Using a Stack to Find an AI Needle: Topic Modeling for Cyber Threat Intelligence

SASKIA LAURA SCHRÖER, Liechtenstein Business School, University of Liechtenstein, Liechtenstein
JEREMY D. SEIDEMAN, Computer Science Department, The Graduate Center, CUNY, USA
SHOUFU LUO, Computer Science Department, The Graduate Center, CUNY, USA
GIOVANNI APRUZZESE, Liechtenstein Business School, University of Liechtenstein, Liechtenstein
SVEN DIETRICH, Computer Science Department, Hunter College & The Graduate Center, CUNY, USA
PAVEL LASKOV, Liechtenstein Business School, University of Liechtenstein, Liechtenstein

Cyber Threat Intelligence (CTI) is a fundamental activity to ensure the protection of modern organizations against sophisticated cyberattackers. A large body of literature has addressed problems related to CTI. Despite the scientific validity of such results, the reality is that CTI practitioners rarely deploy advanced CTI methods proposed by the research community and mostly rely on manual processes. We seek to facilitate the manual analyses typical for CTI practice by proposing a novel topic modeling technique that enables analysts to identify specific topics in CTI data sources. We demonstrate how our method, released as an open-source tool, can be used to investigate three case studies revolving around the research question whether attackers are deploying AI for malicious purposes "in the wild," and, if so, what features of AI interest them the most. We analyzed 7 million discussions from 18 underground forums. Our findings reveal that attackers may favor easy-to-use AI toolkits over the sophisticated AI techniques envisioned in research papers. Our contributions are further validated by a user study (N=24) with CTI experts, confirming the relevance of our research. Ultimately, we advocate future endeavors to account for the opinion of CTI practitioners—who should, in turn, try to cooperate.

CCS Concepts: • **Security and privacy → Human and societal aspects of security and privacy**; • **Applied computing → Document searching**; • **Computing methodologies → Natural language processing**.

Additional Key Words and Phrases: Cyber Threat Intelligence, Underground Forums, User Study

ACM Reference Format:
Saskia Laura Schröer, Jeremy D. Seideman, Shoufu Luo, Giovanni Apruzzese, Sven Dietrich, and Pavel Laskov. 2025. Using a Stack to Find an AI Needle: Topic Modeling for Cyber Threat Intelligence. *Digit. Threat. Res. Pract.* 0, 0, Article 0 (2025), 39 pages. https://doi.org/10.1145/3766908

1 INTRODUCTION

The perpetual advances in information technology are a double-edged sword. On the one hand, they bring an unparalleled improvement of life quality for the society [77]. On the other hand, they unveil new possibilities for evildoers to carry out their mischievous deeds. Besides merely increasing the attack surface [15, 41, 140], digital innovation also may

Authors' addresses: Saskia Laura Schröer, Saskia.Schroeer@uni.li, Liechtenstein Business School, University of Liechtenstein, Vaduz, Liechtenstein; Jeremy D. Seideman, jseideman@gradcenter.cuny.edu, Computer Science Department, The Graduate Center, CUNY, New York, NY, USA; Shoufu Luo, sluo2@gradcenter.cuny.edu, Computer Science Department, The Graduate Center, CUNY, New York, NY, USA; Giovanni Apruzzese, Giovanni.Apruzzese@uni.li, Liechtenstein Business School, University of Liechtenstein, Vaduz, Liechtenstein; Sven Dietrich, spock@ieee.org, Computer Science Department, Hunter College & The Graduate Center, CUNY, New York, NY, USA; Pavel Laskov, Pavel.Laskov@uni.li, Liechtenstein Business School, University of Liechtenstein, Vaduz, Liechtenstein.

Digit. Threat. Res. Pract.

provide **new weapons for cyber-attackers**, best exemplified by the recent emergence of "weaponized" or "offensive" AI [92, 138]. In light of such a constantly evolving threat landscape, modern organizations, in both the public and private sector [116], habitually rely on Cyber Threat Intelligence (CTI) to adapt their defenses to novel threats.

The prime objective of CTI is to reveal novel attack traits by analyzing the behavior of cybercriminals. Such insights help defenders to stay up-to-date with current offensive strategies, techniques and tactics, and to potentially defuse incoming attacks before they hit their targets [5, 141]. Despite being a time-consuming activity for security analysts [34], gathering CTI is known to be effective to thwart even advanced forms of cyberattacks [23]. It is hence hardly surprising that a substantial body of research addressed problems related to CTI (see [39, 128, 131] for literature reviews). Exemplary works propose techniques for identification of specific security events, e.g., vulnerability exploitation [94] or malware installation [44], or for extraction of generic knowledge, e.g., in form of knowledge graphs [120], from underground forums or other pertinent data sources.

Recent reports suggest, however, that a large share of CTI extraction is still performed manually [34]. This adversely affects the cost and the timeliness of CTI in practice. While manual investigation is certainly indispensable for precise understanding of complex CTI matters, we argue that such investigation can be facilitated by narrowing the analyst's attention to specific topics of interest. Specifically, we demonstrate how Natural Language Processing (NLP) techniques can be leveraged to learn characterizations of topics of potential interest from reliable data sources, which can be subsequently deployed for extraction of threat intelligence related to such topics.

While using NLP in the context of CTI is by no means unknown (see [17] for a recent review of related literature), most of such works have hardly addressed the practical side of CTI, as shown in this paper. Based on this observation, we developed a novel topic modeling technique which enables one to transfer knowledge *across various data sources* and hence increase the accuracy and the confidence of identified topics. We demonstrate the high practicality of the proposed approach by studying to what extent cybercriminals are interested in applying AI for malevolent purposes, and shed light on their specific interests. Our technique is primarily concerned with the extraction of generic threat intelligence knowledge, yet the generality of our method makes it also applicable for investigation of more specific CTI tasks. In addition, we present a post hoc survey with contributions from 24 CTI practitioners, which justifies and validates our research and design choices.

Contributions. The underlying motif of this work is to simultaneously account for both the *research* and *practical* aspects of CTI, aiming to build a bridge between, and then enhance, each of these ends. Hence, our scientific contributions encompass "technical," "conceptual," and "analytical" areas. Specifically:

- We review prior literature on CTI under a practical lens (Section 2). We find that no technical paper concerned with application of NLP techniques for CTI has fully accounted for practitioners' perspective. We suggest **recommendations** to rectify this disconnection—which we follow in our work.
- We propose TopTran, a new method to facilitate a common CTI task: the **analysis of underground forums** (Section 3). TopTran enables "transferring" topic models built on reliable data with known thematic focus, e.g., legitimate discussion forums, to investigate unknown data—hence **facilitating exploratory work of CTI experts**.
- We demonstrate a **practical implementation** of TopTran (Section 4). We showcase how to develop a topic model that is ready for operational deployment. We publicly release our tool [42].
- We perform a realistic CTI campaign by applying our fine-tuned model (Section 5). Through **three case studies**, entailing 6.9M threads from 18 underground forums spanning over 9 years, we investigate if these communities are interested in abusing AI for malicious purposes. We find scarce evidence of advanced attacks.

- We validate all of the above with a **user study with 24 CTI practitioners** (Section 7) whose findings can guide future research. To the best of our knowledge, this is the first paper on CTI whose technical contributions are *also* validated with a user study with practitioners.

We also provide further assessments and critically analyze our paper under a practical lens (Section 8). Finally, we report extensive technical discussions and in-depth qualitative analyses in Appendix C through Appendix D.

2 THE GAP BETWEEN RESEARCH AND PRACTICE IN CYBER THREAT INTELLIGENCE

As our first contribution, which also serves as a motivation for subsequent development, we survey prior research in CTI (Section 2.1), and then illustrate limitations of scientific literature from the perspective of CTI practitioners (Section 2.2). We conclude by outlining the possibilities for bridging the gap between research and practice in this field (Section 2.3), which we will use as a guide in the remaining sections of the paper.

2.1 The State of Research in CTI

A large amount of scientific works have CTI as their core focus. Aside from literature surveys (e.g., [39, 128, 131]) and SoK papers (e.g., [70]), prior research on CTI can be divided into two categories:

- *Technical papers*, which advance the state-of-the-art by proposing solutions that either facilitate existing CTI tasks, or enable the exploration of new CTI insights. For instance, Shin et al. [121] propose a word-based algorithm to quickly identify critical security events in large data pools; whereas, Jo et al. [72], propose a management system for entire CTI pipelines (from data collection to utilization).
- *Case studies*, which carry out exploratory or longitudinal analyses on a certain data source and contribute with novel findings that are useful to the cybersecurity community (e.g., as an avenue for future research, or to proactively improve operational cyber defense systems). For instance, underground forums [94], darknet marketplaces [57], security feeds [32, 60], as well as social media [1], have all been widely investigated by prior works.

There are also papers that fall into both of these categories. For instance, Sun et al., [129] propose a new solution to investigate private interactions in underground forums, which is then used to analyze data from the real world and yield novel findings usable for CTI. Other notable examples are in the papers by Chiba et al. [38] and Kim et al. [74].

2.2 Practical Shortcomings of CTI Research

Unfortunately, most scientific works connecting NLP with CTI exhibit some intrinsic shortcomings that diminish the *practical value* of the corresponding contributions. Indeed, the two previously identified categories of research papers are affected by two subtle issues which, despite being common in security-related research, are particularly important for CTI. Our reflective analysis of these issues is aimed to rectifying such shortcomings in current and future research.

2.2.1 ***Untimely findings.*** The foundation of CTI is to continuously analyze the digital world in order to be "one step ahead" of cyber criminals [131]. It is hence mandatory that the results of any CTI campaign reach relevant personnel (e.g., decision-makers, or system administrators) as quickly as possible. Unfortunately, scientific literature can hardly achieve such an objective; this is due to the *lengthy publication process* that research papers must undergo before being credited as "peer-reviewed work." For instance, conferences require 2–5 months from submission to acceptance, whereas journals can take years;[1] and this long timespan does not account for potential "rejections" (which not necessarily are due to flaws in the findings themselves). Consequently, while findings provided by research papers can be inspiring

[1] E.g., a recent work [24] was submitted in July 2022, and accepted in August 2023.

for future academic work, their value in practice is questionable.[2] This issue is noteworthy for measurement papers; technical papers are also affected since a given solution may become outdated during the time-span between submission and acceptance of an article.

2.2.2 ***Lack of practitioner input.*** Among the recipients of CTI solutions (or findings), there are *practitioners*, i.e., security operators, policy-makers, and corporate managers [5]. However, prior research overlooks the viewpoint of practitioners—especially *technical papers*. To understand why this lack is problematic, consider the following:

- Any software tool must be developed and maintained. Without *(i)* inquiring practitioners if the tool addresses a meaningful *purpose*, and *(ii)* measuring how *costly* it is to setup such a tool, its practical value would be questionable.
- Any CTI tool requires *data* to be analyzed. Without asking if such data sources are relevant for practitioners, the overall credibility of the findings would be undermined.
- After analyzing some data, the tool must provide an *output*. Such an output may not be easy to interpret for practitioners, which would discourage its usage in the real world.

Put simply, neglecting to account for the practitioners' know-how prevents emphasizing the practical value of a given technical research paper, thereby intrinsically inhibiting its real-world utility.

Table 1. **Technical papers on CTI proposing original solutions/tools reliant on NLP.** For each work, we report: the Data Source (security reports, darknet/marketplaces, underground forums, social media, others) used for the assessment; whether the source code is publicly available; and whether the proposal (i.e., the chosen data source, the envisioned purpose, or the provided output) was scrutinized by practitioners.

Year	Paper (1st author)	Data Source					Open Source?	Practitioner Validation		
		Sec. Rep.	*Dknet / Mktpl*	*Udg. For.*	*Soc. Med.*	*Other*		*Data*	*Purpose*	*Output*
2018	Williams [135]			×						
	Almukaynizi [10]		×	×						
	Ebrahimi [52]		×							
	Tavabi [130]			×		×				
	Deb [44]			×						
	Nunes [97]		×	×						
	Deliu [47]			×						
2019	Zenebe [143]			×						
	Arnold [18]		×	×						
	Schäfer [120]			×						
	Sakar [118]			×						
2020	Marin [86]			×		×				
	Ampel [11]		×	×						
	Yang [142]	×								
	Adewopo [1]			×	×					
	Ebrahimi [51]			×			✓	✓		
	Samtani [117]			×			✓	✓		
	Shin [121]				×					
	Zhao [146]	×		×	×					
2021	Vahedi [133]					×		✓		
	Gao [56]	×								
	Zhang [144]	×								
2022	Ji [71]	×								
	Jo [72]	×								
2023	Bayer [24]				×		✓			
	Moreno-Vera [94]			×			✓			

[2]Unfortunately, uploading a pre-print online is not helpful. Lack of any peer-review makes the resulting findings of dubious utility, especially for practitioners, who would hardly invest their limited time by looking at unpublished works.

2.2.3 ***Validation: Literature Analysis on "practitioner viewpoint".*** To provide exemplary evidence of the gap between CTI research and practice, we carry out a thorough analysis of prior technical papers that use NLP methods to analyze textual data, since this is the focus of our technical contribution (discussed in Section 3). Our analysis was performed by two authors, who frequently interacted and exchanged opinions. At a high-level, we proceeded as follows.

(1) *Search.* We queried Google Scholar between October and November, 2023. We started with a keyword search and applied backwards snowballing to avoid bias in favor of some publishers [136].[3] Since we are interested in recent (peer-reviewed) works, we consider papers published from January 2018 to November 2023.
(2) *Filtering.* Overall, we reviewed over 70 "technical" papers related to CTI, proposing tools or solutions to extract CTI (e.g., we excluded literature reviews or position papers).[4] Our inclusion criteria entail papers that use one or more of the following data sources: *(a)* security reports, *(b)* darknet webpages/marketplaces, *(c)* underground forums, and *(d)* social media. For instance, we exclude works (e.g., [9]) that focus on "internal intelligence" [114] (such as [74]), due to it being an area orthogonal to the one addressed in this paper.[5]
(3) *Analysis.* We eventually collected a total of 26 papers that aligned with our scope, and which we analyze in detail. We questioned whether the paper accounted for the practitioners' viewpoints. These viewpoints entail: whether the data source for the analysis is appropriate; whether the task falls in CTI practitioners' routines; and whether the output of the tool is actionable; importantly, we scrutinize whether the paper provided factual evidence that a CTI practitioner was inquired about any of such dimensions. We also examine whether the paper open-sourced the technical implementation of the proposed method.

We summarize the results of our literature analysis in Table 1. The rightmost columns refer to the bullet-points discussed in the previous subsection (Section 2.2.2). From Table 1, we see that *no paper fully accounted for the viewpoint of practitioners*. Importantly, only three papers explicitly reached out to practitioners to justify some aspects of their research. Specifically:

- Ebrahimi et al. [51] selected four underground forums for their study after having "consulted with cybersecurity experts and researchers well versed in Dark Web analytics and the underground economy." Additionally, to confirm the choice of these forums, they stated that "these platforms are well-known within the Dark Web ecosystem."
- Samtani et al. [117] selected one large underground forum with the motivation that "[it was] suggested by several cybersecurity experts who are well versed in Dark Web analytics from academic and industry perspectives."
- Vahedi et al. [133] identified "three prevailing paste sites for collection based on feedback from cybersecurity experts."

However, neither of these works inquired the practitioners' opinion on the purpose or the output of their technical contribution. We use this finding as an inspiring motivation to try to bridge the gap we brought to light.

2.3 Aligning CTI Research and Practice

Rectifying the disconnect between research and practice is challenging, but it can be done. The long time before publication can be addressed by early dissemination of results (e.g., in workshops, or technical venues); moreover, discovery of critical security issues should be followed by preemptive communication of the uncovered vulnerability to the respective organization (e.g., Bai et al. [21] found some malware in the Android marketplace, and communicated this to Google. With regards to inquiring the practitioners' viewpoint, this can be done via user studies.

[3] We used keywords such as "CTI," "Cyber Threat Intelligence," "Threat Intelligence," combined with "tools" or "solution."
[4] We also exclude works such as the one by Griffioen et al. [60], whose contribution is a longitudinal analysis of CTI feeds (which hence does not fall in our category of "technical" paper).
[5] Internal intelligence includes, e.g., network logs from IDS/IPS, databases, servers, routers and other network devices of an organization.

2.3.1 ***Prior user studies (Related Work).*** Some prior works have carried out user studies with security practitioners. For instance, [7] interviewed 22 SOC analysts, whereas [89] focuses on the security of Smart Grids (interviewing 14 practitioners). However, these (and other [16, 33]) papers have nothing to do with CTI. After extensively analyzing prior literature, we found only three peer-reviewed[6] papers entailing user studies with CTI practitioners. Ahmad et al. [3] carry out a user study revolving around the analysis of information security management and incident response in one organization. Zibak et al. [150] perform a review of CTI literature, and inquire 30 CTI practitioners about quality dimensions in CTI. Kotsias et al. [76] review the integration and adoption practices of CTI in one organization. Finally, we also mention the 2020 USENIX Security paper by Bouwman et al. [31], which focuses on investigating the perceived quality of "commercial threat intelligence" according to interviews with 14 security professionals. These works, despite their undeniable value, do not provide any technical contribution.

2.3.2 ***Proposed Recommendations.*** Our literature review serves to motivate our *call to action*: we seek to improve the practical value of CTI research in the real world. To this purpose, we advocate for the following:

- *Researchers should "go beyond the lab."* First, to account for the "untimely findings," researchers should proactively communicate (via means of dissemination different from research articles) any security-critical finding to the respective stakeholders; then, to account for the "lack of practitioners' input," we endorse researchers to reach out to CTI practitioners. This holds especially true for technical papers, for which we recommend to use the practitioners' know-how to validate the following design choices: *(i)* is the purpose meaningful? *(ii)* is the data source appropriate? *(iii)* is the output useful? *(iv)* is the cost manageable?
- *Practitioners should "lend a hand."* The current disconnect is not just due to researchers who are "lazy." Multiple papers (e.g., [7, 89]) have pointed out that obtaining the cooperation of security professionals can be prohibitive. Given that practitioners will *also* benefit from providing their input, we hence endorse them to be more willing to cooperate with researchers.

