	32,283	15,285	114	cve, vulnerability, issue, user, attacker
C10	CVE Tracker	16,227	64	87	cve, vulnerability, affected, link, products
C11	Cyber Threat Intelligence	30,927	31,267	86	cyber, cve, security, attack, data
C12	Hackmanac Cyber Alerts	1,807	2,738	9	data, group, cyberattack, ransomware, alert
	Total	**145,349**	**242,020**	**186**	cve, vulnerability, affected, cvss, link

***AMD - Average Message per Day**

3.2. Data Crawling

After identifying the target channels as trusted data sources, we proceeded with the crawling phase. For this task, we relied on the official Telegram API and, in particular, the Telethon library[7] (asynchronous MTProto API client for Python). Using this setup, we scraped messages from the selected channels between January 2023 and February 2025, resulting in a dataset of 145, 349 messages. Table 1 summarizes the selected channels along with their respective message volumes, subscriber counts, average message per day, and top five words. Figure 2 illustrates the monthly distribution of posts across all channels. From the figure, it is evident that the volume of messages posted during 2024–2025 was substantially higher compared to 2023.

3.3. Relevant CTI Data Classification

While our corpus comprises 145, 349 messages crawled from CTI-related Telegram channels, not all messages explicitly reference attack incidents, vulnerability, or IoCs. To address this limitation, we introduce an additional filtering stage designed to retain only content that is directly relevant to CTI. Specifically, we develop a classification model to distinguish CTI-relevant messages from unrelated chatter. Before proceeding with the classification tasks, we perform a preprocessing step that reduces noise and produces cleaner text for downstream analysis. This preprocessing stage consists of a sequence of operations: *(i)* **IoC Normalization:** CTI messages often contain IoCs such as IP addresses, URLs, CVE identifiers, and file hashes, which may appear in standard, defanged, or obfuscated forms. We normalized these entities by replacing them with placeholder tokens: `[ip]`, `[url]`, `[cve]`, and `[hash]` to retain key threat information. *(ii)* **Lowercasing:** All text is converted to lowercase to ensure uniform representation, reducing vocabulary sparsity and simplifying downstream processing. *(iii)* **Removal of**

[7] [https://docs.telethon.dev/en/stable]

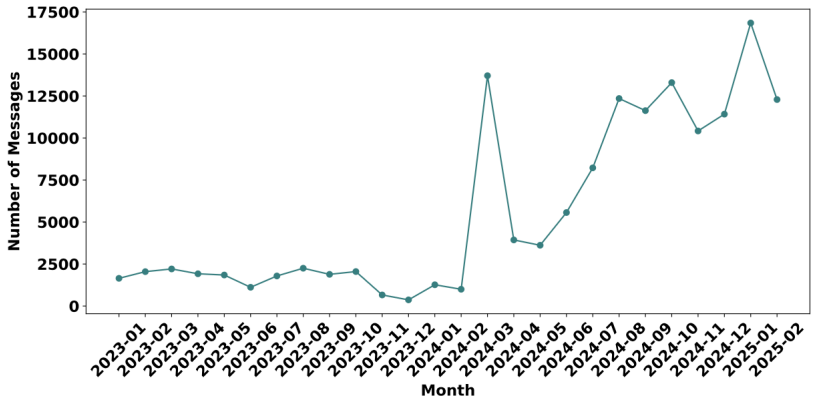

Figure 2: Monthly Message Volume Across All Telegram Channels

emojis and non-essential special characters: While emojis are commonly used on social media to convey user emotions or sentiment, they typically introduce noise rather than substantive content for identifying CTI relevance. Consequently, they were removed to reduce data sparsity and also to remove special characters. *(iv)* **Lemmatization:** We use lemmatization to reduce words to their base form, allowing the model to generalize across related variants (e.g., *attacks*, *attacking*, *attacked* $\rightarrow$ *attack*).

