You Might Have Known It Earlier: Analyzing the Role of Underground Forums in Threat Intelligence

Tommaso Paladini
tommaso.paladini@polimi.it
Politecnico di Milano
Milan, Italy

Lara Ferro
lara.ferro@mail.polimi.it
Politecnico di Milano
Milan, Italy

Mario Polino
mario.polino@polimi.it
Politecnico di Milano
Milan, Italy

Stefano Zanero
stefano.zanero@polimi.it
Politecnico di Milano
Milan, Italy

Michele Carminati
michele.carminati@polimi.it
Politecnico di Milano
Milan, Italy

ABSTRACT

This paper analyzes 88 million hacker forum posts of a publicly available dataset and 75,000 online articles over a 20-year timespan, studying the potential of hacker forums as a proactive Cyber Threat Intelligence (CTI) source. Using a custom Natural Language Processing pipeline with fine-tuned BERT-based models, we extract named entities from forum posts and reports and cross-reference their date of occurrence over different periods. Our analysis reveals that discussions on hacker forums precede official security reports for over 60% of the identified entities in 20 years of data. This highlights the relevance of these platforms as early indicators of cyber threats. However, our longitudinal analysis shows that such a trend has been constantly decreasing since 2012: forum discussions no longer consistently anticipate threats discussed in cybersecurity reports, possibly due to increased scrutiny or the emergence of alternative channels. This suggests that the CTI community should adapt by identifying and monitoring new platforms where threat actors congregate. Despite not being as thriving as in the first decade of 2000, underground communities are still releasing novel malware and showing interest in discussing malware employed in real cyberattacks. Our results highlight the value of hacker forums as early threat indicators and the importance of proactively monitoring them for potential cyberattack detection. This approach addresses the research gap that predominantly focuses on traditional cybersecurity reports.

CCS CONCEPTS

• **Social and professional topics → Malware / spyware crime**;
• **Computing methodologies → Information extraction**.

KEYWORDS

Cyber Threat Intelligence, Hacker Forums, Longitudinal Analysis, Natural Language Processing, Named Entity Recognition

RAID 2024, September 30–October 02, 2024, Padua, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0959-3/24/09
https://doi.org/10.1145/3678890.3678930

ACM Reference Format:
Tommaso Paladini, Lara Ferro, Mario Polino, Stefano Zanero, and Michele Carminati. 2024. You Might Have Known It Earlier: Analyzing the Role of Underground Forums in Threat Intelligence. In *The 27th International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2024), September 30–October 02, 2024, Padua, Italy*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3678890.3678930

1 INTRODUCTION

In 2012, a Remote Access Trojan (RAT) known as DarkComet gained public attention when it was discovered to be used by the Syrian government to spy on activists during the nation's civil war [53]. Still part of some threat actors' arsenals, the latest sightings of variants of the DarkComet malware date to 2023 in cyberattacks conducted against several worldwide organizations [7]. However, back in 2008, on the Hack Forums website, there were already discussions and samples of the potentially malicious tool developed by one of its users. As DarkComet, many other tools have been shared and discussed on online platforms, then deployed in large-scale cyberattacks planned by organized criminal groups or state-sponsored attackers. These platforms, whether found on the surface web or within the darknet, commonly denoted as "hacker forums" in literature, pose as hubs for computer science experts, enthusiasts, or "script-kiddies" (*i.e.*, unskilled attackers who use tools built by others) for sharing malware, stolen data, and tutorials on how to build and use such tools for malicious purposes. To adapt to the rapidly changing cyber threat landscape, private and public organizations have begun actively employing Cyber Threat Intelligence (CTI). This involves acquiring intelligence on novel cyber threats from private vendors who collect threat information related to cyberattacks. The collected information, often in the form of forensic artifacts such as IP addresses or malware signatures, serves as evidence to timely identify ongoing attacks against computing infrastructures and deploy adequate countermeasures. In other cases, intelligence is shared through technical reports, blogs, and newspaper articles that explain the behavior of malware and threat actors in natural language or discuss recently discovered vulnerabilities. Consequently, this research field predominantly adopts a *reactive approach – i.e.,* its strategies are usually based on intervening after an attack has occurred. This involves conducting a retrospective analysis using intelligence derived from cyberattacks recorded against other targets. In order to reverse this trend, several studies have employed social network analysis and trigger-based alerts [4, 17, 43]. These approaches typically monitor social network conversations (e.g., tweets of cybersecurity experts) and raise alarms based on the frequency of specific keywords [42, 43]. However, they often neglect platforms where anonymous individuals discuss malware development and share tools. In DISCOVER [43], hacker forums are involved to monitor possible observations of these novel terms over time, but keywords are extracted from other sources. In Sapienza et al. [42] instead, keywords are extracted from hacker forums and social networks. Despite being a step in the right direction, these studies lack an analysis of historical trends to confirm the validity of identified threats. In addition, these studies cover relatively short time windows (around 4 months of data [42]), which may not provide a sufficient overview of the relevance of such CTI sources. Other works that focus on hacker forums mostly analyze and categorize discussion attachments [5, 19]. There are also approaches to quantitatively analyze the relevance of CTI sources, such as intelligence platforms and feeds [18, 27, 31, 45, 50]. Significantly, none of the existing works evaluate CTI information from hacker forums. This area contrasts with the focus on platforms specifically built for the distribution of structured intelligence data or social networks intended for general discussions. This could involve correlating keywords found in security reports written in natural language and verifying whether hacker forums provide actionable insights and potential early warnings.