We acknowledge that bridging the gap requires efforts *from both research and practice*. However, it **can be done**. This is why, in the remainder of this paper, we will showcase how to embrace our recommendations[7] when *(i)* proposing a new method for CTI—questioning (Section 3.1) whether it addresses real routines; *(ii)* implementing it in a practical tool—questioning (Section 4.1) its purpose as well as the validity of the data source; *(iii)* using such a tool to analyze data from the real world, questioning (Section 5.2) the utility of its output; and *(iv)* assessing its feasibility, questioning (Section 6.1) the work time practitioners would need to accomplish a similar purpose. The answers to all these "questions" are provided in our user study (Section 7.3).

3 ANALYSIS OF UNDERGROUND FORUMS: CHALLENGES, AND OUR SOLUTION

We now shift our focus to our *technical contribution*: TopTran, an original method to facilitate the analysis of *underground forums* – a common activity for CTI practitioners [53, 106, 114, 145]. We first describe underground forums (Section 3.1), and then discuss state-of-the-art techniques to analyze such data source (Section 3.2), and their related deployment challenges (Section 3.3). Finally, we present TopTran (Section 3.4).

[6]We searched Google Scholar based on keywords such as "CTI," "Cyber Threat Intelligence," "Threat Intelligence" combined with "user survey," "expert survey," "study," "expert," "user," "expert." In addition, we also applied the snowball method [136] and reviewed the references of other papers related to CTI, included in this paper. This analysis was done by two authors. We also found only 2 white papers [34, 53], which survey current practices in CTI.
[7]Note: even though our recommendations can be broadly applied to "bridge the research-practice" gap in other domains (e.g., intrusion detection [16]), we stress that our focus is on CTI.

3.1 Overview of Underground Forums (and related NLP applications)

Underground forums can be seen as online social networks for cybercrime [96, 115] where users engage in (potentially illicit) activities, share (potentially criminal) ideas, discuss (potentially offensive) technologies and security incidents [54, 108]. Therefore, *analyzing underground forums provides proactive threat intelligence*, thereby favoring the prevention of security breaches [143]. Previous work examined underground forums from a variety of perspectives. Examples include investigation of trust relationships between mutually distrustful parties [96], analyzing the economy of darkweb shops [99], studying attackers' assets and tools [115], and identification of malicious actors [2, 58, 67, 81, 100, 105]. A recent review [8] has summarized the existing literature on underground forums.

The most common methods for the analysis of underground forums use Natural Language Processing (NLP) techniques. For example, Latent Dirichlet Allocation (LDA) [28] has been applied to identify topics discussed in certain threads [54, 120], to cluster source code [115], or to analyze user's discussion preferences [67]. More advanced NLP techniques, such as word embeddings, have also been applied for classification tasks. Examples include classifying exploit source code [11], predicting exploitability of vulnerabilities [130], and classification of threads as malicious or benign [46].

In the aforementioned works that involved classification tasks, ground truth was available, i.e., samples were associated to pre-defined sets of classes. Unfortunately, *such an assumption may not align with the real world*: such annotation must be done manually,[8] and can be prohibitively expensive [33]. Although these methods can be used in research, they may not be enticing for solutions also meant for CTI practitioners, who already spend abundant time in carrying out manual analyses; additional data labeling would further aggravate their duties. This problem is further aggravated by considering that such methods require additional labeled data for any future update, which is a necessary duty to ensure that the overarching CTI tool provides answers that reflect current trends.

3.2 Topic Modeling (state-of-the-art)

When ground truth is not available, one can resort to *unsupervised* techniques: despite still requiring a training phase, these methods do not require any labeling. The most notable family of unsupervised NLP techniques is *topic modeling* [103], designed to detect thematic structures (e.g., word/phrase patterns, or "topics") in documents. Such a property makes topic modeling methods much more advantageous than searches based on hard-coded keywords [137].

A well-known topic modeling method is LDA [28], but it cannot account for context, i.e., the relationship between adjacent words in the text. The topics in LDA can only include words contained in the training corpus, hence *LDA cannot handle unseen words by design* [27]. Furthermore, using LDA (or its variants [125]) yields topics that are often misaligned with human judgment [63], and are of poor quality for underground forums [54].

A further step in the evolution of topic modeling was to deploy the idea of *pre-trained word embeddings*[9] used for other NLP tasks [91]. Intuitively, pre-training on a large data corpus enables the capture of a multitude of contextual dependencies for any given language. The state-of-the-art is represented by Contextualized Topic Models (CTM) [27], which can handle words and documents not seen at training time [26]. Moreover, CTM can be enhanced to analyze full sentences, and not just words, by combining it with recent advances in transformers, such as SBERT [111].

Takeaway: By combining CTM with SBERT, one can *(i)* account for context and *(ii)* generalize to sentences not seen during training, thereby enabling exploratory analyses in underground forums.

[8] For instance, the authors in [46] manually annotated 10k posts, and hired 5 students to validate only 50 of them—and the margin of error was 10%.
[9] A word embedding maps phrases into high-dimensional spaces used for the final inference. This enables accounting for word context, e.g., realizing that "Air Canada" is a different notion than the set of words "Air" and "Canada" [91].

3.3 Practical Application of Topic Modeling

To appreciate our contribution, let us introduce the problems involved in practical applications of topic modeling methods to analyze underground forums. Suppose that we want to carry out an exploratory analysis, focused on a certain task (e.g., "finding threads that match a given topic"), on a given data source (e.g., an underground forum), for which no labeled data is present.

One can develop an ad-hoc topic model for the data source and then apply it to such a data source. Such an approach is not viable—due to noise. Topic modeling methods are primarily concerned with characterizing prevalent topics of a given set of documents. As an example, the authors of [29] revealed that a straightforward application of topic modeling in the *Yale Law Journal* would yield topics tagged as 'tax,' 'labor,' and 'speech'. However, finding prevalent topics in a given document corpus is not necessarily something that may interest a CTI practitioner, who may rather want to see if the corpus contains *specific* topics.[10]

We argue that one can use pre-trained embedding models (e.g., CTM and SBERT) to "find threads that match a given topic." However, a few questions arise: does the data used for such pre-training enable gathering the desired (specific) CTI insights?[11] Furthermore, how can one be certain that the "blind" application of any pre-trained model on an unknown data corpus (i.e., the underground forum) will yield accurate results?[12] Indeed, applying pre-trained models to carry out exploratory analyses in data sources such as underground forums is not straightforward. This is because we (and especially CTI practitioners!) do not know in advance "what lies in an underground forum." Depending on the given task, some training sets may be inappropriate; this is an issue given that the subjects of interest to CTI practitioners tend to be highly technical.[13]

We argue that, despite the aforementioned difficulties, applying pre-trained topic modeling methods for analyzing underground forums can be done—but two requirements must be fulfilled: *(i)* choosing a suitable dataset for fine-tuning the models, and *(ii)* pre-validating their performance before their deployment. We observe that while the first requirement can be addressed with domain expertise (there are many publicly available data sources usable for fine-tuning), the second presents subtle pitfalls that necessitate an elaborated treatment.

3.4 Our proposed solution: TopTran

Here, we describe TopTran, short for "Topic Transfer," a simple but original[14] protocol to gauge whether pre-trained topic model can be safely *transferred* to analyze unknown data (e.g., underground forums).

The key idea of TopTran, illustrated in Figure 1, is to first build a topic model for the knowledge that may interest a CTI expert from a reliable source of data (step "Train"). Subsequently, the model can be assessed on the CTI data source (potentially unrelated to the "training data") to verify if discovered topics appear in the data (step "Search"). The challenge of the search step, akin to finding a needle in a haystack, is to ensure the validity of a positive search result. To this end, TopTran entails the validation step comprising two control experiments that verify the sufficiency of the true positive and the false positive rates of the derived topic models. The details of the respective steps are elucidated in the ensuing subsections.

[10] As an intriguing (and even somewhat anecdotal) example, Goupil et al. [59] studied IT security job advertisements with the goal of revealing "key technical skills" expected by this industry, and found a large share of topics to be related to general HR issues, employer descriptions, and employee benefits—none of which being useful for the overarching task of identifying "key technical skills."

[11] For instance, using the topic model pre-trained in [29] would yield topics related to "tax," "labor," and "speech."

[12] Topic models simply return the documents in a document corpus that are the closest to the topics they are pre-trained to find—even if the overall relevance of the returned documents to such topics is low.

[13] E.g., one would not search for evidence on the potential usage of AI by attackers by using a model pre-trained on documents older than 30 years.

[14] We are not aware of existing evaluation protocols that have the same goal as ours.

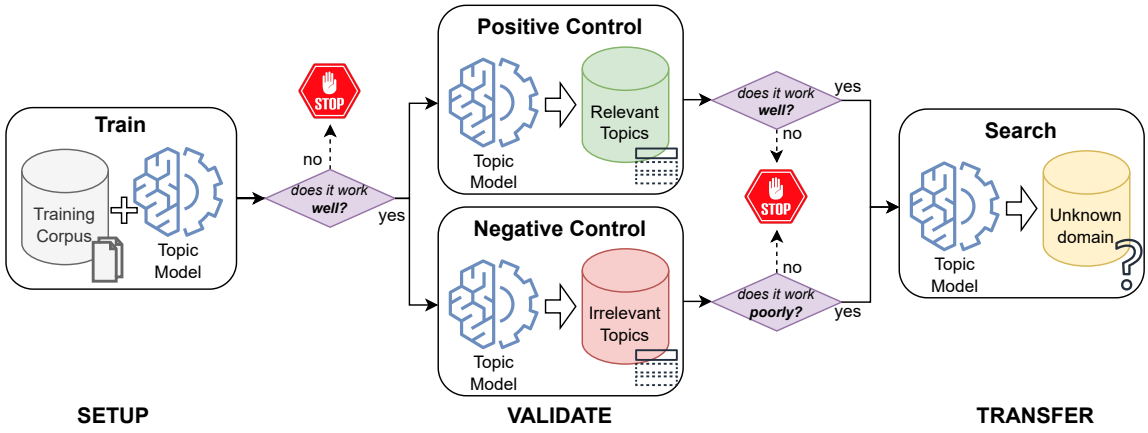


Fig. 1. **Schematic description of the TopTran protocol**. The conventional topic modeling is built based on the training corpus from a known domain and validated on two other corpora from a related and an unrelated domain. Subsequently, the model can be "transferred" for searching for topics from a known domain in the unknown one.

3.4.1 ***Intuition: From Topic Modeling to Topic Transfer.*** Topic models analyze some documents (i.e., textual data) and return the probabilities that such documents match a given topic. However, these probabilities are *normalized on the basis of the analyzed data* [27, 148]. For instance, even if the topics contained in a given document are all irrelevant for the task, a topic model would still return the "most relevant" topic. This may yield impractical findings and a waste of time when transferring the model. To address this issue, *we propose an original heuristic*. Inspired by [62], we observe that if a document is unrelated to the topics captured by the model, the probability distribution computed by the model is approximately uniform. Hence, for n topics inferred by the model, their probabilities are $\approx 1/n$. Conversely, if some topics *are* relevant for the document, their probabilities should be $\gg 1/n$ (we validate these claims empirically in Section C.2.3). Therefore, we can see the probabilities as a measure of the *relevance* of a topic in a given document: if the calculated relevance of a topic is above a certain value ρ, then such a topic can be claimed to be "relevant" for the document. Such ρ can be tuned to optimize the trade-off between true- and false-positive rates.[15] In our demonstration, we show how to calibrate ρ.

3.4.2 ***Validation method.*** Our method is a high-level evaluation protocol used to validate the practical utility of a given topic modeling method for CTI applications. To apply TopTran, the following elements are required: a topic modeling model, $\mathcal{M}$; a CTI *task*,[16] $\mathcal{T}$; the minimum *relevance*, ρ; and four datasets:

$\mathbb{T}$: a *training* dataset, which contains topics relevant for $\mathcal{T}$ (according to ρ), used to fine-tune (i.e., train) $\mathcal{M}$;
$\mathbb{N}$: a dataset for a *negative control* experiment, which should contain topics that are not relevant for $\mathcal{T}$ (according to ρ);
$\mathbb{P}$: a dataset for a *positive control* experiment, which should contain topics that are relevant for $\mathcal{T}$ (according to ρ);
$\mathbb{S}$: the dataset for the actual *search*, which will be used to investigate the CTI task $\mathcal{T}$ by means of $\mathcal{M}$.

It is implicitly assumed that $\mathbb{T} \neq \mathbb{P} \neq \mathbb{N} \neq \mathbb{S}$. These elements are used to apply our proposed evaluation protocol, TopTran, which entails the following steps (depicted in Figure 1):

[15] Note, however, that due to the intrinsic lack of ground truth which entails the application of topic modeling techniques, it may not be possible to systematically compute the false-positive or true-positive rate (even at the development phase). The purpose of ρ is simply to prevent flagging topics that, despite being potentially present in a collection of documents, are not relevant for the overarching task.
[16] We consider a CTI "task" as a collection of topics that are of interest for CTI.

(1) *Prepare.* The model $\mathcal{M}$ is trained on $\mathbb{T}$, and ρ is identified (potentially using a combination of $\mathbb{N}$, $\mathbb{P}$, or $\mathbb{T}$).
(2) *Self Check.* The model $\mathcal{M}$ is assessed on $\mathbb{T}$. The expectation is that $\mathcal{M}$ finds many relevant topics.
(3) *Positive Control.* The model $\mathcal{M}$ is assessed on $\mathbb{P}$. The expectation is that the model identifies many learned topics.
(4) *Negative Control.* The model $\mathcal{M}$ is assessed on $\mathbb{N}$. The expectation is that few matches will be found.
(5) *Search.* The model $\mathcal{M}$ is "transferred" to $\mathbb{S}$: topic transfer is performed to the unknown target domain.

Depending on the outcome of steps 2 and 3, two situations can occur: if the expectations are met, then the topics found during *Search* can be claimed to be present in $\mathbb{S}$; otherwise, any result on $\mathbb{S}$ is of dubious utility (and we recommend not to carry out any analysis). This can be due to a poor selection of $\mathbb{T}$, $\mathbb{P}$, $\mathbb{N}$; but also due to an intrinsic incompatibility between $\mathbb{T}$ or $\mathbb{S}$, or an inappropriate choice of $\mathcal{M}$ or $\mathcal{T}$.

3.4.3 ***Observations.*** We make three observations. First, TopTran is *agnostic* to *(a)* the data type, and *(b)* the topic modeling technique. In other words, TopTran can be used to assess the proficiency of any topic modeling method reliant on sentence embeddings, which must not necessarily be tested on underground forums. Second, the *benefit* of TopTran is that it warns the developer that the results may (or may not) be accurate before the analysis is carried out, saving time for CTI practitioners. Third, once a model $\mathcal{M}$ is validated via TopTran, it can be used on *any* $\mathbb{S}$ (assuming the same CTI task $\mathcal{T}$) without any retraining or revalidation (we will show this in our demonstration).

Disclaimer. The goal of TopTran is to enable the application of topic modeling for CTI. We do not claim (or aim) to "outperform existing methods." TopTran facilitates the workload of CTI experts – researchers and practitioners alike.

4 DEMONSTRATION AND IMPLEMENTATION

As a complementary part of our technical contribution, we demonstrate a practical application of TopTran. Our aim is to showcase how topic transfer can be leveraged to investigate underground forums for CTI. We first define the envisioned CTI task (Section 4.1), and then describe the datasets considered in our case study (Section 4.2). Next, we present our low-level implementation of TopTran (Section 4.3), and conclude with the validation of the results (Section 4.4).

4.1 Goal and Research Question (CTI Task)

We seek to analyze underground forums aiming to answer the research question (RQ): "*is AI on attackers' minds?*" The reasons that led us to consider such an RQ can be summarized as follows:

- **Timeliness.** The advances of artificial intelligence (AI) are apparent [15, 36, 139]. Given that such technologies can also be used by attackers (as suggested by prior works [49, 92, 128]), scrutinizing the "maturity level" of AI-powered offensive strategies is helpful to counter future threats.
- **Novelty.** To the best of our knowledge, there is no prior work which investigated our RQ by using topic modeling in underground forums.
- **Feasibility.** Our case study is likely to yield positive results: despite the lack of ground truth, there are plenty of public sources which we can use to setup our models for our envisioned topic transfer application.

Therefore, the findings related to analyzing our RQ will be a yet another contribution of this paper to the state of the art.

4.2 Datasets (description and rationale)

Recall (Section 3.4) that TopTran requires four datasets: one for training the topic model ($\mathbb{T}$); two for validation – positive ($\mathbb{P}$) and negative ($\mathbb{N}$) control experiments; and one for the actual search ($\mathbb{S}$).

- **Training: `StackEx`.** Stack Exchange is a legitimate question-and-answer discussion forum, with large communities of IT enthusiasts. To derive a representative data corpus to train our models, we considered the forums related to

"Data Science" and "AI." We collected `StackEx` by extracting (and then merging) the corresponding data dumps [126] in September 2022. Overall, `StackEx` contains 92 576 threads, representing our $\mathbb{T}$.

- **Positive control: `Kaggle`.** Kaggle is an online discussion forum that attracts users with expertise in AI/ML to participate in competitions and solve challenges. Users can download datasets, explore codebases, or discuss with other users to receive support or give feedback. We extracted the data [113] in September 2022; the dataset consisted of 242 217 threads, representing our $\mathbb{P}$.
- **Negative control: `Speeches`.** We required a corpus having no relevance with AI. We collected (from [110]) a dataset with 622 speeches given by the Presidents of the United States from 1789 to 2010, which represents our $\mathbb{N}$. We performed a keyword search to ensure that neither ML nor AI was discussed.
- **Search: `CrimeBB`.** CrimeBB is a dataset created [106] by scraping cybercriminal communities (spanning across both clearnet and darknet), making it ideal for investigating our RQ (and to use as $\mathbb{S}$). Notably, `CrimeBB` is constantly updated: each update may add data (e.g., including new forums) or modify data (e.g., change existing threads). We only considered forums that were in the English language, for which data were available starting from 2014 (i.e., when deep learning became popular [77]). We further discuss how we use `CrimeBB` in the next section (Section 5.1).