Subsequently, we leveraged the transformer-based models to identify and filter CTI-relevant content from our collected corpus. Specifically, we employed models including *(i)* standard BERT (Bidirectional Encoder Representations from Transformers) [22], *(ii)* DistilBERT (Distilled version of BERT) [23], *(iii)* RoBERTa (Robustly Optimized BERT) [24], *(iv)* CySecBERT [25], and *(v)* SecBERT[8], which are widely used in Natural Language Processing tasks [26].

3.4. IoC Extraction and Verification

To extract potential threat indicators from Telegram messages, we developed a set of tailored regular expressions (RegEx). These expressions were designed to capture web URLs, IP addresses, domain names, file hashes, and CVE identifiers. Since RegEx-based extraction may also capture benign or irrelevant entries, we performed additional validation. Indicators were cross-checked using VirusTotal[9], and the NVD for CVE verification. Through this two-step enrichment process, we curated a refined set of malicious IoCs suitable for threat intelligence applications.

4. Experiments

In this section, we evaluate the performance of five BERT-based models in identifying relevant CTI texts and also investigate the IoCs collected from the relevant messages. Model effectiveness is assessed using Accuracy and F1-score. To construct a reliable dataset for training the relevance classification model, we estimated the required sample size using the standard statistical approach for finite populations [27]. Based on a 95% confidence level, a 1% margin of error, and an assumed population proportion of 50%, the resulting sample size was approximately $9,009$ messages from the total corpus of $145,349$. The selected messages were manually annotated to create the labeled dataset. Each message was independently reviewed and assigned a label of either *Relevant* (containing actionable threat intelligence) or *Irrelevant* (lacking such content). To ensure annotation reliability, we measured inter-annotator agreement using Cohen's Kappa (k) [28], which yielded 0.90, indicating "almost perfect" agreement and confirming the reliability of annotation. The annotation process, however, revealed an imbalance

[8]https://github.com/jackaduma/SecBERT
[9]https://www.virustotal.com/gui/home/upload

Table 2
Performance Comparison of BERT-based Models

Model	Accuracy	F1-Score (Class 0,1)
DistilBERT	95.83%	0.96, 0.96
CySecBERT	95.42%	0.95, 0.95
RoBERTa	96.00%	0.96, 0.96
SecBERT	95.19%	0.95, 0.95
BERT	**96.64%**	**0.97, 0.97**

between *Relevant* and *Irrelevant* classes. To avoid classifier bias, we applied random under-sampling, yielding a balanced dataset of $8,634$ messages ($4,317$ Relevant, $4,317$ Irrelevant), which served as ground truth for training and evaluation. For model development, the dataset was divided into three subsets: 70% was allocated for training, 10% was used as a validation set, and the remaining 20% was reserved as a test set. Table 2 presents the comparative performance of the five evaluated models: BERT, DistilBERT, RoBERTa, CySecBERT, and SecBERT. Among these, the standard `bert-base-uncased` model achieved the strongest performance on our test data, attaining an Accuracy of 96.6% and an F1-score of 0.97. This high performance highlights the robustness and reliability of the BERT model in accurately identifying threat intelligence content.

After developing the relevant CTI content classification model, we applied it to the remaining unlabeled messages in the corpus. These messages underwent the same preprocessing steps as the training data and were then classified using the fine-tuned binary BERT model. As a result, the BERT model classified the messages into $99,340$ relevant and $42,510$ irrelevant messages.

The relevant messages were further analyzed for IoC extraction to support threat intelligence. Using regular expressions, we initially extracted a total of $188,290$ indicators from the dataset, as summarized in Table 3. Since not all extracted indicators were malicious, we performed verification using VirusTotal and the NVD database, which resulted in a refined set of $86,509$ confirmed malicious indicators (see Table 3). The analysis of this curated collection revealed a notable distribution across different IoC types. Out of the total indicators collected, the majority were CVEs, followed by URLs, IPs, Domains, and Hashes. Specifically, CVEs accounted for about 45.5% of all collected indicators, URLs 50.9%, IPs 2.1%, Domains 1.0%, and Hashes 0.5%. When considering only the malicious indicators, nearly all were CVEs, while URLs, IPs, Domains, and Hashes contributed only a small fraction. This threat indicator dataset serves as a benchmark for further threat analysis.