In this paper, we present an approach for understanding the CTI relevance of information extracted from hacker forums. We investigate the correlation between hacker forum discussions and historical cyberattack events documented in threat reports from traditional sources. Our analysis uses a dataset of over 88 million posts from 34 hacker forums, based on the publicly available hacker forum dataset CrimeBB [36], along with 75,000 articles on attacks, threats, and vulnerabilities from 16 online sources (including newspapers and security reports) spanning from early 2002 to early 2023. To do this, we develop a framework to analyze both data sources and search for potential threats emerging in the discussions under analysis. With our framework, we collect security-related keywords, compare their appearance dates in the two sources and conduct a longitudinal analysis to uncover trends. Our framework implements three steps: *Data Filtering*, *Text Processing*, and *Entities Matching*. First, by exploiting a fine-tuned BERT model [16], we filter hacker forum posts to keep only CTI-relevant data (*Data Filtering*). This step is essential as hacker forum discussions often include general-purpose topics, such as videogames and sports, irrelevant for our analysis. Then, we employ different Natural Language Processing (NLP) techniques to prepare the data, ensuring consistency and comprehensibility, for extracting named entities – *i.e.,* text spans that contain security keywords – with a Machine Learning (ML) model (*Text Processing*). In this phase, we combine and adapt existing approaches originally designed to work on technical reports and security documentation [24, 44]. In particular, we sanitize text for data integrity and filter content on security topics. We remove unrelated content and sentences without direct action or threat references (as in Extractor [44]) to reduce the analysis volume. Subsequently, we apply text normalization techniques to reduce the complexity of the textual data. This involves several steps. First, we convert verbs from passive to active voice, enhancing clarity and making it easier for algorithms to process the information. Next, we perform synonym homogenization, replacing different words that express similar concepts with a single term. This reduces ambiguity and streamlines the content. Additionally, we remove stopwords and internet slang, eliminating elements that do not provide relevant meaning for our purposes. Finally, in the name resolution phase, we ensure the correct identification and understanding of entities within the text. This includes resolving implicit references, handling pronouns and subject ellipses to fill grammatical gaps, and managing aliases for consistent entity representation (as suggested in Vulcan [24]). The last step of *Text Processing* involves solving an NLP task, Named Entity Recognition (NER), to extract relevant CTI keywords from processed text of hacker forum posts and security reports. We compare a popular architecture for NER, the Bidirectional Long Short-Term Memory (BiLSTM) model, against different BERT-based models, such as BERT [16], RoBERTa [30], and two models specifically constructed to handle cybersecurity terminology, SecBERT [29], and DarkBERT [23]. As already discussed in recent CTI research [44, 56], the peculiarities of cybersecurity terminology reduces the performances of traditional NLP models, requiring ad-hoc processing pipelines. In addition, compared to currently popular generative Large Language Models (LLMs), such as GPT-3 [9] and GPT-4 [34], BERT-based models do not suffer from the hallucination problem [25], namely the generation of text with non-factual information, which could hinder the reliability of our analysis. Among the tested models, DarkBERT [23], a LLM fine-tuned on hacker forum data, shows the best performances, extracting relevant entities with weighted average F1 score above 80%. We also conduct an extensive ablation study – *i.e.,* we analyze the performance of the NER model on a pipeline where each text processing step is *turned off* – to confirm their impact on the final performances of the model. Entities from hacker forums and technical reports are then cross-referenced (*Entity Matching*). By comparing the occurrence dates in both sources, for each matched keyword, we calculate its *latency* – *i.e.,* the time elapsed between its first appearance in forum posts and cybersecurity reports. We analyze the latency of the keywords from a global and temporal point of view to obtain a panoramic of the relevance of CTI extracted from underground communities and evaluate its evolution over time. We identify three scenarios, depending on the latency of the keyword occurrence: keywords discussed earlier in forums, keywords discussed earlier in threat reports, and keywords discussed at roughly the same time in both sources.

Our analysis reveals that discussions on hacker forums precede official security reports, with over 60% of identified security entities appearing first in forum discussions over 20 years of data. This highlights the potential of hacker forums as early indicators of cyber threats, especially for specific malware types like trojans and ransomwares, which are frequently discussed on these platforms. This information can help security professionals prioritize their efforts and allocate resources to address the most prevalent threats. However, our longitudinal analysis shows that the timeliness of these discussions has been decreasing since 2012, possibly due to increased scrutiny or the emergence of alternative communication channels. This suggests that while hacker forums remain a valuable source of threat intelligence, the cybersecurity community should adapt by identifying and monitoring new platforms where threat actors congregate. Despite this decline, underground communities are still actively developing and discussing novel malware, emphasizing the continued importance of monitoring these forums. We believe that our approach, which measures the relevance of hacker forums as a source of CTI in terms of the timeliness of the information they provide, can help security professionals stay ahead of emerging threats and improve their ability to detect and respond to cyberattacks.

In summary, we make the following contributions:

- To the best of our knowledge, we are the first to perform a longitudinal analysis of hacker forum discussions and threat reports to identify the correlation between the discussed topics in terms of keyword occurrence latency.
- We analyze and discuss the role of hacker forums as an intelligence source for threat prevention, identifying three discussion trends.

2 PRIMER ON CTI AND NLP

To counter the threats originating from cyberspace, researchers and security practitioners require novel tools capable of keeping pace with the evolving attacks executed by hackers. CTI approaches aim at closing this gap by promoting the sharing of actionable intelligence over the recent cyber threats. As described by R. McMillan [32], CTI is "evidence-based knowledge (*e.g.,* context, mechanisms, indicators, implications and action-oriented advice) about existing or emerging menaces or hazards to assets". Collecting CTI can inform decisions with the aim of preventing an attack or shortening the window between compromise and detection. Depending on the context in which it is applied, we refer to a *proactive approach* if the organizations predict future cyber threat strategies and incorporate these insights into the defense mechanisms of the system [12] for example by identifying CTI from previous threats, analyzing the identified CTI information, and deriving actionable insights that are helpful keys to prepare a system for proactive defense [39]. In essence, a proactive approach uses active measures to prevent attacks before they take place. Opposite to the proactive approach, there is the *reactive approach*, which involves taking action after a cyberattack has taken place somewhere. It entails collecting and analyzing real-time data to identify and mitigate cyber threats promptly, helping other organizations defend against ongoing attacks or anticipate them with vulnerability patching and knowledge of attackers' behavior [41].