We provide detailed descriptions on how we pre-processed each dataset in Appendix C.1. We stress that we used *only* `CrimeBB` to answer our RQ; the other datasets served only to setup TopTran (see Section 3.4.2). Moreover, the datasets used for our setup (i.e., `StackEx`, `Kaggle`, `Speeches`) were chosen because, according to our own knowledge, they met the criteria for their respective role (i.e., $\mathbb{T}$, $\mathbb{P}$, $\mathbb{N}$) in the implementation of TopTran. However, in principle, any dataset that meets such criteria could have been used.

4.3 Application of TopTran (Workflow)

We outline our implementation workflow, now that we have our four required datasets ($\mathbb{T}$, $\mathbb{P}$, $\mathbb{N}$, $\mathbb{S}$) to apply TopTran.

i) Get state-of-the-art "vanilla" topic models, i.e., $\mathcal{M}$ (Section 4.3.1).
ii) Fine-tune (i.e., "train") the topic model $\mathcal{M}$ on `StackEx`, i.e., $\mathbb{T}$ (these technical analyses are covered in Appendix C.2).
iii) Derive a list of topics to search for in $\mathbb{S}$, which should be present in $\mathbb{T}$ and $\mathbb{P}$, but not in $\mathbb{N}$ (reported in Table 2).
iv) Determine an appropriate minimum relevance, i.e., ρ (Section 4.3.2)
v) Validate (i.e., "test") our $\mathcal{M}$ on `StackEx`, `Speeches`, `Kaggle`, i.e., $\mathbb{T}$, $\mathbb{N}$, and $\mathbb{P}$ (Section 4.4)
vi) Transfer $\mathcal{M}$ to `CrimeBB`, i.e., $\mathbb{S}$, and analyze its findings (Section 5).

Disclaimer. Our experiments are meant to *(i)* practically demonstrate TopTran, and *(ii)* find novel CTI insights. Despite our extensive evaluation, we do not claim our implementation to be the only way to apply TopTran.

4.3.1 ***Selection and tuning of topic models.*** We selected CTM as our topic model (for the reasons explained in Section 3.2), for which we use the variant by Bianchi et al. [27]. CTM can be integrated with pre-trained language models, with SBERT representing the state-of-the-art language model. Since we sought to analyze English data, we used a monolingual SBERT model [25]. In particular, we found an SBERT model[17] that was *also* pre-trained on Stack Exchange data, making it an optimal choice to integrate with CTM. We initially assessed our "vanilla" topic model by using it (off-the-shelf) on `StackEx`, and we found it produced good results. Figure 2 illustrates the example word cloud for the topic "Object Detection," which confirms accurate modeling of the topic.[18] We then improved the performance of our models by fine-tuning (i.e., training) them with `StackEx`, identifying the optimal parameter configurations (extensively covered

[17] More precisely, we used the open-source *paraphrase-distilroberta-base-v2* [104].
[18] At first glance, the words "yolo" and "anchor" might be confusing. In Object Detection, however, "anchor" refers to a parameter to identify irregular objects [75], whereas "yolo" is an algorithm for Object Detection [4].

in Appendix C.2). After training, we found that considering the entire list of 50 topics (reported in Table 2) yields models with the best quality. This was confirmed by both quantitative and qualitative analyses (carried out on StackEx, i.e., $\mathbb{T}$).

4.3.2 ***Calibrating the minimum relevance (ρ).*** After training, we must determine an appropriate value of ρ to ensure that the analysis will yield satisfactory results (see Section 3.4.1). Recall that TopTran assumes that $\mathbb{P}$ and $\mathbb{N}$ (and, of course, $\mathbb{T}$) are "known." Hence, we take the fine-tuned model and test it on StackEx (i.e., $\mathbb{T}$) and on Speeches (i.e., $\mathbb{N}$); we do not consider Kaggle (i.e., $\mathbb{P}$). This is because we are aware that in StackEx, there are going to be matches (as also confirmed by our preliminary evaluation and extensive fine-tuning operations); we also know that in presidential speeches there should be no relevance to AI. Therefore, we omit Kaggle from this analysis because it is not necessary, but also for fairness: we expect that our choice of ρ will lead to a positive validation also on Kaggle (discussed in the next section). Hence, for this calibration exercise, our goal is to identify a value of ρ that yields "many matches" on StackEx, and close to no matches on Speeches. For this, we examined the probability distributions of our considered topics (n=50, in Table 2) for each thread in StackEx and Speeches (shown in Figure 3, top row; each line represents a topic). We found that the topic probabilities are very low ($\ll 0.2$) for a large majority of threads (both in StackEx and Speeches). Inspired by this, we set ρ=0.2, and evaluated its effects[19] by measuring the ratio of threads (in both StackEx and Speeches) whose probability is above the minimum relevance, i.e., > 0.2 (shown in Figure 3, bottom row). The results confirm our choice: in StackEx, there are "many" topics above ρ, whereas for Speeches there are almost none. We shed further light on these results in Appendix C.2.3 (which validates our heuristic in Section 3.4.1).

Fig. 2. **Preliminary Assessment** – We ran a vanilla CTM (with SBERT) to search for the topic "object detection" on StackEx, and generated the resulting word cloud. The results confirm that our chosen topic models worked well, and can be further fine-tuned.

Table 2. **Topics and Meta-Topics considered by our topic model** – Topics (50) are found during training (they yield the best results—see Appendix C.2.2) whereas Meta-Topics (10) are derived after internal discussions.

Meta-Topic	Topics
Learning Algorithms	Genetic Algorithms, Tree-Based ML, Reinforcement Learning, Probabilistic Learning, Regression Analysis, Clustering, ML Algorithms
Neural Networks	Convolutional NN, NN, NN Modeling
NLP	Text Mining, NLP: Transformers, Text Analytics
Time Series	Sequence Model, Time Series Analysis
Model Training, Tuning, Evaluation	Model Tuning, Model Train: Data Issues, Cost Function Optimization, Classification, Setup, Model Train.: Train and Test, Model Validation, Model Tun.: Overfitting, Model Train.: NN, Model Tun.: Optimization
Data Preparation	Data Input Formatting, Feature Engineering, Feature Transformation, Data Visualization, Data Representation, Statistics, Image Prepr., Data Prepr.
Python/AI Setup	Python/AI Library Setup, AI Environment/System Setup
AI Education	Skill Requirements/Learning, AI Resources, Object Detection (Toy Prob.), Demand Forecasting (Toy Prob.)
Support Request	Support Request, Recommendation
Human-Techn. Int.	Human-Techn. Interaction

[19]We note that, since we have 50 topics (i.e., n=50), our chosen ρ=0.2 is 10 times higher than the uniform topic probability. See Section 3.4.1

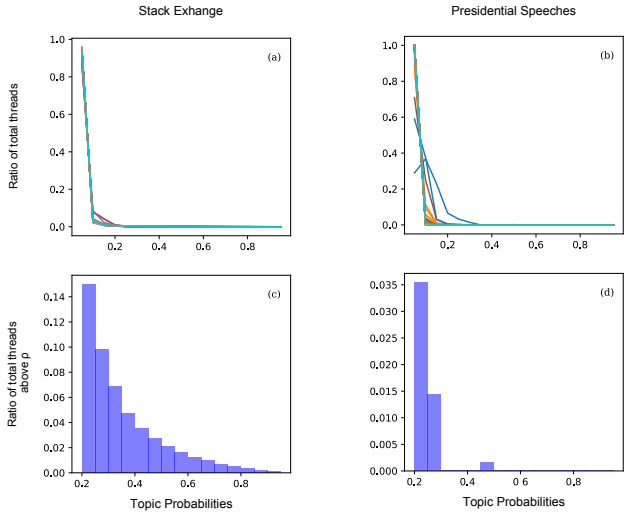


Fig. 3. **Calibration of the minimum *relevance*** (ρ) – [Top] Probability distributions for the topics (n=50, each line is a topic) in each thread, with ρ=0.2. [Bottom] Validation of our selection: for `Speeches` (`StackEx`), few (many) threads are > ρ.

4.4 Validation Results

We validated our topic models by testing them on `StackEx` ($\mathbb{T}$), `Kaggle`, ($\mathbb{P}$) and `Speeches` ($\mathbb{N}$). We note that this operation (i.e., validating our topic model by testing on `StackEx` and `Speeches`, which we used to calibrate ρ) is legitimate: we did not use these datasets to "train" or "fine-tune" our models. We simply used them to find an optimal threshold (rooted on our domain knowledge of these datasets and our proposed heuristics) that would not trigger an unmanageable number of noisy results when deploying the model for the actual investigation (which will be done on data that has never been used in the development of the topic model). We did not use `Kaggle` for calibration.

To simplify the interpretation of the results, we coalesce the 50 identified topics into 10 meta-topics (see Table 2). We further distinguish between three types of meta-topics: AI Core Topics (○); AI Supporting Topics (△), and AI Surrounding Topics (□).[20] These results are reported in Table 3, wherein rows list the meta-topic, and columns report the percentage of threads (in each dataset) for which a match was found. As listed in Table 3, we found the following:

- *Self-Check (✓).* We can see that over 50% of the threads in `StackEx` are related to AI. This result was expected.
- *Positive Control (✓).* Our fine-tuned model found ≈30% threads in `Kaggle` related to AI. This (not very surprising) result confirms that this experiment is successful.
- *Negative Control (✓).* Our fine-tuned model found that only 5% of the threads in `Speeches` are related to AI. We manually reviewed these results, which are due to noise (a more detailed analysis is in Appendix D.1.3). Given the low occurrence of "false positives," we can conclude that this experiment was also successful.

We provide an in-depth analysis of these findings in Appendix D.1.

Takeaway. Our validation was successful. Therefore, if we transfer our fine-tuned topic models to `CrimeBB`, we will find matches *if and only if* they are relevant. We can now investigate our RQ.

[20] AI Supporting Topics are less concerned about the statistical perspective of algorithms (e.g., "loss function optimization"), but rather focus on the application of AI.

Table 3. **Validation experiments.** – We report the distribution of meta-topics found in the threads of each dataset (𝕋, ℙ, ℕ).

Meta-Topic	Self-Check StackEx	Positive Kaggle	Negative Speeches
Learning Algorithms ○	11.66	2.08	0.32
Neural Networks ○	3.93	0.55	0.00
Model Training/Tuning/Eval ○	9.48	3.19	0.00
NLP ○	3.52	0.56	0.00
Time Series ○	1.81	0.47	0.00
Data Preparation △	8.29	3.89	0.00
AI Education △	6.66	5.41	0.00
Python/AI Setup △	2.28	8.68	0.00
Support Request □	0.84	3.72	0.00
Human-Tech. Interaction □	2.09	0.36	4.82
Total Percentage	**50.55**	**28.91**	**5.14**

5 INVESTIGATION AND RESULTS

The fourth contribution of our paper are the results of our exploratory analysis on underground forums, whose relevance was confirmed by CTI practitioners (see Section 4.1). We first describe how we structured our investigation (Section 5.1), and then present the results of our case studies (Section 5.2–5.4).

5.1 Design of our Exploratory Analysis

To investigate our RQ, we searched for evidence of AI-related threads among the underground forums in CrimeBB.

5.1.1 ***Approach.*** To simulate a realistic CTI scenario (and also to provide more intriguing findings), we carried out our exploratory analysis at two different points in time (see Figure 4). The first was in Sept. 2022, *before the public release of ChatGPT*. The second was in Sept. 2023, *after the release of ChatGPT*. In this way, we can also perform a temporal comparison of our results, thereby allowing one to identify trends in the attackers' interest towards AI, which is likely to have changed after the Nov. 30th, 2022 rollout of ChatGPT. Such an approach was enabled by CrimeBB's authors constantly maintaining the dataset with periodic updates using new scrapes from underground forums. We stress that our analysis is *fair*: we were not aware, in Sept. 2022, of ChatGPT's release (in a sense, we were "lucky").

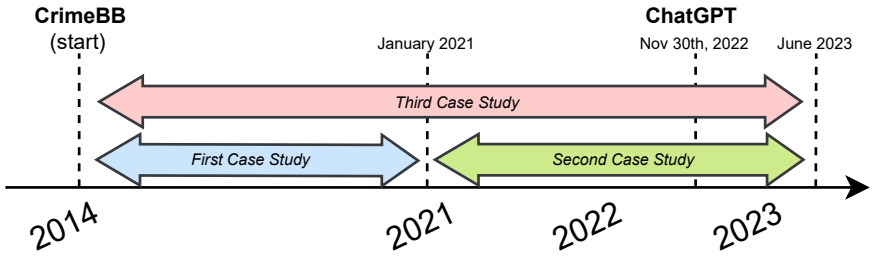


Fig. 4. **Timeline of our investigation**. We collected two snapshots of underground forums from CrimeBB [106] (one from 2014 to Jan.'21, and another from Jan.'21 to Jun.'23) which we used for our case studies. We sought to identify (and analyze) threads about AI-related topics in these communities.

5.1.2 ***Categories of Underground Forums.*** Recall that CrimeBB spans both darknet and clearnet underground forums (Section 4.2). For a fine-grained analysis, we grouped them into five categories:

- *Darknet Cybercrime Forums (DCF),* focusing on malware, hiring hackers, selling drugs, social engineering.
- *Clearnet Cybercrime Forums (CCF),* with the same focus as DCF, but for clearnet.
- *Clearnet Gaming Forums (CGF),* focusing on game cheating.
- *Clearnet Mixed Forums (CMF),* focusing on malware, advanced hacking, hacking tutorials (including games), scams.
- *Clearnet Other Forums (COF),* focusing on illegal downloads of movies, and black hat marketing techniques.

To preserve the confidentiality of CrimeBB (for which we signed a Non-disclosure Agreement), we cannot provide specific details on "who" posted any given message we found during our investigation, nor the names of these forums. Nonetheless, we can report that, overall, our categories encompass 18 individual forums.

5.1.3 ***Case Studies.*** We organized our search by devising three case studies (depicted in Figure 4). The first two leveraged TopTran: we ran the fine-tuned topic model (tasked to identify the 50 AI-related topics in Table 2) on two "snapshots" of CrimeBB. The third was done with a manual search. Let us outline the methodology and focus on each case study.

(1) **Pre-ChatGPT (Section 5.2).** We downloaded the CrimeBB dataset in Sept. 2022. This snapshot included data until January 2021, containing 18 underground forums with a total of 6.8M threads. The goal of this case study was to investigate the viewpoint of underground communities on the possibility of using AI for offensive purposes, as of January 2021.
(2) **Pre/Post-ChatGPT (Section 5.3).** We downloaded the CrimeBB data again in September 2023. As a result of an update to CrimeBB in June 2023, eight underground forums were populated with new entries. The goal of this case study was to detect potential changes in the attractiveness of AI for attackers due to the release of ChatGPT. Hence, we perform a temporal split of the CrimeBB snapshot: "before," having threads from January 2021 to November 29th 2022 (≈130K threads); and "after," having data from November 30th 2022 to June 2023 (≈3K threads).
(3) **Offensive AI tools in the wild (Section 5.4).** We scrutinized whether underground communities talked about the offensive AI tools discussed *in the literature.* To this purpose, we *(i)* identified a list of well-known AI-powered "attacks" proposed in prior work, and *(ii)* performed a keyword search on the latest CrimeBB snapshot (having data until June 2023).

In Table 4, we provide the details of the CrimeBB snapshots (threads and number of forums) of our case studies.

Table 4. **Comparison of CrimeBB datasets** – We report the # of threads (and # of forums) included in each group of underground forums (columns) for the three subsets of CrimeBB considered in our case studies.

Timespan	DCF	CCF	CGF	CMF	COF	*Total*
Jan. 2014→Jan. 2021	91 462 (4)	4 738 647 (7)	924 674 (3)	400 248 (2)	656 323 (2)	6 811 354 (18)
Jan. 2021→Nov 30th, 2022	N/A	47 808 (4)	18 263 (2)	8 808 (1)	52 889 (1)	127 768 (8)
Nov 30th, 2022→June 2023	N/A	1 039 (2)	306 (2)	848 (1)	753 (1)	2 946 (6)
Jan. 2014→June 2023	91 462 (4)	4 787 494 (7)	943 243 (3)	409 904 (2)	709 965 (2)	6 942 068 (18)

5.2 First case study (Pre-ChatGPT)

"Do underground communities talk about using AI for malicious purposes?" We ran our topic model on the CrimeBB data spanning from 2014 until January 2021; we report coverage rates (i.e., percentages of matching threads) in Table 5. We first analyze these results quantitatively (Section 5.2.1) and then underscore three relevant findings for CTI (Section 5.2.2). Detailed qualitative analyses, which also encompass discussions entailing AI-specific threats, are provided in Appendix D.2.

5.2.1 ***Overview (quantitative).*** By analyzing Table 5, we see that the coverage rates for AI-related topics in underground forums range between 10% and 20%. This reveals that such topics are subjects of discussion in these communities, but, of course, to a lesser degree than in AI-specific communities (such as those of StackEx and Kaggle). Besides the two general topics *Support Request* and *Human-Technology Interaction*, a large share of discussions focus on using or learning AI techniques, such as *Python/AI Setup* (20.94%), and *AI Education* (8.90%), followed by discussions on AI Core Topics like *Learning Algorithms* (1.61%), *Data Preparation* (0.91%), *NLP* (0.64%). Moreover, to determine if any "false positives" were raised by our topic model, we manually analyzed (by reading through all posts) the top-10 ranked discussions for each topic: we found that the only instances that had little to do with AI were a large minority (around 5%), and predominantly included those in the category of human-computer interaction (a finding which aligns with that of our validation experiment on Speeches, discussed in Appendix D.1.3).[21] Overall, we can conclude that these findings support the soundness of our implementation of TopTranwhen applied to this use-case.