Table 3
IoCs extracted and validated per channel. T - Total indicators extracted, **M** - Malicious indicators

Channel	Domain		IP		URL		Hash		CVE	
	T	M	T	M	Total	M	T	M	T	M
C1	237	19	731	25	24	1	173	0	19153	19145
C2	354	36	605	22	26935	2	341	0	15827	15827
C3	419	4	10	0	315	13	0	0	0	0
C4	105	3	2	0	186	24	2	0	0	0
C5	65	3	630	77	0	0	0	0	1	1
C6	86	5	276	11	11476	160	30	1	5348	5348
C7	21	8	2	0	4972	125	0	0	55	54
C8	0	0	4	1	3452	17	0	0	475	474
C9	277	24	923	38	18617	13	182	0	25503	25503
C10	208	16	640	23	191	2	151	0	16309	16309
C11	185	7	74	3	28977	51	0	0	3074	3073
C12	7	0	0	0	628	9	0	0	32	32

5. Conclusion

In this work, we present a large-scale CTI dataset comprising $145,349$ messages collected from 12 selected Telegram channels between January 2023 and February 2025. We train and evaluate a binary

BERT-based classifier, designed to automatically filter cybersecurity-relevant messages with a high accuracy of 96.64%. From the curated messages, we assembled a comprehensive dataset of $86,509$ malicious IoCs. Our contributions include the systematic identification of high-value CTI channels on Telegram, the creation of a CTI message dataset with IoCs, and the development of an automated filtering pipeline that enhances the quality and usability of CTI data for research and operational purposes. In the future, we plan to expand this dataset, considering also blogs in the Dark/Deep Web and other Social Network scenarios. Moreover, we plan to evaluate IoCs using other threat intelligence feeds such as AlienVault, MalwareBazaar, etc., and also extract information about attack behaviours.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] R. Brown, R. M. Lee, 2021 sans cyber threat intelligence (cti) survey, Tech. Rep. SANS Institute (2021).

[2] M. Arazzi, D. R. Arikkat, S. Nicolazzo, A. Nocera, R. R. KA, M. Conti, et al., Nlp-based techniques for cyber threat intelligence, Computer Science Review 58 (2025) 100765.

[3] H. Shin, W. Shim, S. Kim, S. Lee, Y. G. Kang, Y. H. Hwang, # twiti: Social listening for threat intelligence, in: Proceedings of the Web Conference 2021, 2021, pp. 92–104.

[4] K. Ravi, A. E. Vela, E. Jenaway, S. Windisch, Exploring multi-level threats in telegram data with ai-human annotation: a preliminary study, in: 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE, 2023, pp. 1520–1527.

[5] A. K. Threatt, Some analysis of common vulnerabilities and exposures (cve) data from the national vulnerability database (nvd), repository.uncw.edu (2024).

[6] F. Tabatabaei, D. Wells, Osint in the context of cyber-security, Open Source Intelligence Investigation: From Strategy to Implementation (2017) 213–231.

[7] P. Koloveas, T. Chantzios, C. Tryfonopoulos, S. Skiadopoulos, A crawler architecture for harvesting the clear, social, and dark web for iot-related cyber-threat intelligence, in: 2019 IEEE World Congress on Services (SERVICES), volume 2642, IEEE, 2019, pp. 3–8.

[8] A. Singh, N. Goyal, Malcrawler: A crawler for seeking and crawling malicious websites, in: Distributed Computing and Internet Technology: 13th International Conference, ICDCIT 2017, Bhubaneswar, India, January 13-16, 2017, Proceedings 13, Springer, 2017, pp. 210–223.