Natural Language Processing. The extraction of CTI requires processing vast amounts of natural language data from diverse sources to identify specific patterns associated with cyberattacks, events, and vulnerabilities [39]. Natural language sources, due to their irregular nature, are particularly complex to be analyzed and are a subject of ongoing research. The most common solutions leverage NLP techniques to extract information from text, a field of artificial intelligence that focuses on enabling machines to understand, interpret, and generate human language.

Named Entity Recognition. Security entities (*i.e.,* cyber threat-related keywords) can be identified by analyzing each sentence at the token level, where text spans within the sentence are processed and classified into predefined categories (*e.g.,* Malware, Vulnerability, and Threat Actor). This task involves solving a NER problem, a NLP task, where each token is assigned a single class c from a limited set of classes C [26]. This problem can be solved directly by searching for specific words from a dictionary, but this approach ignores the semantic meaning of the word in the sentence. For example, depending on the context of a given sentence, the word 'gates' may be the surname of a person [33]. Thus, recent advancements [16] have focused on building complex deep learning models to process natural language text and infer the meaning of words according to their context. These models are trained on huge corpora of generic texts, such as newspapers, books, Internet discussions, and blogs. The knowledge learned from these generic sources can be transferred to the cyber security domain. The pre-trained models can be *fine-tuned*, namely adapted to a specific task or dataset by further training, improving their performance on that task.

3 PROBLEM STATEMENT

The dynamic nature of cyberattacks requires a shift from a mainly reactive approach in CTI to a more proactive one. Traditional CTI sources, in fact, such as vulnerability disclosures and threat reports, provide insights after an attack has occurred. While helpful, this leaves organizations exposed to novel menaces. As shown in a recent survey [39], most of the relevant research works focus on traditional threat reports or social networks, leaving the real role of hacker forums unclear. Hacker forums, where potential cybercriminals discuss malware development, vulnerabilities, and attack techniques, offer a potential source of proactive CTI.

This research aims to (1) determine if discussions on hacker forums can provide early indicators of emerging cyber threats, anticipating traditional threat reports, and (2) study the relevance of hacker forums in the cybersecurity domain over the years. To do so, we develop a method to process and analyze the unstructured, informal language of hacker forums to extract reliable CTI insights. Let us note that the goal of this work is not to propose a novel keyword extraction methodology but to perform a longitudinal analysis of hacker forums to understand their relevance in the cybersecurity domain.

Furthermore, aware of the ethical concerns that our research may raise (see Section 10), we focus our analysis on security-related entities while anonymizing individuals and platforms to prevent disclosing sensitive data and avoiding downloading harmful content. Finally, our results are presented objectively, emphasizing cybersecurity implications without discussing specific behaviors in underground forums.

3.1 Challenges

Extracting actionable intelligence from underground forums and cyberthreat reports requires overcoming several challenges.

Specialized Terminology. The cybersecurity domain has a unique vocabulary, making it challenging to identify and extract relevant threat information. We summarize the main challenges related to this aspect below.
Technical Language. Security texts often contain technical elements, such as IP addresses and code, that hinder traditional NLP techniques, such as tokenization, Part-of-Speech (POS)-tagging, and

dependency parsing [44, 56]. We solve this challenge by adopting an LLM, capable of automatically obtaining a correct representation of these texts (see Section 5.2).

Inconsistent Writing Style. Threat reports are authored by different vendors, which follow distinct writing styles and conventions [44]. Similarly, hacker forum posts are written by a multitude of users. To solve this challenge, we normalize the natural language texts of both sources using a common processing pipeline (see Section 5.2).

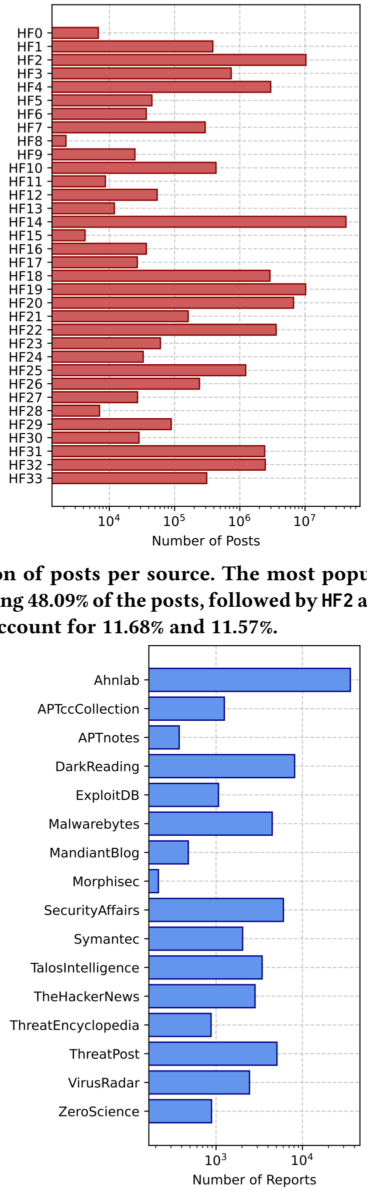
Informal Language. This challenge is related to hacker forum content. Forum discussions are often rife with slang, misspellings, and technical jargon, hindering traditional NLP techniques. Our findings show that automatic correction of misspellings may fail . We solve this challenge by adopting an LLM that has been fine-tuned on hacker forum texts (see Section 5.2).

Data Volume. The massive amount of discussions makes manual analysis impractical, requiring automated methods. We solve this challenge by filtering relevant posts with a ML model (see Section 5.1).