Table 5. **Results of the first case study (Pre-ChatGPT)** – Cells report the coverage rate of the 50-topic model (fine-tuned on StackEx) when transferred to the underground forums in CrimeBB (until Jan.'21).

Case Study 1 (CrimeBB)	Jan. 2014→Jan. 2021				
Meta-Topic	**DCF**	**CCF**	**CMF**	**CGF**	**COF**
Learning Algorithms ○	0.04	**0.48**	0.18	**0.87**	0.04
Neural Networks ○	0.00	0.00	0.00	0.01	0.00
NLP ○	0.03	0.08	0.09	0.04	**0.40**
Time Series ○	0.00	0.00	0.00	0.00	0.00
Model Training/Tuning/Eval ○	0.01	0.02	0.02	0.08	0.00
Data Preparation △	0.04	0.38	0.09	0.29	0.12
AI Education △	**1.51**	**1.06**	**1.13**	**0.52**	**4.68**
Python/AI Setup △	**1.58**	**6.28**	**2.92**	**7.65**	**2.50**
Support Request □	**6.97**	**10.75**	**6.55**	**9.69**	**7.75**
Human-Techn. Interaction □	**0.63**	**1.51**	0.23	**0.66**	**0.80**
Total Percentage	**10.80**	**20.55**	**11.22**	**19.81**	**16.31**

5.2.2 ***Intriguing Findings.*** By analyzing the threads found by our topic model, we made three intriguing discoveries.

- *Lack of innovation.* Users in CrimeBB seem to favor the adoption of existing (and ready-to-use) AI techniques/tools rather than developing new ones. We found a substantial amount of requests inquiring for support on well-known AI libraries (e.g., scikit-learn, PyTorch, Pandas).
- *Low technical expertise.* We found it striking that the amount of Support Requests in underground communities is 10x larger than in AI-specific communities, cf. 41.71% in CrimeBB (Table 5) with 0.84% in StackEX and 3.72% in

[21] We do not attempt to estimate the "false-positive rate" of our topic model, given the size of our considered datasets and the lack of ground truth.

`Kaggle` (Table 3). We speculate that this is due to an overall low technical expertise of most underground forums' users (who sometimes receive responses such as "Google it, this kind of sh*t is all over the surface net").

- *(Offensive) AI-as-a-Service.* Users in `CrimeBB` ask for "AI services" which, at those points in time, were not provided in the underground economy. This is a problem: the more attackers request for a certain service, the more likely a similar service will be released on the market [69].

Nevertheless, our analysis reveals that (at least as of 2021) most "attackers" were just approaching the "offensive AI" domain. This is in stark contrast with some claims made by contemporary researchers (e.g., [92], whose first submission dated-back to 2021) which warned about sophisticated adversarial threats.

Takeaway. Underground communities talk about AI. However, most discussions entail simple applications of AI.

5.3 Second case study (Pre/Post-ChatGPT)

"Did the situation change after January 2021 (and what are the effects of ChatGPT)?" We ran our topic model (the exact same one used for the first case study) *only* on the new `CrimeBB` data added in the June 2023 update. We distinguish "before ChatGPT" (Jan. 2021→Nov. 29th, 2022) and "after ChatGPT" (Nov. 29th, 2022→June 2023). We report the coverage rates in Table 6 (note that DCF is excluded because none of those forums were updated; see Table 4). We summarize the results (Section 5.3.1) and then present intriguing findings (Section 5.3.2). Detailed analyses are in Appendix D.3.

Table 6. **Results of the second case study (Pre/Post-ChatGPT)** – We compare the coverage rate of the 50-topic model (fine-tuned on `StackEx`) transferred to the underground forums in `CrimeBB` ("Before ChatGPT" and "After ChatGPT")

Case Study 2 (`CrimeBB`)	Jan. 2021→Nov.29th, 2022				Nov 30th, 2022→Jun. 2023			
Meta-Topic	**CCF**	**CMF**	**CGF**	**COF**	**CCF**	**CMF**	**CGF**	**COF**
Learning Algorithms ○	0.22	0.09	**1.53**	0.06	0.00	0.24	**1.96**	0.13
Neural Networks ○	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
NLP ○	0.08	0.18	0.01	0.33	0.00	0.12	0.00	**0.53**
Time Series ○	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
Model Train/Tuning/Eval ○	0.08	0.09	0.05	0.01	0.29	0.12	0.00	0.00
Data Preparation △	0.21	0.14	0.27	0.05	0.00	0.12	0.00	0.00
AI Education △	**0.55**	0.11	**0.49**	**0.65**	**0.58**	0.12	0.33	0.40
Python/AI Setup △	**4.33**	**6.46**	**13.69**	**1.90**	**4.23**	**1.77**	**8.82**	**1.73**
Support Request □	**7.78**	**6.73**	**9.75**	**10.86**	**4.62**	**4.01**	**10.78**	**13.94**
Hum-Tech. Int. □	**0.61**	0.09	0.37	**1.21**	**1.51**	0.00	0.00	0.27
Total Percentage	**13.86**	**13.90**	**26.19**	**15.09**	**10.88**	**6.49**	**21.90**	**17.00**

5.3.1 ***Overview (quantitative).*** By observing Table 6, we see that our topic model found matches in both subsets: on average, there are matches for 17.26% of the threads in the "before" subset, and for 14.06% of the "after" subset. As we did for the first case study (Section 5.2.1), we carried out a (small-scale) manual validation of these matches, and the outcome was similar, i.e., the majority were indeed relevant to AI. By comparing our results over the temporal axis, we notice some changes over the findings of our first case study (cf. Table 5 with Table 6). For instance, 13.69% of the threads in CGF that matches the *Python/AI Setup* topic during the Jan. 2021→Nov.29th, 2022 timeframe, whereas the percentage was 7.65% during Jan. 2014→Jan. 2021. A statistical test confirms the change to be statistically significant ($p \approx 0$). Intriguingly, the percentage decreases to 8.82% *after* the rollout of ChatGPT: this may indicate that users of underground communities

are using ChatGPT to setup their development environments. There are also other statistically significant "decreasing" trends. For instance, in CCF, there are 20.55% overall matches for January 2014→January 2021, which goes down to 13.86% in January 2021→November 29th, 2022, and further drop to 10.88% during November 30th, 2022→June 2023. This may indicate that the overall interest towards AI in CCF is relatively decreasing, being replaced by different and more enticing subjects for its users.

5.3.2 ***Intriguing Findings.*** We highlight three intriguing findings derived from our manual review.

- *AI-powered evasion.* After ChatGPT, we saw many discussions entailing the (ab)use of AI for highly technical tasks, such as circumventing CAPTCHAs, or building malware with ChatGPT (one user wrote "I just made ChatGPT build me a python malware that resulted in 0/58 on virus total").
- *Breaking AI.* After ChatGPT, many threads discuss possibilities to "break" AI-based systems.
- *Future of malware.* Before ChatGPT, we found a "philosophical" conversation on whether AI will better benefit the defensive or offensive side of malware in the long run.

Altogether, our analyses confirm our hypotheses during the first case study: "attackers" *are* turning their attention to AI-techniques, which are facilitated by the easiness of some "cheaply" available solutions, such as ChatGPT. Finally, a pivotal finding is that we obtained all these results by using *the exact same* topic model used for the first case study. This denotes the power of topic model transfer for CTI, which is enabled by our proposed TopTran.

Takeaway. The interest towards AI has changed in the last two years. The topic model developed with our TopTran can be used in the future with no effort to gather AI-related CTI.

5.4 Third case study (offensive AI tools in the wild)

A recent paper claimed that "Real attackers don't compute gradients," [14] but the evidence supporting this claim was scarce. Hence, we question whether users of underground forums *at least mention* the names of sophisticated AI-powered attacks proposed in prior work (thereby indicating that real attackers "may" use gradients!).

5.4.1 ***Method.*** We are interested in "Offensive AI tools," i.e., techniques or software toolkits that *(i)* have unique names, *(ii)* rely on AI, and *(iii)* have been (or can be) used to carry out cyberattacks. As a starting point, we looked for tools mentioned in papers about offensive AI [35, 61, 73, 92, 138], and expanded our analysis by looking at references on Google Scholar. We only included entries for which the authors gave a name[22] (e.g., PassFlow [101]). The final list of offensive AI tools, reported in Table 7, includes research papers (e.g., [147]), industry publications (e.g., [88]), and even repositories (e.g., [45]). After identifying 19 tools, we performed a manual keyword search covering the entire `CrimeBB` data (from 2014 until June 2023), and then inspected any thread that yielded a match.

5.4.2 ***Results.*** Out of the 19 tools, only 3 yielded some matches. We found the following:

- EagleEye: we had many hits for this term in Gaming forums. However, manual review revealed that such hits were *all false positives* (unrelated to TopTran): the discussions were about a Playstation3 game converter that enables the usage of a keyboard and mouse to play instead of a traditional controller.
- DeepLocker: we had less than 5 hits, entailing discussions about the capabilities of this tool.
- Lyrebird: we also had less than 5 hits. Discussions were about its usage for eWhoring and identity obfuscation.

Takeaway. Underground forum communities scarcely discuss the offensive AI tools proposed in prior works.

[22] From this, it follows that we had to exclude a high number of papers, such as [107].

Table 7. **Offensive AI tools (third case study)**– We list the 19 offensive AI tools identified during our literature review, including the publication date and their intended (or potential) offensive purpose.

Offensive AI Tool	Pub. Date	Purpose
CodeQL [43]	2007	Vulnerability Discovery
PhishGen [102]	2015	Social Engineering/Phishing
DeepDGA [13]	2016	Fake Domain Generation
DeepGenerator [45]	2017	Web Appl. Hacking
DeepHack [109]	2017	Web Appl. Hacking
EagleEye [50]	2017	Social Engineering
LyreBird [85]	2017	Social Engineering
DeepLocker [127]	2018	Evasive Malware
GyoiThon [88]	2019	Web Appl. Hacking
PassGAN [64]	2019	Password Attacks
uriDeeP [132]	2019	Fake Domain Generation
attackGAN [147]	2021	Evasive Malware
CyberBattleSim [90]	2021	Autonomous Agent
CyberEvo [95]	2022	Autonomous Agent
IDSGAN [84]	2022	Evasive Malware
MalwareGAN [66]	2022	Evasive Malware
PassFlow [101]	2022	Password Attacks
BlackMamba [122]	2023	Evasive Malware
EyeSpy [123]	2023	Cognitive Threat Agent

6 FURTHER ANALYSES AND REFLECTIONS

So far, we have presented TopTran (Section 3), implemented it in a practical tool (Section 4), and used it to investigate a meaningful RQ for real-world CTI. Here, we further substantiate our contributions. We first perform a an assessment of the costs to develop TopTran (Section 6.1), and then demonstrate the impracticality of using ChatGPT as an alternative to topic transfer (Section 6.2). Next, we carry out an ablation study to justify TopTran (Section 6.3).

6.1 Cost Assessment (how much time does it require?)

To investigate our RQ, we had to invest resources, including *time*. We kept track of the time-expenditure required to accomplish most of the operations required for our implementation of TopTran.

- *Qualitative analyses.* We manually labeled the word clouds of the 50 identified topics. The initial annotations for each of the 50 topics were separately conducted by three experts (duration 3h each) and then discussed as a group (duration 1h). For 45 of the 50 topics, experts suggested the same label, in some cases with differences in language. For the remaining 5 topics, differences in the labeling were discussed and resolved by consensus. Therefore, labeling of the 50 topics took a total of 12 human working hours.
- *Definition of meta-topics.* The definition of the 10 meta-topics (from the 50 topics) entailed three experts. For the definition, each expert coded their own meta-topics (for 30m). Then, a group discussion (of 1h) was held to derive the final list. This totaled 4.5 human working hours.
- *Computational runtime.* To carry out all the experiments discussed in this paper (training and testing of the topic models), our server (see Appendix C for the hardware specifications) required ≈51h of computation. Specifically, it took: 1h to analyze `StackEx`; 1h for `Kaggle`; 0.5h for `Speeches`, 48.5h for `CrimeBB` (with 45h for the first case study, and 3.5h for the second case study).

Hence, in terms of overall human working hours, "setting-up-the-stage" to investigate our RQ required 16.5 hours (≈2 workdays). As demonstrated with our case studies (Section 5), our fine-tuned topic model (developed thanks to our proposed TopTran) can be used without any additional human effort to analyze new data streams.

6.2 ChatGPT: an alternative? (negative result)

Since the release of ChatGPT the question of "could it also be done with ChatGPT?" almost naturally comes up. Therefore, we briefly illustrate why we cannot use ChatGPT for the tasks performed in this paper (at least "today"). To do this, we performed a proof-of-concept experiment; we present a technical analysis.

6.2.1 ***Experiment.*** We scrutinize whether we can use ChatGPT for a crucial part of our analysis: Identify 50 topics related to AI in discussion forums, which we do via topic modeling. We asked ChatGPT to identify the 50 most prevalent topics in AI discussion forums, such as "Stack Exchange AI" and "Stack Exchange Data Science." We instructed ChatGPT (using GPT Turbo 3.5) to provide the 50 AI topics requested, including the top keywords defining each topic, as well as the percentages that describe the relevance of the keyword for the topic. We performed the following step-by-step analysis (on Oct. 19th, 2023), whose results are publicly observable (currently provided as an HTML file in our repository [42]). The steps we took, and the output we received for each, were:

(1) When we asked ChatGPT to perform the aforementioned task for the first time, we only got five topics related to AI. We had to ask ChatGPT two more times to get 50 topics (and not five).
(2) When ChatGPT finally provided 50 AI-related topics, only the first four are directly related to AI, and from Topic 5 onward the results only listed topics like: "AI in Real Estate" (which appears twice) or "AI in agriculture." Also, most percentages *exceed* 100%.
(3) When we asked ChatGPT again and raised our concern that these topics are unlikely to be the most prevalent topics in those AI forums, it said "I apologize" and *made the same mistake again.*

Disclaimer. This is only an attempt to use ChatGPT (in Q3 2023) for the tasks envisioned in our investigation. We do not claim any more recent versions of ChatGPT (or other commercial models) to yield the same results.

6.2.2 ***Analysis.*** There are no other alternatives to obtain the 50 most prevalent AI topics from ChatGPT: We cannot upload a large file for analysis and we cannot copy paste the text from Stack Exchange AI and Stack Exchange Data Science due to their size (>90 000 threads). The limit of input that can be provided to ChatGPT has also been highlighted in previous research [119]. According to OpenAI, the maximum input length is limited to 4 096 tokens (≈3 000 words) or 16, 385 tokens (≈12 300 words) for GPT-3.5, considering that one token is on average 1.33 words. For GPT-4 the maximum input length is 25 000 words. Therefore, *ChatGPT cannot be used for the topic modeling task* and, for sure, not for the analysis of underground forums – our `CrimeBB` dataset had, at the time we carried out our analysis (Q4 2023), over 7M threads.

6.2.3 ***Other uses.*** Previous research also raised concerns regarding keywords on specific topics provided by ChatGPT [119]. However, they also highlight that it is difficult to objectively evaluate such results. After all, we do not know which *exact data* ChatGPT uses to determine the 50 AI-related topics. ChatGPT can, however, be used for the interpretation of topics [112]. Although it is not a "one-size-fits-all" solution, ChatGPT can reduce the time it takes to label topics created by topic modeling, and some research even shows that it creates better topic labels than humans [82].[23]

6.3 Ablation Study

For our first two case studies (Section 5.2 and Section 5.3) we used the fine-tuned topic model developed through our TopTran (see Section 4). A legitimate question arises: "is TopTran necessary *in practice*?" Indeed, setting the minimum

[23] We did not leverage ChatGPT for topic labeling in our study since we performed the labeling with three experts to ensure human-based scientific results. However, in practice, the use of ChatGPT can dramatically decrease labeling time: we will do so for our ablation study (Section 6.3)

relevance (ρ), fine-tuning (on StackEx) and validating (on Kaggle and Speeches) requires some effort (quantified in Section 6.1). Hence, to justify the contribution of TopTran, we carried out an ablation study. We measured the performance of the "vanilla" topic model when not performing any changes on the original model from [27], i.e., keeping the default parameters and removing the minimum relevance ρ. The performance was assessed both on StackEx and on CrimeBB, and we compared it with the performance of the topic model obtained with TopTran, both qualitatively and quantitatively (refer to Appendix C.2.1 for a detailed explanation, which we also used to setup TopTran). We proceed as follows:

- *Quantitative comparison on* StackEx. The common quantitative criteria to gauge a topic model is the topic quality (TQ), which is expressed as the product of the normalized pointwise mutual information [30] with the topic diversity [48]. Higher TQ implies a better topic model. By running our vanilla topic model on StackEX, we obtained TQ=0.072, whereas TopTran yielded a model with TQ=0.096, showing that TopTran provides a factual improvement.
- *Qualitative comparison on* StackEx. The vanilla model identified 10 topics in StackEx, resulting in a lower granularity than the 50 topics identified by the fine-tuned model. We used ChatGPT for topic labeling and reviewed the labels manually, as suggested in Section 6.2. Manual analysis of these 10 topics showed that *(i)* the topics are very general, e.g., the topics "Computing Resources" or "Machine Learning Metrics"; and *(ii)* the words used to characterize these topics lack specificity, thereby leading to ambiguity in the interpretation of the results. These findings highlight that lack of fine-tuning leads to a (very likely) worse topic model, which is impractical for analyzing our RQ. Moreover, from a broader viewpoint, the lack of specificity of the topics found by the vanilla topic model suggests that such a topic model is unlikely to be of much utility for CTI-related tasks.
- *Real-world assessment on CrimeBB*. Lastly, we applied the vanilla topic model to an excerpt of CrimeBB. The excerpt includes 25 forum threads for which the model fine-tuned with TopTran identified an AI-related topic and 25 forum threads for which it did not identify any topic. The vanilla model identified one topic for each of the 50 provided threads, despite 25 of the threads from CrimeBB being unrelated to AI (i.e., these are all false positives). This emphasizes the importance of the minimum relevance ρ introduced in TopTran.