[9] Y. Zhou, Y. Ren, M. Yi, Y. Xiao, Z. Tan, N. Moustafa, Z. Tian, Cdtier: A chinese dataset of threat intelligence entity relationships, IEEE Transactions on Sustainable Computing 8 (2023) 627–638.

[10] X. Wang, S. He, Z. Xiong, X. Wei, Z. Jiang, S. Chen, J. Jiang, Aptner: A specific dataset for ner missions in cyber threat intelligence field, in: 2022 IEEE 25th international conference on computer supported cooperative work in design (CSCWD), IEEE, 2022, pp. 1233–1238.

[11] D. R. Arikkat, P. Vinod, R. R. KA, S. Nicolazzo, A. Nocera, M. Conti, Relation extraction techniques in cyber threat intelligence, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2024, pp. 348–363.

[12] Q. Le Sceller, E. B. Karbab, M. Debbabi, F. Iqbal, Sonar: Automatic detection of cyber security events over the twitter stream, in: Proceedings of the 12th International Conference on Availability, Reliability and Security, 2017, pp. 1–11.

[13] F. Alves, A. Bettini, P. M. Ferreira, A. Bessani, Processing tweets for cybersecurity threat awareness, Information Systems 95 (2021) 101586.

[14] A. Rodriguez, K. Okamura, Generating real time cyber situational awareness information through social media data mining, in: 2019 IEEE 43rd annual computer software and applications conference (COMPSAC), volume 2, IEEE, 2019, pp. 502–507.

[15] L.-M. Kristiansen, V. Agarwal, K. Franke, R. S. Shah, Cti-twitter: gathering cyber threat intelligence

from twitter using integrated supervised and unsupervised learning, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 2299–2308.
[16] S. Horawalavithana, A. Bhattacharjee, R. Liu, N. Choudhury, L. O. Hall, A. Iamnitchi, Mentions of security vulnerabilities on reddit, twitter and github, in: IEEE/WIC/ACM International Conference on Web Intelligence, 2019, pp. 200–207.
[17] T. Vahedi, B. Ampel, S. Samtani, H. Chen, Identifying and categorizing malicious content on paste sites: a neural topic modeling approach, in: 2021 IEEE International Conference on Intelligence and Security Informatics (ISI), IEEE, 2021, pp. 1–6.
[18] A. Dutta, S. Kant, An overview of cyber threat intelligence platform and role of artificial intelligence and machine learning, in: Information Systems Security: 16th International Conference, ICISS 2020, Jammu, India, December 16–20, 2020, Proceedings 16, Springer, 2020, pp. 81–86.
[19] M. R. Rahman, R. M. Hezaveh, L. Williams, What are the attackers doing now? automating cyberthreat intelligence extraction from text on pace with the changing threat landscape: A survey, ACM Computing Surveys 55 (2023) 1–36.
[20] J. Zhao, Q. Yan, J. Li, M. Shao, Z. He, B. Li, Timiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data, Computers & Security 95 (2020) 101867.
[21] S. S. Roy, E. P. Vafa, K. Khanmohamaddi, S. Nilizadeh, {DarkGram}: A {Large-Scale} analysis of cybercriminal activity channels on telegram, in: 34th USENIX Security Symposium (USENIX Security 25), 2025, pp. 4839–4858.
[22] J. Lee, K. Toutanova, Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 3 (2018) 8.
[23] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[25] M. Bayer, P. Kuehn, R. Shanehsaz, C. Reuter, Cysecbert: A domain-adapted language model for the cybersecurity domain, arXiv preprint arXiv:2212.02974 (2022).
[26] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: China national conference on Chinese computational linguistics, Springer, 2019, pp. 194–206.
[27] H. Ahmad, H. Halim, Determining sample size for research activities: the case of organizational research, Selangor Business Review (2017) 20–34.
[28] S. Kılıç, Kappa testi, Journal of mood disorders 5 (2015) 142–144.