4 DATASETS OVERVIEW

Hacker Forum Data. Our hacker forum discussion dataset expands on the CrimeBB [36] dataset, which contains over 91 million posts collected from 32 platforms in different languages, such as English, Russian, and Spanish. To the best of our knowledge, this dataset represents the currently largest collection of hacker forum posts and threads. Multiple works [14, 20, 22, 35, 37, 51] have already analyzed this dataset from different points of view (see Section 8). For simplicity, we discard data from forums with discussions in languages other than English. In addition, we include hacker forum posts collected by the AZSecure portal [40], publicly available leaked datasets [15], and data scraped from another forum, extending the CrimeBB dataset with more than 600,000 posts. Let us note that some of the forums are already present in the CrimeBB dataset; therefore, we merge all the data and drop duplicate posts. For privacy reasons (see Section 10), we obscure the names of the forums, replacing them with identifiers (e.g., HF0, HF1, HF2, etc.). Forum posts cover both relevant (*e.g.,* malware advertisement and development) and non-relevant topics (*e.g.,* videogames and sports). Figure 1a illustrates the contribution of each forum to the composition of the final dataset. The final dataset contains a total of 88,323,254 posts from 34 platforms, distributed across 7,052,097 threads, which are further categorized into 79,694 subforums. The number of users engaged in discussions on these forums is 3,077,399. As mentioned earlier, we consider a time period that goes from January 1, 2002, up to April 30, 2023.

CTI Report Data. We create the report dataset by considering sources renowned for their significance in the CTI community and newspapers or websites with dedicated sections on hacker activities, malware, and vulnerabilities. Figure 1b illustrates the contribution of each of the 16 sources to the dataset. We acquire 75,433 documents written in natural language, covering the same time period of the forums data, with the earliest date being January 1, 2002, and the most recent being April 30, 2023. Our dataset contains, on average, 3,362 documents per year and 289 documents per month.



(a) Distribution of posts per source. The most populous source is HF14, comprising 48.09% of the posts, followed by HF2 and HF19, which respectively account for 11.68% and 11.57%.

(b) Distribution of reports per source. The greatest contribution is provided by Ahnlab, accounting for 47.58% of the total acquired documents, followed by DarkReading with 10.76%, and SecurityAffairs with 7.99%.

Figure 1: Distribution of data per source type in our experimental dataset.

A comparison of the temporal distribution of the collected hacker forum posts and reports is shown in Figure 2. We have a higher number of posts compared to the technical reports, between 2 and 3 orders of magnitude. These values are realistic, as there is a high probability the discussions occurring on online platforms are considerably more numerous than technical documentation and news based on the knowledge of potential or occurred threats. With respect to the temporal distribution of posts, we observe that 2002 is the year with the lowest number of discussions, with

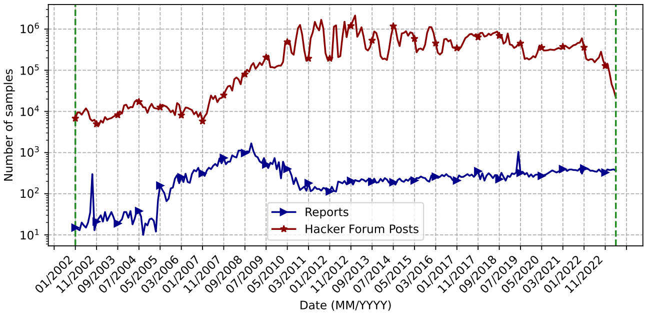


Figure 2: Monthly distribution comparison of hacker forums posts and reports. The two vertical green lines, respectively, point to the temporal beginning and end of the dataset. The two points are set at January 2002 and April 2023.

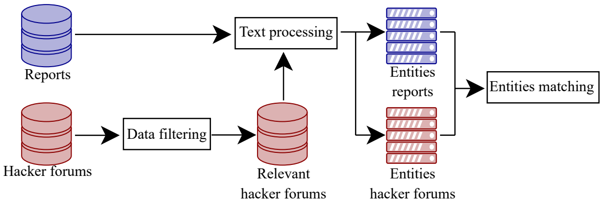


Figure 3: Overview of our analysis framework.

93,947 posts, while there is a peak in 2012, with 9,980,073 posts. Meanwhile, the year with the fewest documents at our disposal is 2004, with 318 reports, while the peak was reached in 2008, with 11,733 reports. From Figure 2, we also observe an increase in the volume of discussions between late 2007 and the end of 2011. This phenomenon could be attributed to a rising interest in underground communities. A similar increase can be observed for the volume of collected threat reports, specifically between mid-2005 and the first months of 2009. However, we do not have sufficient information to correlate this quasi-concurrent increase in the volume of data for both sources. Such oscillations could also be attributed to an increased activity of specific threat report vendors during that period of time.

5 ANALYSIS FRAMEWORK

The architecture of our framework consists of three main logical steps (see Figure 3). The first step is *Data Filtering*, exclusively performed on data collected from hacker forums. Given the vast number of posts covering various topics beyond malware and threats, we employ trained models to retain only discussions relevant to CTI investigations, reducing the data volume. The filtered posts, as well as the reports, serve as input in the *Text Processing* phase. Here, the data undergoes a series of transformations through an NLP pipeline tailored to optimize the structure of sentences for training a NER model. This model extracts entities from the text using a LLM. We conduct an ablation study to determine the best combination of steps for achieving optimal performance in recognizing concepts in sentences. We also test different ML models. The final step involves *Entities Matching*, where keywords obtained from one source are employed to search for correspondences in the other source. Finally, we analyze the final results to identify the temporal trends.

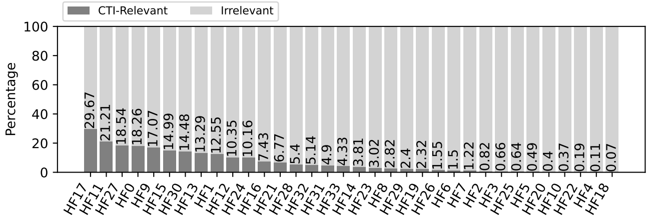


Figure 4: Percentage of relevant posts in relation to the total volume of posts. The forum with the highest percentage is HF17, with 29.67% relevant posts, while the lowest is HF18, with only 0.07%.