Takeaway. Using TopTran ensures that transferring topic models yields practical results (e.g., low false positives).

7 USER STUDY WITH PRACTITIONERS

A complementary contribution of this paper is the expert survey conducted with real practitioners of CTI. As identified by Ainslie et al. [5], researchers need to bridge the theory-practice gap by collecting evidence from practitioners. Besides serving as a validation of our claims, our survey is also a source of *inspiration* for future work, shedding additional insight in the practical domain.

7.1 Description and Design

Our user study is rooted on transparency and fairness. We will now describe how we performed it.

Eligibility Criteria. It is notoriously difficult to obtain the cooperation of professionals (especially in cybersecurity) for research purposes (e.g. [7, 33, 89]). Hence, we recruited our participants via convenience sampling [55] (as also done in [16]). We looked for practitioners with experience in CTI-related roles. To increase diversity, we reached out to practitioners associated with various organizations, spanning across both the public and private sectors, and including both small and large companies. For instance, we contacted practitioners with experience in *(i)* CTI research, *(ii)* law enforcement, and *(iii)* CTI services. We did not offer any form of compensation, and participation was voluntary. We

only reached out to practitioners with more than 5 years of experience in CTI-related roles. To ensure *fairness*, we were unaware of the participants' opinions on any of the subjects touched by our user study beforehand.

Survey Design. We carried out our user study during Oct. and Nov. 2023. During this time frame, we contacted CTI practitioners (via email, or private messaging) and asked them to participate in an anonymous online questionnaire. The questionnaire did not have any time limit. After informing the participant about the goals of the study, the first three questions focused on demographics. Then, 11 closed (and 2 optional open) questions related to CTI followed. The questionnaire (available in our repository [42]) has four pages:

(1) *Demographics* (gender, work area, and knowledge of CTI).
(2) Rating of the *interest on our research question* (used to drive our investigation in Section 5).
(3) *Cyber threat analysis*: How would the practitioner investigate our RQ, how long it would take, and whether the output of our analysis and tool would be of interest to them.
(4) Rating of practitioner's *opinion on academic research* in CTI.

To avoid duplicate answers, we asked each participant to give us feedback after they filled the questionnaire.

Timeline. Importantly, we carried out our user study *after* having done our analyses, but *before* writing this paper. This is why, when writing this paper, we aligned it with the practitioners' viewpoints, but our results are due to our own efforts as researchers.

7.2 Sample Description.

The final population includes 24 practitioners, 83% male and 17% female. 79% of the survey participants work in a role related to CTI, and 21% work in a broader cybersecurity role (despite having experience in CTI). 54% of the survey participants consider their CTI knowledge as expert level ("I have extensive knowledge and experience in CTI"), 25% as advanced level ("I have a solid understanding of CTI and its practices"), and 21% as intermediate level ("I have a basic understanding of CTI"). Nobody identified as a beginner. Our practitioners work in organizations located in the USA, the UK, and the EU. The participants have worked or do work at public law enforcement, SOC, or world-leading cybersecurity companies and/or have a reputable track record of publications at industry conferences in the domain. To preserve participants' privacy, we cannot reveal more information about our sample.

7.3 Main Results (what do practitioners say?)

The following are the major results of our user study, which we use to *validate* the various claims made throughout our paper (see also end of Section 2). We provide additional results in Appendix A.

- **What do practitioners think about research?** We inquired our participants if they "...ever turn to research papers to improve [their] expertise or discover new tools/methods." While 12 (50%) answered that they do so regularly,[24], the remainder have a more skeptical opinion of research, stating that "*there is a disconnect between research papers and the reality*," that "*academic papers can become obsolete due to publishing time*" (stated twice), and also that research "*is not practical/necessary as no attackers are doing that* [...] *it is theory that might be used in the future*."
- **Do CTI practitioners perform manual analyses?** Yes. In our user study, 21 (87.5%) participants resort to manual analyses for analyzing forums (potentially by coding simple scripts). Only 3 (12.5%) leverage full-fledged tools.
- **Is our chosen CTI purpose meaningful?** Yes. According to our user study, 23 (95.8%) CTI practitioners considered our RQ[25] to be of interest. Only 1 (4.2%) rated it as not very interesting.

[24] Even these, however, stated that "*papers on threat actors are more interesting than technical papers*" and "*academic papers are not super useful.*"
[25] The actual phrasing of our RQ in the questionnaire was "what cybercriminals think about using Artificial Intelligence (AI) for their malicious deeds."

- **Is our data-source appropriate?** Yes. In our user study, 23 (95.8%) CTI practitioners would have considered[26] "underground forums" for our RQ, with "darknet websites" being the second most popular option (20 votes).
- **How costly would it be in practice?** We inquired CTI practitioners how much time it would take to investigate our RQ: 18 (75%) reported one week or more; 4 (17%) two to four days; 2 (8%) stated it would take one day (see Figure 7).
- **Is our tool interesting?** We inquired CTI practitioners whether they would be interested in "a tool that identifies discussion trends among cybercriminals in underground forums." Out of 24 participants, 21 (87.5%) stated they were interested, with 14 (58%) reporting to be "very interested" (see Figure 8).
- **Is our visualized output useful?** We presented a snippet of a "draft" version of Table 5 to CTI practitioners, and asked whether they would find the insights to be of interest. The opinion varied, with half of our participants leaning towards a yes, and half towards a no.[27] (see Figure 9)

Altogether, the results of our user studies can serve to validate our research. For instance, practitioners agree that research and practice in CTI may be disconnected; that our research question is interesting; that our data source is appropriate; and that even though only half of our participants found our output to be of use, their comments (i.e., that they would rather look at trends) supported the idea of carrying out the analysis discussed in our second case study (in Section 5.3). We stress that the participants of our user study were not aware of our investigations.

8 DISCUSSION, LIMITATIONS, AND FUTURE WORK

We now perform a reflective exercise wherein we scrutinize our "contributions." We pinpoint limitations, highlight practical implications, and identify avenues for future work.

- **Literature review:** To ensure broad coverage, two authors extensively reviewed and critically analyzed related literature. However, we acknowledge that we may have overlooked some works that met our inclusion criteria (and which should have been considered in Table 1).
- **Performance and quality:** We performed extensive parameter searches (see Appendix C), but we acknowledge that our model could have been further tuned; moreover, we did not manually review *every* thread identified by the model developed via TopTran (overall, our datasets include over 7M threads; see Section 4.2 and Table 4). We acknowledge that more manual analyses and further tuning could have led to "better" results (albeit at higher implementation costs; see Section 6.1). As we stated in Section 4.3, we do not claim our implementation to be the best, but our results, confirmed by our qualitative analyses (see Appendix D), suggest that our transferred topic model works well.
- **Data source and RQ:** The intent of our RQ is to investigate what *attackers* want to do with AI. However, users of underground forums in `CrimeBB` are not necessarily "attackers." Yet, our survey revealed that even CTI practitioners would have looked into underground forums for investigating our RQ (see Section 4.1). Furthermore, we consider `CrimeBB` due to it being widely recognized in the research community. However, we acknowledge that there exists other sources that could be used to investigate our RQ. For instance, it is possible to scrape additional underground forums, as was done in [120], whose proposed scraping tool is, unfortunately, not publicly available.
- **Utility of our findings**: We do not claim that our tool can provide "actionable" CTI over a single run (as also hinted by one of our participants; see Section 5). To provide "actionable" CTI in the context of analyzing underground forums, topic transfer is useful if run over time, so that it allows an analyst to identify trends. Our case studies

[26] Intriguingly, only 8 (33%) CTI practitioners knew `CrimeBB`—which is only available to researchers upon explicit request to its creators.
[27] However, among the *nos*, some reported that "they would look at trends on fora." Our second case study (in Section 5.3) shows that TopTran enables these analyses.

revealed that our topic model retains its performance even after years, thereby making it appropriate for long-term deployment (without any sort of maintenance). Nevertheless, our tool still enables one to ascertain whether "it makes sense" to even carry out some (manual) analyses: if it yields no matches (which can be done without any manual effort), then it would be pointless to even begin an investigation.

- **User study:** A total of 24 CTI practitioners participated in our survey. Such a population does not allow one to make general claims about the whole spectrum of CTI practitioners, but we never made this claim. Regardless, it is acknowledged that user studies with security practitioners are hard (e.g., two recent USENIX Security papers interviewed 22 and 14 practitioners, respectively [7, 31]).
- **Applicability of our method:** At a high level, our proposed methodology can be used in two orthogonal ways. On the one hand, one can develop a given topic model by fine-tuning it on AI-related topics (like we did) and, if it passes the validation of TopTran, then use such a model to analyze a wide array of data-sources of similar nature that are of interest for CTI (e.g., multiple underground forums including those not included in `CrimeBB`, and also certain Telegram channels or Discord servers which may be frequented by malicious actors [65]). On the other hand, one can develop another topic model by fine-tuning it on data that have little in common with AI (e.g., quantum computing) and, if it passes the validation of TopTran, then use such a model to analyze one or more data sources with proven traces of malicious actors. Regardless, it is implicitly assumed that, to extract CTI, our model should be used on data sources that are frequented by malicious actors (such as underground forums). For instance, using our fine-tuned topic model on Stack Exchange may find some topics, but such findings would hardly have any value for CTI given that the users in this platform are unlikely to be malicious.
- **Practical implementation of our tool:** In this work, we have proposed TopTran, and we have used it to carry out an exemplary (but original) analysis of underground forums for CTI "as academic researchers." However, our tool *can be used by industry practitioners*, too. Let us explain how this can be done. First, the implementation of TopTran requires access to underground forum data, which practitioners working on these topics do have (as also evidenced by our user study in Section 7.3). Next, practitioners need to define their topic of interest (such as AI in our work) and identify legitimate discussion forums focusing on their topic of interest (in our work: StackExchange "Data Science" and "AI," i.e. the `StackEx` dataset, see Section 4.2). StackExchange offers a wide array of different sub-forums and may thus also represent a valuable source for various technical topics (beyond AI, e.g., "Quantum Computing") that can be attractive for cybercriminals. Then, practitioners should apply TopTran by following the steps proposed in Section 3.4: Setup, Validation, and Transfer. Based on our work focusing on AI-related discussions in underground forums (see Section 6.1), we can estimate that practitioners can implement TopTran within two workdays. Once implemented, TopTran allows practitioners to track any changes in discussions over time, and without any additional human effort. This facilitates a strategic analysis of changes in the threat landscape.

In this paper, we made many "disclaimers" but we want to make one last one to avoid harmful misunderstandings:

Disclaimer. In our recommendations (Section 2.3.2), we advocate for future research to account for the practitioners' viewpoint. Yet, from a scientific perspective, a "technical paper" does not need to undergo any form of validation by practitioners to be valuable. However, doing so would increase the impact of the paper's contributions.

9 ETHICAL CONSIDERATIONS

We discuss the precautions we took to ensure that our research complies with established ethical standards, and then elaborate on a potential "ethical dilemma" that may trouble future researchers attempting to carry out similar efforts.

We consulted with our institutions' IRB to determine any ethical concerns related to the data and research protocols used. We were granted an exemption from IRB oversight, since all data were anonymized by the source and the analysis was considered "secondary research." For our user study, we followed the Menlo report [22]: our participants were transparently informed of the purpose of our questionnaire, and the participation was voluntary. Their data is anonymous, and they know our identity so they can ask us to delete their responses. We cannot disclose actual data from `CrimeBB` due to NDA with its creators [106].

Nevertheless, when tackling topics similar to ours, a researcher may wonder: *is it ethical to use data from underground forums for CTI?* Research may be regarded as ethical when the potential benefits surpass the possible harms [106]. In this paper, we do not scrape the data ourselves, but we analyze the data provided from `CrimeBB`. Unfortunately, obtaining informed consent from individuals (who typically use handles or pseudonyms to conceal their real-world identity) participating in underground forums with illegal activities is not feasible [106]. According to established ethical principles and guidelines, such as the statement of ethics of the British Society of Criminology [98], informed consent may not be explicitly required when the analyzed data are publicly accessible on the Internet, and the focus of the research lies on the analysis of collective behavior, without the aim of identifying particular members [106]. In our work, the objective is to extract information about novel attack traits by analyzing discussions in underground forums to allow defenders to stay up to date with offensive strategies and ultimately prevent cybercrime. Thus, in this context and with these premises, employing underground forums for CTI is ethically justifiable from a research viewpoint.

Finally, for responsible disclosure, we disseminated our most critical findings prior to submission (following our recommendations in Section 2.3.2).

10 CONCLUSIONS

By scrutinizing prior literature on Cyber Threat Intelligence (CTI), we found that the viewpoint of practitioners tends to be overlooked. This is not acceptable given that CTI is meant to be used by security analysts to protect real assets.

In this paper, we took the first step towards aligning research and practice in CTI. We proposed an original method, TopTran, that enables the application of "topic modeling" to analyze unknown data in underground forums. We then implement TopTran in a proof-of-concept tool, for which we released the source code [42]. Next, we used our tool to investigate whether communities of underground forums talk about AI, potentially for offensive purposes. We found scarce interest in sophisticated AI-powered attacks. We also show that our tool retains its capabilities over time and requires no maintenance. All these contributions are connected to a user study that we carried out with 24 CTI practitioners, which we used as a guide to demonstrate the relevance, validity and interest of our research.

We endorse future endeavors in CTI to embrace our overarching message. Researchers should reach out to CTI practitioners and inquire about their routines. Practitioners should facilitate the collaboration with researchers and share their insights.

ACKNOWLEDGMENTS

This research was partially funded by the Hilti Foundation and the Research Foundation at the City University of New York (RFCUNY). We thank the Cambridge Cyber Crime Centre for data access to CrimeBB.

REFERENCES

[1] Victor Adewopo, Bilal Gonen, and Festus Adewopo. 2020. Exploring Open Source Information for Cyber Threat Intelligence. In *IEEE International Conference on Big Data*. IEEE, 2232–2241.

[2] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. Doppelgänger Finder: Taking Stylometry to the Underground. In *IEEE Symposium on Security and Privacy*. IEEE, 212–226.

[3] Atif Ahmad, Kevin C Desouza, Sean B Maynard, Humza Naseer, and Richard L Baskerville. 2020. How integration of cyber security management and incident response enables organizational learning. *Journal of the Association for Information Science and Technology* 71, 8 (2020), 939–953.

[4] Tanvir Ahmad, Yinglong Ma, Muhammad Yahya, Belal Ahmad, Shah Nazir, and Amin ul Haq. 2020. Object detection through modified YOLO neural network. *Scientific Programming* 2020, 1 (2020).

[5] Scott Ainslie, Dean Thompson, Sean Maynard, and Atif Ahmad. 2023. Cyber-threat intelligence for security decision-making: a review and research agenda for practice. *Computers & Security* 132 (2023).

[6] Bader Al-Sada, Alireza Sadighian, and Gabriele Oligeri. 2023. MITRE ATT&CK: State of the art and way forward. arXiv:2308.14016

[7] Bushra A Alahmadi, Louise Axon, and Ivan Martinovic. 2022. 99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms. In *31st USENIX Security Symposium*. 2783–2800.

[8] Abdullah Albizri, Alaa; Nehme, and Antoine Harfouche. 2022. A Systematic Review on Using Hacker Forums on the Dark Web for Cyber Threat Intelligence. In *Americas Conference on Information Systems, AMCIS 2022 TREOs. 92.*

[9] Mahathir Almashor, Ejaz Ahmed, Benjamin Pick, Jason Xue, Sharif Abuadbba, Raj Gaire, Shuo Wang, Seyit Camtepe, and Surya Nepal. 2023. Unraveling threat intelligence through the lens of malicious URL campaigns. In *18th Asian Internet Engineering Conference*. 78–86.

[10] Mohammed Almukaynizi, Ericsson Marin, Eric Nunes, Paulo Shakarian, Gerardo I Simari, Dipsy Kapoor, and Timothy Siedlecki. 2018. DARKMENTION: A Deployed System to Predict Enterprise-Targeted External Cyberattacks. In *IEEE International Conference on Intelligence and Security Informatics*. IEEE, 31–36.

[11] Benjamin Ampel, Sagar Samtani, Hongyi Zhu, Steven Ullman, and Hsinchun Chen. 2020. Labeling Hacker Exploits for Proactive Cyber Threat Intelligence: A Deep Transfer Learning Approach. In *IEEE International Conference on Intelligence and Security Informatics*. IEEE, 1–6.

[12] Priya Anand, Jungwoo Ryoo, Hyoungshick Kim, and Eunhyun Kim. 2016. Threat Assessment in the Cloud Environment: A Quantitative Approach for Security Pattern Selection. In *10th International Conference on Ubiquitous Information Management and Communication*. 1–8.

[13] Hyrum S Anderson, Jonathan Woodbridge, and Bobby Filar. 2016. DeepDGA: Adversarially-tuned domain generation and detection. In *ACM Workshop on Artificial Intelligence and Security*. 13–21.

[14] Giovanni Apruzzese, Hyrum S Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. 2023. "Real Attackers Don't Compute Gradients": Bridging the Gap between Adversarial ML Research and Practice. In *IEEE Conference on Secure and Trustworthy Machine Learning*. IEEE, 339–364.

[15] Giovanni Apruzzese, Pavel Laskov, Edgardo Montes de Oca, Wissam Mallouli, Luis Brdalo Rapa, Athanasios Vasileios Grammatopoulos, and Fabio Di Franco. 2023. The role of machine learning in cybersecurity. *Digital Threats: Research and Practice* 4, 1 (2023), 1–38.