Table 1: List of keywords used to label the *CTI-Relevant* class

CTI-relevant Keywords
Adware, Backdoor, Botnet, Bruteforce, Bypass, Chargeware, Crack, Crimeware, Crypter, CVE, Cyberweapon, DDoS, Downloader, Dropper, Exploit, Firewall, Flood, Hack, Hijack, Infect, Inject, Keylogger, Logic bomb, Malware, Monetizer, Password, Payload, Phishing, RAT, Ransomware, RCE, Reverse shell, Riskware, Rootkit, Scanner, Security, Shell code, Spam, Spoof, Spyware, SQLi, Steal, Trojan, Virus, Vulnerability, WAF, Worm, 0day, Zeus

5.1 Data Filtering

This phase is exclusively related to the hacker forums dataset: In darknet forums, as in any other online community, the topics discussed range from discussions covering interests such as sports or video games to threads specifically dedicated to sharing malware and tutorials related to hacking techniques. We model this as a binary classification problem: for a specific post, a machine learning algorithm has to classify it as *CTI-Relevant* or *Irrelevant*. The CTI-relevant class refers to discussions whose content may be potentially relevant to cyber security (e.g., posts that describe malware and its development, attack techniques, etc.). To build our training dataset, we refer to the methodology proposed in Deliu et al. [15]. In particular, we mark posts as irrelevant if they do not contain any of the terms presented in Table 1 and additionally, the text includes non-security related keywords such as those related to sports, music, movies, and drugs. The resulting dataset comprises 1,200 entries, distributed in 400 relevant posts and 800 irrelevant ones, following a partition that appears representative of a real-case scenario in the distribution of topics. We apply a series of language processing steps to each post, following the procedures outlined in papers addressing similar tasks [10, 15]. In particular, we lowercase the text, tokenize it, and remove stopwords, punctuation, tokens with non-ASCII characters, and tokens with many repeated letters (e.g., laughter), and lastly, we apply lemmatization. We then evaluate two machine learning models, BERT [16] and Support Vector Machines. For the former, we incorporate an additional layer for binary classification, while for the latter, we evaluate two different approaches to represent the data: Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). To assess the performances of the models, we hold out 20% of the dataset defined formerly for validation and 10% for testing. We choose the model based on the F1-score calculated on the validation set. We choose the model based on metrics such as Accuracy, F1-score, Matthews Correlation Coefficient (MCC), and Receiver Operating Characteristic (ROC) curve calculated on the validation set. We obtain the best performances with the BERT-based model, which achieves an F1-score of 87.8%, an Accuracy of 91.7%, an MCC of 0.815, and an area under the ROC curve of 0.980 on the test set. It shows excellent predictive capabilities for both relevant and non-relevant occurrences, with correct true positives at 92.50% and relevant class true positives at a value of 90.00%. Finally, we employ this classifier to filter the CTI-relevant data in our hacker forum dataset.

5.2 Text Processing

The goal of this step is to extract a list of CTI named entities from hacker forum posts and reports. The following steps are thus applied to both datasets, as generalized in Figure 3. Firstly, we define an NLP pipeline for data processing, taking inspiration from the processing steps presented in other works [4, 24, 42, 44]. Building upon previous work on technical reports and security-related documentation, we categorize the tasks into three main areas, as outlined below.

(1) Sanitization. First, we perform a few processing steps to reduce the noise present in the textual content to be analyzed. In particular, we eliminate redundant text that lacks actions or references to threats, thus reducing the data load to be processed in the pipeline (*Unrelated content removal*). We employ a binary classifier trained on a set of sentences labeled as *informative* or *uninformative*. We consider as informative sentences those that potentially contain entities and actions related to cyber threats. Then, we identify and correct spelling errors in the text, as done by Adewopo et al. [4], using the Jamspell library[1] (*Misspelling correction*).

(2) Text normalization. This phase aims to standardize the representation of text by simplifying linguistic structures and concepts and making them easier to process. First, we convert passive verbal structures to active, e.g., "the file was downloaded by the user" becomes "the user downloaded the file" (*Passive/active conversion*). We leverage the algorithms used by Satvat et al. [44], which uses the Spacy [1] library. Then, we homogenize synonyms, reducing eventual ambiguities proper of CTI texts, e.g., "C2", "C&C", and "Command and Control" (*Synonym homogenization*). For this task, we use the dictionaries provided by Satvat et al. [44]. Then, we remove stopwords by using the NLTK library [54]. These words may not contribute significantly to the context or meaning of the text, thus leaving only potentially relevant terms without distorting the sense of the sentence (*Stopwords removal*). With the same motivations, we remove slang expressions or abbreviations typical of informal internet language since these elements might not be understood by the model (*Internet slang removal*).

(3) Name resolution. This phase reconciles implicit references that refer to the same entity with the actual referent. Making these implicit references explicit is essential, as they can otherwise reduce the accuracy of subsequent steps, leading to ambiguous and imprecise final results. In this step, first, we resolve pronouns and subject ellipsis to their corresponding words (*Pronouns and subject ellipsis resolution*). We develop an approach based on Part-of-Speech (POS) tagging to recognize which sentences lack an explicit subject or use a pronoun and replace it with the correct entity present in the text [44]. Lastly, we resolve aliases of CTI-related entities, such as malware names and threat actors, to a single name, e.g., WannaCry can also be referred to as WCry or WanaCryptor [24] (*Aliases handling*). To address this issue, we employ a dictionary consisting of a list of well-known malware and threat actors with their respective aliases. Our dictionary has been manually curated by experts working in the threat intelligence division of a major Italian telecommunication company.

These steps contribute to making the text more consistent for solving a NER task, an NLP problem, with a machine learning model. A named entity is defined as a word or a phrase that identifies one item from a class of items that share similar attributes [26]. Essentially, solving the NER task means to assign to each token in a sentence a class from a set of predefined classes (e.g., Malware, Threat Actor, Tool, etc.). For a more formal definition of NER, we refer the reader to Li et al. [26]. In our approach, we respectively apply NER to recognize CTI-related entities from textual data of hacker forums and threat reports. In order to train our machine learning models, we use a publicly available dataset containing annotated sentences from threat intelligence reports, APTNER [52]. This dataset contains 21 CTI-related classes of entities, such as *Malware*, *Threat participant*, and *Vulnerability name*. As for hacker forums, we manually annotate a set of 5,000 sentences randomly extracted from our hacker forum dataset, following the same scheme of APTNER [52]. We use these datasets to train a set of machine learning models and conduct an ablation study to discover the combination of steps that maximize the performances of such models for each type of dataset. This means that we assess the importance of the various components of our framework by examining how the removal of each step impacts the overall performance of the models. We consider various classifiers used in the literature for entity recognition tasks: BERT [16], RoBERTa [30], and a Bidirectional Long Short-Term Memory (BiLSTM), along with models specifically trained for CTI tasks such as DarkBERT [23] and SecBERT [29]. We evaluate their performance by calculating the weighted averaged F1 score metric on the test sets obtained by holdout from the starting datasets (annotated hacker forum posts, APTNER [52]). This metric is computed by calculating the mean of all F1-scores per class while considering the support of each class, where support refers to the number of actual examples of the class in the dataset. This approach adjusts the contribution of each class to the final average F1-score based on its size, providing a more balanced perspective. The results of our ablation study are reported in Table 2 and Table 2, respectively evaluated on hacker forum data and security reports. The tables show the results of the complete pipeline and how each step negatively affects the performance of each model. First, we find that for both types of sources, the optimal framework utilizes the DarkBERT model [23]. Then, we notice that for both data sources, the *Misspelling correction* step negatively affects the predictive performance. The JamSpell library seems to distort the meaning of the sentences and undermine the entity recognition model. For example, sensible terms like *spamming* are corrected to *slamming*. Therefore, we remove it from the final processing

[1] https://github.com/bakwc/JamSpell

Table 2: Results of the ablation study for the two datasets. The 'Complete NLP pipeline' column shows the performances of the model when all the steps are included. Other columns show the performances when the processing step is removed. Each percentage refers to the weighted average F1-score calculated on the respective test set. Higher values indicate that the considered step negatively impacts the performances.

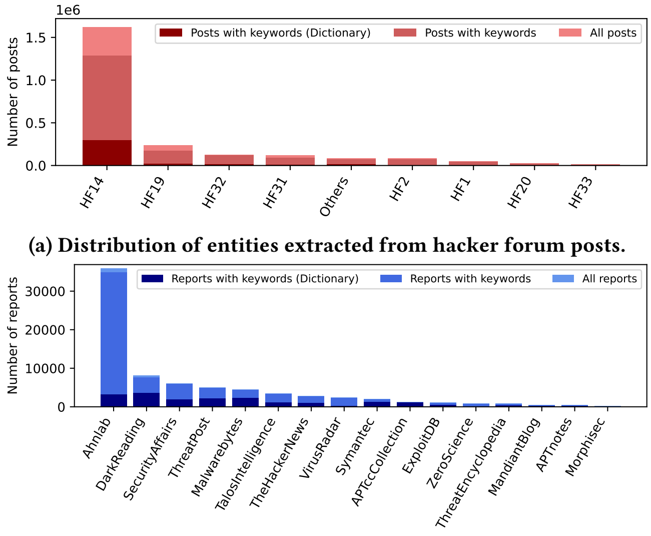
(a) Results of the ablation study of the NLP pipeline for hacker forums data.

Model	Tokenization	Complete NLP pipeline	NLP pipeline without:							
			Unrelated content removal	Misspelling correction	Active/passive conversion	Synonym homogenization	Stopwords removal	Internet slang removal	Pronouns and subject ellipsis resolution	Aliases handling
BERT (base) [16]	Cased	72.24%	74.16%	72.77%	74.10%	73.20%	72.16%	73.43%	**74.48%**	72.48%
	Uncased	68.48%	71.63%	71.66%	71.63%	70.25%	70.79%	70.30%	70.77%	**73.15%**
BiLSTM	Cased	52.74%	53.59%	52.89%	49.75%	**54.33%**	53.81%	52.40%	50.62%	50.58%
	Uncased	51.06%	50.75%	48.03%	50.60%	**51.61%**	50.01%	49.71%	47.71%	48.20%
DarkBERT [23]	Cased	78.42%	78.14%	**79.95%**	77.82%	77.74%	78.61%	78.98%	79.11%	79.25%
	Uncased	73.71%	77.45%	**77.63%**	75.41%	76.30%	73.96%	76.77%	75.67%	77.30%
RoBERTa [30]	Cased	76.28%	77.42%	**78.98%**	77.01%	77.38%	78.65%	76.47%	76.76%	77.93%
	Uncased	74.81%	75.79%	75.93%	73.81%	72.58%	76.14%	74.77%	75.30%	**77.23%**
SecBERT [29]	Cased	64.33%	64.33%	**64.99%**	63.48%	63.77%	64.82%	64.33%	64.65%	66.64%
	Uncased	63.15%	64.15%	64.81%	63.32%	63.62%	**64.82%**	64.17%	64.47%	66.48%

(b) Results of the ablation study of the NLP pipeline for cybersecurity reports.

Model	Tokenization	Complete NLP pipeline	NLP pipeline without:							
			Unrelated content removal	Misspelling correction	Active/passive conversion	Synonym homogenization	Stopwords removal	Internet slang removal	Pronouns and subject ellipsis resolution	Aliases handling
BERT (base) [16]	Cased	84.49%	84.66%	**85.83%**	84.50%	84.99%	84.50%	84.65%	83.82%	80.33%
	Uncased	81.50%	81.00%	**82.99%**	80.50%	80.99%	81.50%	81.17%	80.33%	79.33%
BiLSTM	Cased	63.09%	62.17%	**65.04%**	64.06%	64.32%	64.34%	63.47%	62.57%	59.06%
	Uncased	59.84%	59.90%	**63.54%**	60.69%	61.89%	62.57%	58.86%	60.00%	56.16%
DarkBERT [23]	Cased	85.66%	85.33%	**86.33%**	85.17%	85.66%	85.83%	85.80%	84.49%	82.17%
	Uncased	82.00%	82.33%	**83.17%**	82.00%	81.99%	82.00%	82.16%	82.33%	81.66%
RoBERTa [30]	Cased	82.97%	84.33%	**85.82%**	85.66%	85.33%	85.00%	84.83%	84.83%	82.15%
	Uncased	81.83%	**82.00%**	**82.00%**	81.33%	82.00%	80.99%	81.66%	81.83%	80.66%
SecBERT [29]	Cased	77.49%	77.50%	**78.49%**	77.32%	77.49%	76.99%	76.83%	77.33%	76.48%
	Uncased	77.33%	77.33%	**78.33%**	77.16%	77.33%	76.83%	76.66%	77.16%	76.32%