[16] Giovanni Apruzzese, Pavel Laskov, and Johannes Schneider. 2023. SoK: Pragmatic Assessment of Machine Learning for Network Intrusion Detection. In *8th European Symposium on Security and Privacy*. IEEE, 592–614.

[17] Marco Arazzi, Dincy R Arikkat, Serena Nicolazzo, Antonino Nocera, Mauro Conti, et al. 2023. NLP-based techniques for cyber threat intelligence. (2023). arXiv:2311.08807

[18] Nolan Arnold, Mohammadreza Ebrahimi, Ning Zhang, Ben Lazarine, Mark Patton, Hsinchun Chen, and Sagar Samtani. 2019. Dark-net Ecosystem Cyber-Threat Intelligence (CTI) tool. In *IEEE International Conference on Intelligence and Security Informatics*. IEEE, 92–97.

[19] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. 2022. Dos and Don'ts of Machine Learning in Computer Security. In *31st USENIX Security Symposium*. 3971–3988.

[20] Yan Lin Aung, Martín Ochoa, and Jianying Zhou. 2022. ATLAS: A Practical Attack Detection and Live Malware Analysis System for IoT Threat Intelligence. In *International Conference on Information Security*. Springer, 319–338.

[21] Chongyang Bai, Qian Han, Ghita Mezzour, Fabio Pierazzi, and VS Subrahmanian. 2019. DBank: Predictive behavioral analysis of recent Android banking trojans. *IEEE Transactions on Dependable and Secure Computing* 18, 3 (2019), 1378–1393.

[22] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. 2012. The Menlo Report. *IEEE Security & Privacy* 10, 2 (2012), 71–75.

[23] Sean Barnum. 2012. Standardizing cyber threat intelligence information with the structured threat information expression (stix). *Mitre Corporation* (2012).

[24] Markus Bayer, Tobias Frey, and Christian Reuter. 2023. Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence. *Computers & Security* 134 (2023), 103430.

[25] Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Mono and multi-lingual embeddings. https://contextualized-topic-models.readthedocs.io/en/latest/language.html#mono-and-multi-lingual-embeddings. Accessed: June 2, 2023.

[26] Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: contextualized document embeddings improve topic coherence. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.

[27] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

[28] David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. *Advances in Neural Information Processing Systems* 14 (2001).

[29] David M Blei. 2012. Probabilistic topic models. *Communications of ACM* 55, 4 (2012), 77–84.

[30] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *International Conference of the German Society for Computational Linguistics and Language Technology*, Vol. 30. 31–40.

[31] Xander Bouwman, Harm Griffioen, Jelle Egbers, Christian Doerr, Bram Klievink, and Michel Van Eeten. 2020. A different cup of {TI}? the added value of commercial threat intelligence. In *29th USENIX security symposium (USENIX security 20)*. 433–450.

[32] Xander Bouwman, Victor Le Pochat, Pawel Foremski, Tom Van Goethem, Carlos H Gañán, Giovane CM Moura, Samaneh Tajalizadehkhoob, Wouter Joosen, and Michel Van Eeten. 2022. Helping Hands: Measuring the Impact of a Large Threat Intelligence Sharing Community. In *31st USENIX Security Symposium*. 1149–1165.

[33] Tobias Braun, Irdin Pekaric, and Giovanni Apruzzese. 2024. Understanding the Process of Data Labeling in Cybersecurity. In *ACM Symposium on Applied Computing*. 1596–1605.

[34] Rebekah Brown and Pasquale Stirparo. 2022. *Cyber threat intelligence survey.* Technical Report. SANS.

[35] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The malicious use of Artificial Intelligence: forecasting, prevention, and mitigation. arXiv:1802.07228

[36] Fabrício Ceschin, Marcus Botacin, Albert Bifet, Bernhard Pfahringer, Luiz S Oliveira, Heitor Murilo Gomes, and André Grégio. 2020. Machine learning (in) security: A stream of problems. *Digital Threats: Research and Practice* 5, 1 (2020), 1–32.

[37] Daiki Chiba, Mitsuaki Akiyama, Takeshi Yagi, Kunio Hato, Tatsuya Mori, and Shigeki Goto. 2018. DomainChroma: Building actionable threat intelligence from malicious domain names. *Computers & Security* 77 (2018), 138–161.

[38] Daiki Chiba, Takeshi Yagi, Mitsuaki Akiyama, Toshiki Shibahara, Takeshi Yada, Tatsuya Mori, and Shigeki Goto. 2016. DomainProfiler: Discovering Domain Names Abused in Future. In *46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 491–502.

[39] Collins Uchenna Chimeleze, Norziana Jamil, Roslan Ismail, and Kwok-Yan Lam. 2021. A Review on malware variants detection techniques for threat intelligence in resource constrained devices: existing approaches, limitations and future direction. In *Advances in Cyber Security*. Springer.

[40] Geumhwan Cho, Jusop Choi, Hyoungshick Kim, Sangwon Hyun, and Jungwoo Ryoo. 2019. Threat modeling and analysis of voice assistant applications. In *Information Security Applications*. Springer, 197–209.

[41] Kim-Kwang Raymond Choo. 2011. The cyber threat landscape: Challenges and future research directions. *Computers & Security* 30, 8 (2011), 719–731.

[42] cti-offensiveai 2025. *cti-offensiveai: Github Repository.* https://github.com/jseideman/cti_offensiveai

[43] Oege De Moor, Mathieu Verbaere, Elnar Hajiyev, Pavel Avgustinov, Torbjorn Ekman, Neil Ongkingco, Damien Sereni, and Julian Tibble. 2007. Keynote address: QL for source code analysis. In *Seventh IEEE International Working Conference on Source Code Analysis and Manipulation*. IEEE, 3–16.

[44] Ashok Deb, Kristina Lerman, and Emilio Ferrara. 2018. Predicting cyber-events by leveraging hacker sentiment. *Information* (2018).

[45] DeepGenerator 2017. DeepGenerator. https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/Generator Accessed: November 22, 2023.

[46] Isuf Deliu, Carl Leichter, and Katrin Franke. 2017. Extracting Cyber Threat Intelligence from Hacker Forums: Support Vector Machines versus Convolutional Neural Networks. In *2017 IEEE International Conference on Big Data*. IEEE, 3648–3656.

[47] Isuf Deliu, Carl Leichter, and Katrin Franke. 2018. Collecting Cyber Threat Intelligence from Hacker Forums via a Two-stage, Hybrid Process using Support Vector Machines and Latent Dirichlet Allocation. In *2018 IEEE International Conference on Big Data*. IEEE, 5008–5013.

[48] Adji B Dieng, Francisco J R Ruiz, and David Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics* 8 (2020), 439–453.

[49] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, and Anand Jeyaraj et al. 2023. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71 (2023), 102642.

[50] Eagle Eye 2018. Eagle Eye. https://github.com/ThoughtfulDev/EagleEye. Accessed: June 10, 2024.

[51] Mohammadreza Ebrahimi, Sagar Samtani, Yidong Chai, and Hsinchun Chen. 2020. Detecting Cyber Threats in non-English Hacker Forums: an Adversarial Cross-lingual Knowledge Transfer Approach. In *IEEE Security and Privacy Workshops*. IEEE, 20–26.

[52] Mohammadreza Ebrahimi, Mihai Surdeanu, Sagar Samtani, and Hsinchun Chen. 2018. Detecting Cyber Threats in non-English Dark Net Markets: A Cross-lingual Transfer Learning Approach. In *IEEE International Conference on Intelligence and Security Informatics*. IEEE, 85–90.

[53] Jared Ettinger. 2019. *Cyber intelligence tradecraft report: The state of cyber intelligence practices in the United States.* Technical Report. Carnegie Mellon University: Software Engineering Institute.

[54] Yong Fang, Yusong Guo, Cheng Huang, and Liang Liu. 2019. Analyzing and identifying data breaches in underground forums. *IEEE Access* (2019).

[55] Alison Galloway. 2005. *Non-Probability Sampling.* Elsevier.

[56] Peng Gao, Xiaoyuan Liu, Edward Choi, Bhavna Soman, and Chinmaya Mishra et al. 2021. A system for automated open-source threat intelligence gathering and management. In *International Conference on Management of Data*.

[57] Dimitrios Georgoulias, Jens Myrup Pedersen, Alice Hutchings, Morten Falch, and Emmanouil Vasilomanolakis. 2023. In the market for a Botnet? An in-depth analysis of botnet-related listings on Darkweb marketplaces. In *Symposium on Electronic Crime Research*.

[58] Joobin Gharibshah and Michalis Faloutsos. 2019. Extracting actionable information from security forums. In *World Wide Web Conference*. 27–32.

[59] Francois Goupil, Pavel Laskov, Irdin Pekaric, Michael Felderer, Alexander Dürr, and Frederic Thiesse. 2022. Towards understanding the skill gap in cybersecurity. In *27th ACM Conference on on Innovation and Technology in Computer Science Education*. 477–483.

[60] Harm Griffioen, Tim Booij, and Christian Doerr. 2020. Quality evaluation of cyber threat intelligence feeds. In *Applied Cryptography and Network Security*. Springer, 277–296.

[61] Blessing Guembe, Ambrose Azeta, Sanjay Misra, Victor Chukwudi Osamor, Luis Fernandez-Sanz, and Vera Pospelova. 2022. The emerging threat of ai-driven cyber attacks: A review. *Applied Artificial Intelligence* 36, 1 (2022).

[62] Dan Hendrycks and Kevin Gimpel. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

[63] Geoffrey E Hinton and Russ R Salakhutdinov. 2009. Replicated Softmax: an Undirected Topic Model. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 22.

[64] Briland Hitaj, Paolo Gasti, Giuseppe Ateniese, and Fernando Perez-Cruz. 2019. PassGAN: A Deep Learning Approach for Password Guessing. In *Applied Cryptography and Network Security*. Springer, 217–237.

[65] how discord 2024. How Discord is Abused for Cybercrime. https://intel471.com/blog/how-discord-is-abused-for-cybercrime Accessed: 20-05-2025.

[66] Weiwei Hu and Ying Tan. 2022. Generating Adversarial Malware Examples for Black-box Attacks Based on GAN. In *International Conference on Data Mining and Big Data*. Springer, 409–423.

[67] Cheng Huang, Yongyan Guo, Wenbo Guo, and Ying Li. 2021. HackerRank: Identifying key hackers in underground forums. *International Journal of Distributed Sensor Networks* 17, 5 (2021).

[68] Jiajia Huang, Min Peng, Pengwei Li, Zhiwei Hu, and Chao Xu. 2020. Improving biterm topic model with word embeddings. *World Wide Web Conference* (2020).

[69] Keman Huang, Michael Siegel, and Stuart Madnick. 2018. Systematically understanding the cyber attack business: A survey. *Comput. Surveys* (2018).

[70] Martin Husák, Tomáš Jirsík, and Shanchieh Jay Yang. 2020. SoK: Contemporary issues and challenges to enable cyber situational awareness for network security. In *International Conference on Availability, Reliability and Security*.

[71] Zhengjie Ji, Edward Choi, and Peng Gao. 2022. A knowledge base question answering system for cyber threat knowledge acquisition. In *IEEE International Conference on Data Engineering*. IEEE.

[72] Hyeonseong Jo, Yongjae Lee, and Seungwon Shin. 2022. Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text. *Computers & Security* (2022).

[73] Nektaria Kaloudi and Jingyue Li. 2020. The AI-based Cyber Threat Landscape: A Survey. *Comput. Surveys* 53, 1 (2020), 1–34.

[74] Eunsoo Kim, Kuyju Kim, Dongsoon Shin, Beomjin Jin, and Hyoungshick Kim. 2018. CyTIME: Cyber Threat Intelligence ManagEment framework for automatically generating security rules. In *13th International Conference on Future Internet Technologies*. 1–5.

[75] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. 2020. FoveaBox: Beyound Anchor-Based Object Detection. *IEEE Transactions on Image Processing* 29 (2020), 7389–7398.

[76] James Kotsias, Atif Ahmad, and Rens Scheepers. 2023. Adopting and integrating cyber-threat intelligence in a commercial organisation. *European Journal of Information Systems* 32, 1 (2023), 35–51.

[77] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* (2015).

[78] Martin Lee. 2023. *Cyber Threat Intelligence*. John Wiley & Sons.

[79] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies* 105 (2017), 28–42.

[80] Valentine Legoy, Marco Caselli, Christin Seifert, and Andreas Peter. 2020. Automated Retrieval of ATT&CK Tactics and Techniques for Cyber Threat Reports. (2020). arXiv:2004.14322

[81] Gastón L'huillier, Hector Alvarez, Sebastián A Ríos, and Felipe Aguilera. 2011. Topic-based social network analysis for virtual communities of interests in the dark web. *ACM SIGKDD Explorations Newsletter* 12, 2 (2011), 66–73.

[82] Dai Li, Bolun Zhang, and Yimang Zhou. 2023. Can Large Language Models (LLM) label topics from a topic model? SocArXiv. https://doi.org/10.31235/osf.io/23x4m

[83] Xiaojing Liao, Kan Yuan, XiaoFeng Wang, Zhou Li, Luyi Xing, and Raheem Beyah. 2016. Acing the IOC Game: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence. In *ACM Conference on Computer and Communications Security*. 755–766.

[84] Zilong Lin, Yong Shi, and Zhi Xue. 2022. Idsgan: Generative adversarial networks for attack generation against intrusion detection. In *Pacific-asia conference on knowledge discovery and data mining*. Springer, 79–91.

[85] Lyrebird 2024. Lyrebird - Descript. https://www.descript.com/lyrebird Accessed: June 1, 2024.

[86] Ericsson Marin, Mohammed Almukaynizi, and Paulo Shakarian. 2020. Inductive and deductive reasoning to assist in cyber-attack prediction. In *10th Annual Computing and Communication Workshop and Conference*. IEEE, 262–268.

[87] Daisuke Mashima, Derek Kok, Wei Lin, Muhammad Hazwan, and Alvin Cheng. 2020. On design and enhancement of smart grid honeypot system for practical collection of threat intelligence. In *13th USENIX Workshop on Cyber Security Experimentation and Test*.

[88] Masafumi Masuya, Toshitsugu Yoneyama, and Isao Takaesu. 2019. GyoiThon: Next generation penetration test tool. https://github.com/gyoisamurai/GyoiThon. Accessed: May 20, 2024.

[89] Jacqueline Meyer and Giovanni Apruzzese. 2022. Cybersecurity in the smart grid: practitioners' perspective. In *Annual Industrial Control Systems Security Workshop*.

[90] Microsoft Defender Research Team. 2021. CyberBattleSim. https://github.com/microsoft/cyberbattlesim. Accessed: April 15, 2024.
[91] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* (2013).
[92] Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, and Deng Gelei et al. 2023. The threat of offensive ai to organizations. *Computers & Security* (2023).
[93] Ajay Modi, Zhibo Sun, Anupam Panwar, Tejas Khairnar, and Ziming Zhao et al. 2016. Towards automated threat intelligence fusion. In *International Conference on Communications in Computing*. IEEE.
[94] Felipe Moreno-Vera, Mateus Nogueira, Cainã Figueiredo, Daniel S Menasché, and Miguel Bicudo et al. 2023. Cream skimming the underground: identifying relevant information points from online forums. In *International Conference on Cyber Security and Resilience*. IEEE, 66–71.
[95] Stephen Moskal, Erik Hemberg, and Una-May O'Reilly. 2022. CyberEvo: Evolutionary search of knowledge-based behaviors in a cyber attack campaign. *Genetic and Evolutionary Computation Conference Companion* (2022), 2168–2176.
[96] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M Voelker. 2011. An analysis of underground forums. In *ACM SIGCOMM Conference on Internet Measurement*. 71–80.
[97] Eric Nunes, Paulo Shakarian, and Gerardo I Simari. 2018. At-risk system identification via analysis of discussions on the darkweb. In *Symposium on Electronic Crime Research*. IEEE.
[98] British Society of Criminology. 2015. Statement of ethics. http://www.britsoccrim.org/ethics/ Accessed: March 2, 2025..
[99] Kris Oosthoek, Mark Van Staalduinen, and Georgios Smaragdakis. 2023. Quantifying Dark Web Shops' Illicit Revenue. *IEEE Access* 11 (2023), 4794–4808. https://doi.org/10.1109/ACCESS.2023.3235409
[100] Rebekah Overdorf, Carmela Troncoso, Rachel Greenstadt, and Damon McCoy. 2018. Under the Underground: Predicting Private Pnteractions in Underground Forums. arXiv:1805.04494
[101] Giulio Pagnotta, Dorjan Hitaj, Fabio De Gaspari, and Luigi V Mancini. 2022. Passflow: guessing passwords with generative flows. In *IEEE International Conference on Dependable Systems and Networks*. IEEE, 251–262.
[102] Sean Palka and Damon McCoy. 2015. Dynamic phishing content using generative grammars. In *IEEE International Conference on Software Testing, Verification and Validation Workshops*. IEEE, 1–8.
[103] Christos H Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. 1998. Latent semantic indexing: A probabilistic analysis. In *ACM Symposium on Principles of Database Systems*. 159–168.
[104] Paraphrase Data 2022. Paraphrase Data. https://www.sbert.net/examples/training/paraphrases/README.html. Accessed: June 2, 2023.
[105] Sergio Pastrana, Alice Hutchings, Andrew Caines, and Paula Buttery. 2018. Characterizing Eve: Analysing Cybercrime Actors in a Large Underground Forum. In *Research in Attacks, Intrusions, and Defenses*. Springer, 207–227.
[106] Sergio Pastrana, Daniel R Thomas, Alice Hutchings, and Richard Clayton. 2018. CrimeBB: enabling cybercrime research on underground forums at scale. In *World Wide Web Conference*.
[107] Will Pearce, Nick Landers, and Nancy Fulda. 2020. Machine learning for offensive security: sandbox classification using decision trees and artificial neural networks. In *Intelligent Computing*. Springer.
[108] Ildiko Pete, Jack Hughes, Andrew Caines, Anh V Vu, Harshad Gupta, Alice Hutchings, Ross Anderson, and Paula Buttery. 2022. PostCog: A tool for interdisciplinary research into underground forums at scale. In *IEEE European Symposium on Security and Privacy Workshops*. IEEE, 93–104.
[109] Dan Petro and Ben Morris. 2007. Weaponizing machine learning: humanity was overrated anyway. In *DEF CON*.
[110] Presidential Speeches: Downloadable Data 2022. Presidential Speeches: Miller Center of Public Affairs. data.millercenter.org. Accessed: November 2, 2022.
[111] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.
[112] Emil Rijcken, Floortje Scheepers, Kalliopi Zervanou, Marco Spruit, Pablo Mosteiro, and Uzay Kaymak. 2023. Towards Interpreting Topic Models with ChatGPT. In *20th World Congress of the International Fuzzy Systems Association (IFSA)*.
[113] Megan Risdal and Timo Bozsolik. 2022. Meta Kaggle. https://doi.org/10.34740/KAGGLE/DS/9 Accessed: November 22, 2022.
[114] Sagar Samtani, Maggie Abate, Victor Benjamin, and Weifeng Li. 2020. Cybersecurity as an industry: A cyber threat intelligence perspective. *The Palgrave Handbook of International Cybercrime and Cyberdeviance* (2020).
[115] Sagar Samtani, Ryan Chinn, and Hsinchun Chen. 2015. Exploring hacker assets in underground forums. In *IEEE Intelligence and Security Informatics*. IEEE.
[116] Sagar Samtani, Weifeng Li, Victor Benjamin, and Hsinchun Chen. 2021. Informing cyber threat intelligence through dark Web situational awareness: The AZSecure hacker assets portal. *Digital Threats: Research and Practice* 2, 4 (2021), 1–10.
[117] Sagar Samtani, Hongyi Zhu, and Hsinchun Chen. 2020. Proactively identifying emerging hacker threats from the dark web: A diachronic graph embedding framework (d-gef). *ACM Transactions on Privacy and Security* 23, 4 (2020), 1–33.
[118] Soumajyoti Sarkar, Mohammad Almukaynizi, Jana Shakarian, and Paulo Shakarian. 2019. Mining user interaction patterns in the darkweb to predict enterprise cyber incidents. *Social Network Analysis and Mining* 9, 1 (2019), 57.
[119] Mark Scanlon, Frank Breitinger, Christopher Hargreaves, Jan-Niclas Hilgert, and John Sheppard. 2023. ChatGPT for digital forensic investigation: The good, the bad, and the unknown. *Forensic Science International: Digital Investigation* 46 (2023).

[120] Matthias Schäfer, Markus Fuchs, Martin Strohmeier, Markus Engel, Marc Liechti, and Vincent Lenders. 2019. BlackWidow: Monitoring the Dark Web for Cyber Security Information. In *11th International Conference on Cyber Conflict (CyCon)*, Vol. 900. IEEE, 1–21.

[121] Hyejin Shin, WooChul Shim, Jiin Moon, Jae Woo Seo, Sol Lee, and Yong Ho Hwang. 2020. Cybersecurity Event Detection with New and Re-emerging Words. In *ACM Asia Conference on Computer and Communications Security (AsiaCCS)*. 665–678.

[122] Jeff Sims. 2023. BlackMamba: AI-synthesized, polymorphic keylogger with on-the-fly program modification. Hyas Research.

[123] Jeff Sims. 2023. EyeSpy Proof-of-Concept. Hyas Research.

[124] Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Niklas Elmqvist, and Leah Findlater. 2017. Evaluating visual representations for topic understanding and their effects on manually generated topic labels. *Transactions of the Association for Computational Linguistics* 5 (2017), 1–16.

[125] Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

[126] Stack Exchange Data 2022. Stack Exchange Data Dump. https://archive.org/details/stackexchange. Accessed: November 22, 2022.

[127] Marc P Stoecklin, Jiyong Jang, and Dhilung Kirat. 2018. DeepLocker: how AI can power a stealthy new breed of malware. *Security Intelligence* (2018).

[128] Nan Sun, Ming Ding, Jiaojiao Jiang, Weikang Xu, and Xiaoxing Mo et al. 2023. Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives. *Communications Surveys & Tutorials* (2023).

[129] Zhibo Sun, Carlos E. Rubio-Medrano, Ziming Zhao, Tiffany Bao, Adam Doupé, and Gail-Joon Ahn. 2019. Understanding and Predicting Private Interactions in Underground Forums. In *Ninth ACM Conference on Data and Application Security and Privacy (CODASPY '19)*. ACM, 303–314. https://doi.org/10.1145/3292006.3300036

[130] Nazgol Tavabi, Palash Goyal, Mohammed Almukaynizi, Paulo Shakarian, and Kristina Lerman. 2018. Darkembed: Exploit prediction with neural language models. In *AAAI Conference on Artificial Intelligence*, Vol. 32.

[131] Wiem Tounsi and Helmi Rais. 2018. A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Computers & Security* 72 (2018), 212–233.

[132] UriDeep 2019. UriDeep. https://github.com/mindcrypt/uriDeep Accessed: April 20, 2024.

[133] Tala Vahedi, Benjamin Ampel, Sagar Samtani, and Hsinchun Chen. 2021. Identifying and categorizing malicious content on paste sites: a neural topic modeling approach. In *IEEE Intelligence and Security Informatics*. IEEE.

[134] Christopher J C H Watkins and Peter Dayan. 1992. Technical Note: Q-Learning. *Machine Learning* (1992).

[135] Ryan Williams, Sagar Samtani, Mark Patton, and Hsinchun Chen. 2018. Incremental hacker forum exploit collection and classification for proactive cyber threat intelligence: An exploratory study. In *IEEE Intelligence and Security Informatics*. IEEE.

[136] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *18th International Conference on Evaluation and Assessment in Software Engineering*. 1–10.

[137] Tien-Hsuan Wu, Ben Kao, Felix Chan, Anne SY Cheung, Michael MK Cheung, Guowen Yuan, and Yongxi Chen. 2021. Semantic search and summarization of judgments using topic modeling. In *Legal Knowledge and Information Systems*. IOS Press, 100–106.

[138] Muhammad Mudassar Yamin, Mohib Ullah, Habib Ullah, and Basel Katt. 2021. Weaponized AI for cyber attacks. *Journal of Information Security and Applications* (2021).

[139] Helin Yang, Kwok-Yan Lam, Liang Xiao, Zehui Xiong, and Hao Hu et al. 2022. Lead federated neuromorphic learning for wireless edge artificial intelligence. *Nature Communications* (2022).

[140] Lu-Xing Yang, Pengdeng Li, Xiaofan Yang, and Yuan Yan Tang. 2018. A risk management approach to defending against the advanced persistent threat. *IEEE Transactions on Dependable and Secure Computing* 17, 6 (2018), 1163–1172.

[141] Wenzhuo Yang and Kwok-Yan Lam. 2020. Automated Cyber Threat Intelligence Reports Classification for Early Warning of Cyber Attacks in Next Generation SOC. In *Information and Communications Security*. Springer, 145–164.

[142] Wenzhuo Yang and Kwok-Yan Lam. 2020. Automated Cyber Threat Intelligence Reports Classification for Early Warning of Cyber Attacks in Next Generation SOC. In *International Conference on Information Systems*. Springer, 145–164.

[143] Azene Zenebe, Mufaro Shumba, Andrei Carillo, and Sofia Cuenca. 2019. Cyber Threat Discovery from Dark Web. *International Conference on Software Engineering and Data Engineering* (2019), 174–183.

[144] Huixia Zhang, Guowei Shen, Chun Guo, Yunhe Cui, and Chaohui Jiang. 2021. Ex-action: Automatically extracting threat actions from cyber threat intelligence report based on multimodal learning. *Security and Communication Networks* (2021).

[145] Yiming Zhang, Yujie Fan, Yanfang Ye, Liang Zhao, and Chuan Shi. 2019. Key Player Identification in Underground Forums over Attributed Heterogeneous Information Network Embedding Framework. In *28th ACM International Conference on Information and Knowledge Management*. 549–558.

[146] Jun Zhao, Qiben Yan, Jianxin Li, Minglai Shao, and Zuti He et al. 2020. TIMiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data. *Computers & Security* (2020).

[147] Shuang Zhao, Jing Li, Jianmin Wang, Zhao Zhang, Lin Zhu, and Yong Zhang. 2021. attackGAN: Adversarial Attack against Black-box IDS using Generative Adversarial Networks. *Procedia Computer Science* 187 (2021), 128–133.

[148] Weizhong Zhao, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC Bioinformatics*, Vol. 16. Springer, 1–10.

[149] Ziming Zhao, Gail-Joon Ahn, Hongxin Hu, and Deepinder Mahi. 2012. SocialImpact: Systematic analysis of underground social dynamics. In *European Symposium on Research in Computer Security*. Springer.

[150] Adam Zibak, Clemens Sauerwein, and Andrew C Simpson. 2022. Threat intelligence quality dimensions for research and practice. *Digital Threats: Research and Practice* 3, 4 (2022), 1–22.

Appendix A USER STUDY: RESULTS

In Figure 5 to Figure 9 we show the responses to the questions included in our user study. We cannot include open answers to protect the privacy of our participants. However, we reviewed the tools listed by our participants (summarized in Table 8). Moreover, we report below our "re-elaborated" version of some comments we received.

The single participant who was **not interested in our RQ** stated that they were interested in attack monetization and coordination.

Participants who found our **output visualization to be not interesting** stated that quantitative analyses are not useful, and that results should be pre-filtered, for example, for threat actor reputation; another stated that their job only entails qualitative analyses (for which even a single positive hit would be precious); another reported that discussions are not actionable because sophisticated threat actors are not active therein; another stated that a single discussion with a powerful threat actor could be a more relevant find than simple percentages.

Participants who commented on **why they do not read research papers** stated that they would rather go to venues; or that there is a disconnection between research and reality; or that data analysis is less important than the fact itself; two participants stated that research papers become obsolete quickly; another stated that research papers are not practical.

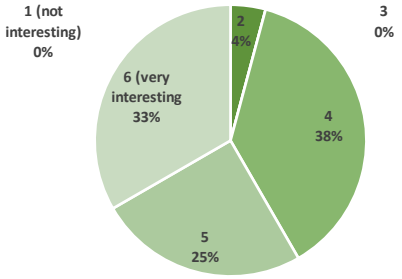


Fig. 5. Would the question of what cybercriminals think about using AI for their malicious deeds be of interest to you as a CTI analyst?

Table 8. **CTI tools listed by practitioners in the user survey** – For each tool, we report: its data sources; whether underground forums are analyzed; whether ML/AI is used, and whether details on the exact use case and or application of ML/AI are provided.

Tool	Data Sources	Udg. For. Incl.	Use of ML/AI?	Any details?
CyberSixGill	Clear, deep and darknet sources, incl. forums and markets, messaging groups, code repositories, paste sites	Yes	Yes	No
FlashPoint	Social media, messaging apps, finished intelligence, deep and darknet data, threat actor communities	Yes	Yes	No
Ceberus	Deep, and darknet sources	Yes	No	No
DarkIQ	Darknet marketplaces, forums, and onion sites, code repositories, chats, CVEs, domains, phishing sites	Yes	Yes	No
Intel471 - Titan	Darknet marketplaces, forums, and chat rooms	Yes	Yes	No
Maltego	Social media, company data, other OSINT data; darknet data only as integration	No	No	No
DarkOwl	Deep and darknet sources, incl. forums, markets, and chat rooms, Telegram	Yes	Yes	No

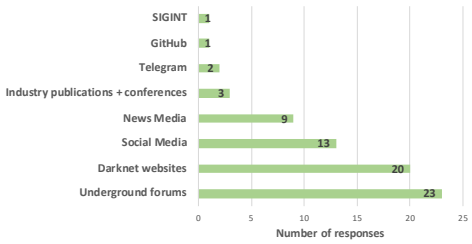


Fig. 6. In most cases, as an analyst, you have access to different data sources. Which of the following sources would you select to investigate what cybercriminals think about AI?

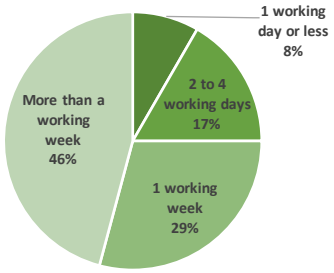


Fig. 7. How long would it take to investigate this question and make a confident statement about what cybercriminals think about using AI for their purposes? Also account for the time used by your colleagues.

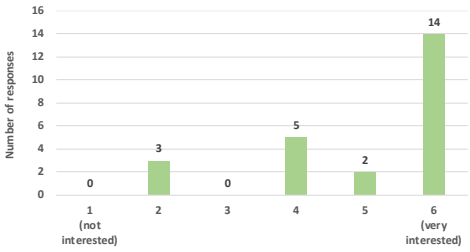


Fig. 8. Would you be interested in a tool that identifies discussion trends among cybercriminals in udg. forums or similar sources?

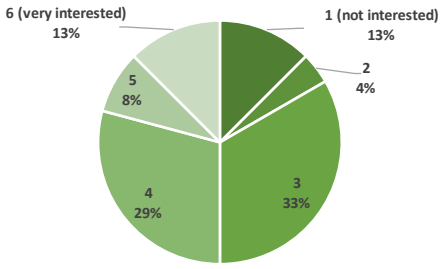


Fig. 9. The results table (which is similar to Table 5) describes different topics related to AI and the percentages of posts discussing the specific topics in underground forums. Would this insight be of interest to you?

Appendix B BACKGROUND ON CTI

Although the term CTI is commonly used in the cybersecurity domain, practitioners and researchers have a varying understanding of the term [5, 78, 80]. We denote CTI as the process and outcome of collecting and analyzing cyber threat-related data to *(i)* understand the motivation, targets, and attack behaviors of adversaries, and *(ii)* empower short- and long-term decision making of organizations. CTI should not be confused with Threat Modeling, a proactive approach to identify security risks in systems, e.g. voice assistants [40], or assessing threats, e.g. in cloud environments [12].

CTI can be divided into three types: *strategic*, *operational*, and *tactical* (or *technical*) [5, 78, 80, 114, 131]. *Strategic CTI* refers to the analysis of longer-term changes in adversary behavior to understand future risks.[28] *Operational CTI* refers to short- or medium-term changes in adversary behavior to manage day-to-day priorities. *Tactical CTI* is the most actionable, describing current threats, e.g., in the form of IP addresses or file hashes [80].

The collection of CTI data is a complex process that ideally combines different data sources [6], such as in [93]. CTI data can be extracted from security blogs, as in [83], cyber threat reports, as in [80], DNS logs, as in [38], URLs, as in [9], honeypots, as in [20, 87], underground forums, as in [129, 149], or similar.

Some works even focus on the "next step" of applying CTI in an organization, e.g. how to best defend against malicious domain names [37] or how to aggregate CTI data from external sources to generate automatic security rules [74]. Another common research area in CTI focuses on predicting future threats, such as in [38].

Appendix C TECHNICAL DETAILS

We provide further details of our implementation. We ran our experiments on a 32 core/64 thread Xeon Gold 6226R at 2.90GHz equipped with 384 GiB of RAM and an NVIDIA Quadro RTX8000 GPU with 48 GiB of RAM. The OS was Ubuntu 20.04.5.

C.1 Data Preprocessing

We preprocessed our data to improve its quality. Specifically:

- *Thread-level merge.* Our analysis is at the thread level, since entire discussions yield more powerful topics compared to single posts [68]. To create threads, we merge posts (taken from the same thread in the actual source) into a single thread (i.e., as seen from our dataset). For `StackEx`, we merge based on *post id*, for `Kaggle` on *forum topic id*, and for `CrimeBB` on *thread id*.
- *Data cleansing.* While each dataset provider already removed the so-called "junk" data, we performed additional data cleansing steps to reinforce consistency. Removing extraneous information ensures that the topic model derives high-quality topics. We stripped out attachment information, e.g. file attachments, web URIs, etc., removed named XML or HTML tags, special characters such as hex characters, and forum-specific markup tags such as ****IMG****, added during the CrimeBB data collection process. We also removed special characters such as hex characters that indicate newlines or text flow.
- *Filtering.* Following the recommendations of prior work [68], we filtered out threads that *(a)* contain less than 50 words, or *(b)* contain more than 10k words, which overall accounts for 0.15% of all threads. We filtered out these threads for two reasons: *(a)* they were not long enough for the proper discussion of any topic, or *(b)* they were too long, potentially containing large code samples or numerous diverse discussions.
- *Stop words.* As suggested by [27] we removed stop words only for the topic-word matrix creation, which defines the words belonging to a specific topic. To ensure that no stop words were included in our topic-word matrix, we

[28] Our solution falls in the category of strategic CTI.

retrieved stop words from several renown AI frameworks (namely NLTK, WEKA, SpaCy, and Onix), examined their union, and removed duplicates, resulting in 714 stop words.

C.2 Fine-tuning (Training)

A variety of parameters can be tuned to increase the performance of our "vanilla" models. In this Appendix, we discuss how we carried out our analyses. We stress that these operations were performed by taking the "vanilla" models (i.e., CTM using SBERT) and using them to analyze `StackEx` (i.e., our training dataset, $\mathbb{T}$). This ensures no "data-snooping" [19], yielding a fair assessment.

C.2.1 ***Model Evaluation Criteria.*** The quality of topic models can measured quantitatively and qualitatively, and taken as a reference for hyperparameter tuning. In our paper, we measure topic quality (TQ) by following the guidelines of prior works. Specifically:

- **Quantitative**. We empirically measure topic coherence via *normalized pointwise mutual information* (NPMI) [30] and *topic diversity* (TD) [48]. A coherent topic has words with high mutual information, which implies better interpretability by humans [124]. Topic diversity refers to the percentage of unique words among the top k words of all topics. Overall topic quality is measured as the product of NPMI and TD [48], i.e.: TQ = NPMI × TD.
- **Qualitative.** An intuitive visual representation of the semantics of extracted topics is provided by *word clouds*. The words are randomly positioned in a picture, their size is chosen according to the importance of a word in a topic [148]. The assessment is done by manually reviewing the word clouds, gauging how much they reflect the considered topic. To remove bias, such an assessment can entail the participation of multiple individuals. In our case, three authors independently reviewed the word clouds, and discussions were held to reach a consensus.

C.2.2 ***Parameter Selection.*** There are three parameters that we extensively analyze: *vocabulary size* (*Vocab*), denoting the most frequent words from the documents; *number of topics* (*TopNum*), denoting the number of topics the model attempts to identify; and *number of samples*, denoting the number of samples to estimate the final distribution of topics. Here, we discuss *TopNum* and *Vocab*, for which we proceed as follows. **(1)** We first set *Vocab*=5 000 and then vary *TopNum*=(20,30,40,50), using our quantitative metrics to find a "sweet spot," which we found with *TopNum*=50. **(2)** Next, we do the same by fixing value of *TopNum* and varying *Vocab*=(1 000,2 000,5 000,10 000), and study the resulting performance to find an optimal value (which we identified as *Vocab*=5 000). **(3)** For both of these analyses, we repeat the quantitative experiments 10 times to mitigate bias, and we validate our selection by qualitatively analyzing the resulting model. We report in Table 9 the results.[29]

For number of samples, we set it to 100 (up from 20) to enhance consistency and reduce variance [26] (we confirm this empirically). For other parameters (e.g., number of epochs) we kept the default values provided by the authors of CTM [27] (our source code has more details [42]). Finally, we report in Table 2 the list of topics[30] and Meta-Topics; see Section 4.4) considered by our fine-tuned model.

C.2.3 ***Verification of the Heuristic (from Section 3.4.1).*** Inspired by [62], topic probabilities of incorrect and out-of-distribution examples tend to be lower than the topic probability for correct samples. Therefore, we review the topic probabilities provided for `StackEx` (regarded as correct samples) and for `Speeches` (regarded as incorrect samples). We manually review the histograms of the provided probabilities for the threads of `StackEx` and `Speeches`

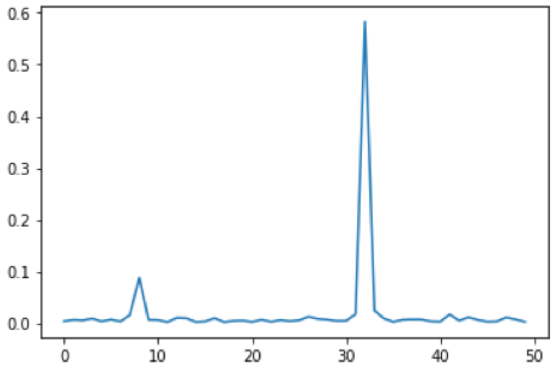
[29] NPMI ranges from −1.0 to 1.0; higher values correspond to higher coherence [79]. A TD close to 1.0 indicates more diverse topics [48].

[30] We use the term "toy problems" when a complex problem is explored based on a simplified use case. For AI, toy problems are used to test functionalities of Python libraries, e.g., trying features for Object Detection, resembling AI tutorials.

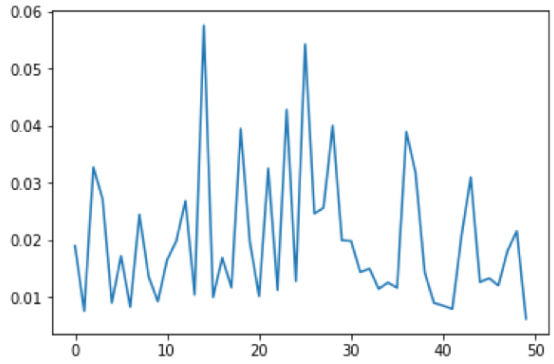
Table 9. Evaluation of hyperparameters (avg. 10 runs)– We select *TopNum*=50 since it provides the highest coherence (NPMI) and more granular topics (while having high TD); we also examined the word clouds and observed more interpretable topics for this value. We choose *Vocab*=5 000 since the overall TQ is similar and we prefer a higher number of vocabulary.

Hyperparameter	NPMI	TD	TQ = NPMI×TD
TopNum=20	0.098	0.840	0.0821
TopNum=30	0.117	0.817	0.0959
TopNum=40	0.129	0.783	0.1013
***TopNum*=50**	**0.134**	**0.740**	**0.0995**
Hyperparameter	**NPMI**	**TD**	**TQ = NPMI×TD**
Vocab=1 000	0.152	0.492	0.0749
Vocab=2 000	0.149	0.646	0.0960
***Vocab*=5 000**	**0.131**	**0.740**	**0.0964**
Vocab=10 000	0.117	0.756	0.0884

to define the minimum relevance ρ, and visualize these results in Figure 10; we stress that we consider 50 topics (as explained in Section 4). We see that, as expected, for `StackEx` there are some topics that have very high relevance (y-axis). In contrast, for `Speeches`, all topics tend to have a comparatively low-but-similar relevance. Importantly, for `Speeches`, we see that all topics tend to have relevance around $0.02{=}1/50{=}1/n$ (which confirms our conjecture in Section 3.4.1).



(a) `StackEx`: We see that the relevance of one topic is comparatively higher than the relevance of the other topics.



(b) `Speeches`: We cannot identify that the relevance of any topic is comparatively higher than the relevance of the other topics.

Fig. 10. **Validation of our Heuristic**– We plot the relevance (y-axis) for the 50 chosen topics on `StackEx` and `Speeches`.

Appendix D IN-DEPTH ANALYSIS OF OUR ORIGINAL FINDINGS

We provide an in-depth analysis of the results of all our experiments.

D.1 Validation Experiments (on $\mathbb{T}, \mathbb{P}, \mathbb{N}$) [Section 4.4]

We extensively discuss (qualitatively) the results shown in Table 3.

D.1.1 ***Self-check (`StackEx`).*** The main outcome is that the 50 topics derived from `StackEx` represent ≈ 50% of its threads. Stack Exchange forums are very heterogeneous and are likely to contain more than 50 topics. Hence, it is not surprising that only about half of this data corpus is covered by a 50-topic model. A thread can remain without

a topic assignment if *(i)* it discusses some topics unrelated to AI,[31] *(ii)* it discusses some AI-related topics which are different from the ones selected during model training, or *(iii)* it is weakly related to a large number of learned topics and hence none of them passes the likelihood threshold. A characteristic example for the latter case is a thread with a topic probability of 18.70% for *Data Input Formatting*, meta-topic *Data Preparation*, which does not exceed the minimum relevance ρ. This thread discusses image dataset in Python and how similar datasets can be manually created. A weak discussion about data preparation is present, but due to the focus on the manual part, we correctly denoted it as unclassified. It is also interesting to see that discussions on `StackEx` have a substantial focus on technical topics, either related to core AI or the supporting infrastructure. Topics related to "surrounding" questions, such as *Support Requests* and *Human-Technology Interaction*, are not very prominent in `StackEx`.

D.1.2 ***Positive Control Experiment (`Kaggle`).*** The goal of the positive control experiment is to demonstrate that a topic model attains similar coverage rates on a different AI-related data set. Table 3 reveals that 28.91% of threads in `Kaggle` discuss at least one of the 50 topics learned from `StackEx`. The difference in coverage rates is in line with our expectations. Even though Stack Exchange and Kaggle are both platforms for AI-related discussions, they also have some inherent differences. Kaggle's focus lies on AI challenges and competitions, and it is hence not surprising that a large share of threads on Kaggle is not among the 50 topics considered in our study (refer to Table 2). Also the difference in coverage rates of specific meta-topics highlights the diversity of these two forums. For example, the coverage rates of discussions on Python/AI Setup and Support Requests are higher on `Kaggle` than on `StackEx`, aligning with the intrinsic characteristics of Kaggle competitions.

D.1.3 ***Negative Control Experiment (`Speeches`).*** The coverage rates in the negative control experiment can be interpreted as false positive rates, and hence we expect them to be low. Our results show that "AI-related topics" are found in about 5% of `Speeches`. These hits were mainly on meta-topics *Human-Technology Interaction* (4.82%), and very few on *Learning Algorithms* (0.32%), the topic *Reinforcement Learning* to be precise. Our manual examination revealed some correlations between the vocabulary of `Speeches` and certain AI-related topics. For example, words such as *world, people, human, law, life,* and *system* are part of the topic *Human-Technology Interaction* and obviously not uncommon in `Speeches`. The two hits on *Reinforcement Learning*, barely exceeding minimum relevance (ρ=0.2), contained words such as *policy, action,* and *reward*.

D.2 First Case Study (pre-ChatGPT) [Section 5.2]

We discuss the results in Table 5, focusing on each meta-topic.

D.2.1 ***AI Core Topics.*** We observed a relatively small percentage of threads associated with *AI Core Topics*. Still, the identified threads for *AI Core Topics* cover a large spectrum of AI techniques.

- *Learning Algorithms.* For the meta-topic *Learning Algorithms*, we observed threads in DCF, referring to the use of experimentation techniques that help researchers in testing Tor's performance and discovery of related security problems, specifically highlighting Hidden Markov Modeling. Threads from CCF discuss Regression Analysis, Binary Tree Search, Linear Tree Search, Genetic Algorithms, Q-learning, and Reinforcement Learning. The discussions are very precise, e.g., about the difference between Q-Learning [134] and Temporal Difference Learning in the context of Reinforcement Learning. On the other hand, threads on gaming forums discuss path-finding algorithms, such as A^* or Dijkstra's Algorithm. In some cases, the discussions even include shared Python code.

[31] E.g., we have seen discussions related to driver's decision-making (in the context of autonomous driving). A decision of which topics to discuss is made entirely by users.

- ○ *NLP.* In this category, we observed many threads discussing password cracking and dictionary lists, and a strong interest in word semantics. As an example, a user was looking for "good" dictionary lists, which are commonly used for password brute-force attacks. In another thread, someone offers to sell the Oxford English Dictionary as an XML file. This thread advertises that the dictionary contains 290K words and compares it to the well-known WordNet lexical database, which has 150K words. In other threads, the use of statistical algorithms for password cracking was discussed, with one thread referring to a GitHub repo related to "Probable Wordlists." The repository organizes the most probable passwords and provides statistical probabilities for the use of different passwords, with the goal to the understand the human reasoning behind them. We also observed a high interest in open-source intelligence, e.g., for phishing as well as (ab)use of tools designed for defensive purposes. As NLP can be instrumental for crafting phishing content, it is not surprising to see NLP discussions related to phishing kits in cybercrime forums. Other threads assigned to *NLP* discuss black-hat Search Engine Optimization techniques used to improve search engine ranking based on content analysis of sites. Users discuss practices such as keyword usage, keyword bloating, keyword density, and TF-IDF.
- ○ *Model Training/Tuning/Evaluation.* This topic is rarely discussed. We have identified, for instance, threads discussing how an optimal model can be obtained via grid search and further tuned for a programming contest. Overall, we identified that adversaries are mostly interested in Learning Algorithms and NLP techniques when it comes to *AI Core Topics*. We also understand that even if *AI Core Topics* are not a prevalent topic in underground forums in general, based on the above examples we can deduce that underground forums are still a place to seek answers in regard to *AI Core Topics*. This is also confirmed by results presented below in the section regarding *AI Education*.

D.2.2 ***AI Supporting Topics.*** We categorize *Data Preparation*, *AI Education*, and *Python/AI Setup* as topics supporting AI.

- △ *Data Preparation.* We saw clear examples of interest in data analytics among discussion threads, and threads mostly fall into this category as they focus on preliminary data processing. Threads were found related to logarithmic scaling, vectorization, dimensionality reduction, and transformation through matrix multiplication. Overall, discussions on *Data Preparation* are mostly related to three main areas: formatting issues; explanations on the mathematical handling of data; and regex. Many of these discussions refer to the handling of data dumps for malicious purposes, preparing data for phishing campaigns or specific code injection attacks. This reveals attackers' activities related to handling leaked data, raising the question of what adversaries will do next with the preprocessed data dumps. Interestingly, when a question classified as *Data Preparation* is too specific, e.g., addresses Python coding related to data preprocessing, users are often referred to Stack Overflow.
- △ *AI Education.* There are four topics in this category: *Skill Requirements/ Learning*, *AI Resources*, *Object Detection (AI Toy Problem)*, and *Demand Forecasting (AI Toy Problem)*. Object Detection and Demand Forecasting are common toy problems referred to by AI tutorials to illustrate relevant Python libraries. However, even for such elementary problems we could observe characteristic traits of offensive intent. For instance, in one conversation tagged as *Object Detection*, a user wants to apply color and image recognition to detect an item and click on it, further revealing that he wants to gain this understanding to build a bot. In another thread, a user asks about circumventing image plagiarism detection and inquires if this is achievable with AI. Discussions on AI-related programming languages and on how to learn these languages are also prevalent in underground forums. A common recommendation for implementation of various hacking techniques is Python, and specifically Luna for game hacking. These discussions often go beyond programming languages, e.g., users inquire about AI usage in general, discuss specific AI techniques,

and ask about easy-to-use AI libraries, such as numpy, pandas, keras, pytorch, scikit-learn, etc. Also, the use of CPU vs. GPU for specific algorithms and libraries is compared. Similarly to the topic *Data Preparation*, we also see users referring others to the clearnet if they think the questions are too AI-specific. Some users also ask for advice on AI and hacking techniques in the same thread. One user, for instance, is looking for general advice on Neural Networks, rootkits, making money online, and hiding online presence. These terms clearly reveal an offensive purpose behind the intended use of AI. Some users do not just ask for advice or directions, but rather share knowledge from certain AI techniques, e.g., Google's research on determining the location of any image, weather forecasting using Time Series and Neural Networks, and cheating in online games using Neural Networks. An example of AI-assisted online game cheating from the threads is AimBot, which uses object detection to automatically identify a target and to point the crosshair at the target. Additionally, we observed a large number of posts looking for developers specializing in AI. Python programming, Big Data and AI are the most common requirements in such threads. Some users express explicit interest in hiring individuals with proficiency in Data Mining, NLP, and Database Management, with an additional required knowledge of the darknet.

△ *Python/AI Setup.* The meta-topic *Python/AI Setup* often appears in underground forums in connection with the Python Interactive Disassembler, various Python packages, AI library issues as well as system setup for AI development. Similarly, there were discussions on setting up or configuring the system environment for various hacking purposes, not necessarily in a clear context of AI. Discussions about non-AI related environment setups in the underground forum is of no surprise, as exemplified in a thread in which the Python setup of a phishing tool named "SocialFish" to clone websites or login pages is explored.

D.2.3 ***AI Surrounding Topics.*** Surrounding topics often occur in AI discussions but do not directly address AI techniques. Examples of such surrounding topics are threads in which users ask for help or discuss the impact of technology on society.

□ *Support Request.* For CMF, we observed a very high occurrence of *Support Requests*. The support requests on the darknet forums are of diverse nature, e.g., getting access to someone's website, hacking someone's account, asking about the correct link to another darknet forum, or phishing. For CGF, the discussions are more related to attacking gaming servers or asking about specific hacks. We also did observe interest in the setup of virtual environments, as well as to Python related topics such as code requirements, library updates and installation.

□ *Human-Technology Interaction.* Threads on this topic discuss the threat of AI to society from different viewpoints, e.g., the role/impact of robots or whether robots can outsmart humans.

D.3 Second case study (Pre/post-ChatGPT)[Section 5.3]

We stress that these analyses entail a data corpus having a different size than the one considered in the first case study (see Table 4). Therefore, some comparisons may be not possible (i.e., there is no new darknet data added in the June 2023 update to `CrimeBB`).

For the **"before" ChatGPT period** (January 2021→November 29th, 2022), we identified a conversation about the future of malware in the context of AI for the topic *Human-Technology Interaction* in CCF. This thread discusses whether the defensive or offensive side benefits more from AI. One side argues that malware will be eliminated by the advanced AI capabilities of the defense and the centralization of traffic, while the other side argues that it is a constant cat-and-mouse game, emphasizing that AI will cause malware to become more sophisticated. In another thread from CGF, assigned to the topic *Learning Algorithms*, a user is designing a deep reinforcement learning bot with Python and

wonders how he can best extract certain data from a game. Other threads discuss AI bots or AI-driven cheating for gaming, e.g., an algorithm that learns the behavior of players to mimic them and then mimics their gameplay.

For the **"after" ChatGPT period** (November 30th, 2022→June 2023), the amount of datapoints is small, which requires cautious interpretation and additional analysis. Reviewing the identified threads, we found new discussion topics related to DANs (Do Anything Now) exploits and jailbreaking ChatGPT for *AI/Python Setup*. For *AI Education* in CN-CF, we observed threads in which a user asks for a Data Scientist, another referring to Data Science as the perfect balance of hacking and math skills, and others looking for advice on hacking with Python. For *Human-Technology Interaction* in CCF, we see threads asking if anyone has used AI for anything nefarious before. While some users believe that AI is still "too much in its infancy" to be useful for the average threat actor, others bring up the use case of ChatGPT and similar tools for social engineering. For social engineering, however, the users emphasize that some level of human supervision is still required. Apart from AI being used to circumvent CAPTCHAs, users do not provide any answer to whether they have used AI for malicious purposes. Other threads indicate discussions related to using different large language models and how they can be used to create viruses or malware, with explicit reference to keyloggers. Specifically, a user leveraged ChatGPT to write a keylogger in PowerShell with startup and persistence using scheduled tasks. The user highlights that ChatGPT suggested a (commonly known) User Account Control bypass (to evade security measures). One user even states that it is only a matter of time until we see a huge hack that would never have been possible without AI. For *Learning Algorithms* in CGF, we observe similar discussions to the previous data sets, e.g., related to AimBot or autonomous survival bots.