pipeline, considering that the best solution consists of employing a specialized model, such as DarkBERT [23]. Furthermore, we repeat the ablation study on both sources with the DarkBERT model, greedily removing the steps that lower the performances. The final results are included in Appendix A, under Table 4b and Table 5. We observe that by removing the *Synonym homogenization* step when processing hacker forum data, its final F1-score increases up to 82.15%. Probably, synonym homogenization does not adapt well to the writing style of hacker forums, stressing the necessity to adopt different processing steps for the two data sources. In conclusion, we adopt a fine-tuned DarkBERT model for processing forum data, without *Misspelling correction* and *Synonym homogenization*, and achieve a weighted average F1-score of 82.15%, 82% Precision, and 85% Recall on the test set. For reports, we fine-tune DarkBERT and obtain 86.33% F1-Score, 86% Precision, and 88% Recall on the test set. In this case, the text processing pipeline excludes only the phase of *Misspelling correction*. With the identified optimal combinations of NLP pipeline steps and the chosen LLMs respectively fine-tuned on the two datasets, we extract from the datasets described in Section 4 a list of named entities and the corresponding dates of occurrences. The source code of our framework is available at https://github.com/necst/underground_forums_analysis_code.



(a) Distribution of entities extracted from hacker forum posts.

(b) Distribution of entities extracted from threat reports.

Figure 5: Distribution of extracted entities per source type.

5.3 Entities Matching

In this step, we verify whether the entities extracted from hacker forums also appear in reports' discussions contained in our dataset. We treat each entity extracted from hacker forums as a keyword and search for its exact matches in the textual content of threat reports. To reduce the number of false negatives, we perform this search in the opposite direction as well, starting from the reports and searching the keywords in forum posts. For each keyword, we obtain a pair of lists where one contains the dates of each occurrence of the keyword in hacker forum posts, and the other contains the dates of reports and articles. In this step, we extract more than 55,000 entities from nearly two million relevant posts. As shown in Figure 5a, the majority of these keywords is extrapolated from HF14 posts, covering 67.51% of the total. A significant contribution is also made by HF19, contributing 9.21%, followed by HF32 at 6.36%. The remaining 16.92% of the keywords are extracted from the other platforms. Similarly, we can analyze the sources present in the dataset of the reports, from which we extract more than 40,000 entities from 73,000 reports containing keywords. Figure 5b shows how each source contributes to the total reports containing entities. Most of the entities are extracted from Ahnlab's reports, accounting for 47.56%. Other notable sources include DarkReading, SecurityAffairs, Malwarebytes, and ThreatPost, with values ranging between 6% and 11%.

We conduct further analysis by focusing on a subset of entities extracted from the datasets. We employ the dictionary of known malware and threat actors mentioned in Section 5.2. Then, we filter the extracted keywords, considering those that fall within our list. Using the dictionary, we obtain 1,554 entities extracted from just under 400,000 posts. The distribution of this subset of entities is very similar to the one obtained by considering all keywords, except for an increase of HF14's contribution (see Figure 5a). Additionally, for the malware entities belonging to this subset of keywords, we provide supplementary information regarding the types of malware most frequently discussed. From Table 3, we can observe that the majority of discussions are about the *Trojan* category, accounting for 42.26% on hacker forums and 43.42% on reports, followed by a significant number of discussions about *Ransomware*. By restricting the entities with the dictionary, the distributions and proportions of reports with entities slightly change. Less than 20,000 reports contain interesting entities, and the contribution of Ahnlab decreases to 16.26%, while the contribution of Malwarebytes, Symantec, and TalosIntelligence increases significantly.

6 SECURITY ENTITY LATENCY ANALYSIS

In this section, we describe our experimental analysis, specifically designed to answer the following research questions:

- **RQ1** *Do hacker forum discussions anticipate the topics discussed in cyber threat reports?*
- **RQ2** *How did historical trends evolve over time?*

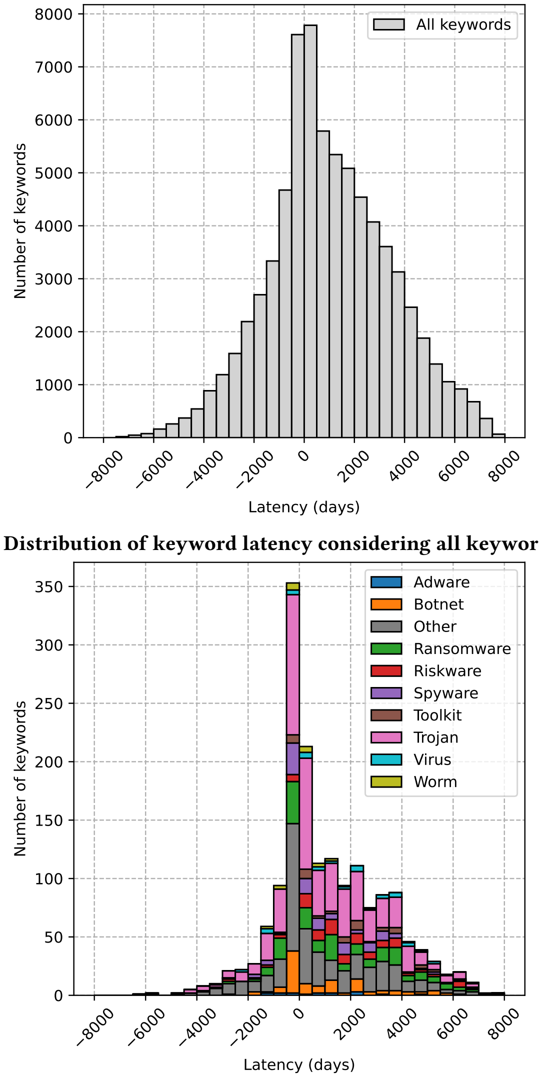
Latency metric. We define latency as the time elapsed between the occurrence of a keyword in cybersecurity reports and forum posts. For a specific keyword w_i, we consider t_{w_i} the date of its first appearance recorded in forum posts, and $t^*_{w_i}$ its first appearance in cybersecurity reports, and express the latency l of w_i as follows:

$$l(w_i) = t^*_{w_i} - t_{w_i} \tag{1}$$

Positive latency values indicate that a certain keyword appears first in forum posts and subsequently in reports. Conversely, for negative values, the keyword is first recorded in technical reports.

6.1 Global Latency Analysis (RQ1)

In this experiment, we first analyze the distribution of the latency of the keywords extracted with our framework. We then identify and discuss three different scenarios in the trends of keyword appearance. For each keyword extracted with our framework, we calculate its latency and evaluate the overall distribution of the observed values. The distribution of data in Figure 6a shows that 64.69% of the keywords have appeared earlier in underground forum posts preceding public feeds. This phenomenon highlights the fact that hacker forums mostly precede potential sightings of threats and their subsequent reporting on security-related platforms. We repeat this process also for keywords that only appear in our dictionary of known threats and threat actors, defined in Section 5.2. As shown by Figure 6b, we obtain similar distribution values for this subset of entities, with no significant changes in the latency distribution. For this subset, 62.77% of the entities have latency greater than 0.



(a) Distribution of keyword latency considering all keywords.

(b) Comparison of the latency distribution of all keywords and the subset of keywords filtered with our manually curated dictionary.

Figure 6: Distribution of keyword latency considering all population samples from 2002 to early 2023.

Table 3: Categorization of malware-related entities extracted per source.

	Adware	Botnet	Ransomware	Riskware	Spyware	Toolkit	Trojan	Virus	Worm	Other
Hacker forums	1.11%	8.08%	13.86%	7.14%	7.06%	3.66%	42.26%	3.49%	2.21%	11.13%
Reports	1.19%	8.07%	13.51%	7.48%	7.14%	3.82%	43.42%	2.72%	1.95%	10.70%

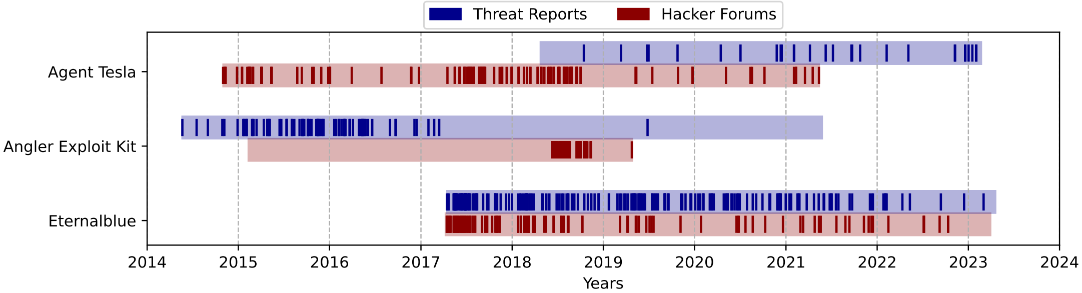


Figure 7: Notable examples of the three different identified trends. Occurrences from hacker forums are highlighted in red, and those from reports are in blue. Horizontal bars indicate the period in which the keyword has been observed in the corresponding source type. Vertical ticks represent an observation of the keyword.

Both distributions show a significant peak around 0, with latency between -500 and 500 days. This hints that many keywords appear in discussion almost concurrently in underground communities and security news. Figure 6b also shows that malware categories could be analyzed separately, as they exhibit different values of latency. The majority of the classes that have positive latency belong to the Trojan and Ransomware categories, signaling an intense activity on these types of malware. This consideration reflects the distribution of categories shown in Table 3. Finally, we can examine the volume of keywords concurrently discussed in both sources by considering those that have latency $l \in [-\hat{t}, \hat{t}]$, where $\hat{t}$ is an adjustable parameter that can be tuned to align with an institution's preferences regarding the timeliness of information collection. With $\hat{t} = 7$, we observe that around 2.46% keywords fall within this category, while with $\hat{t}$ equal to 30 days, this value increases to 3.95%. This leaves more than 50% keywords with latency greater than 0. Therefore, the majority of keywords still make their first appearance in forum discussions. Based on the previous observations on the latency distribution, we identify and discuss three scenarios in the trends of keyword appearance in discussions.

Keywords first discussed in hacker forums. As an example of this scenario, we discuss Agent Tesla, a RAT written in .NET that has been actively targeting users with Windows machines since 2014. With our framework, we were able to trace the 2014 release of Agent Tesla. As shown in Figure 7, this malware is a clear example of the first type of trend identified, where the keyword is discussed first in hacker forums and subsequently in reports. In this case, the distinction of activity periods is well-defined, showing a 3-year period during which Agent Tesla is exclusively discussed in forums. Later on, it is picked up by various reports in the following years, continuing to be mentioned to these days. This malware became of particular interest to security experts in the subsequent years due to a series of phishing campaigns during the COVID-19 pandemic [8]. This scenario suggests that users often discuss the malware on such forums before releasing it.

Keywords first discussed in security reports. Angler Exploit Kit represents an example of the second type of trend (see Figure 7). It is one of the most sophisticated exploit kits used by cybercriminals to deliver diverse malware to compromised machines, and it first appeared in late 2013. We extract this entity initially from CTI reports, starting from early 2014. As for hacker forums, there are scattered occurrences from 2015 onwards, with a peak between late 2018 and early 2019. In this scenario, the malware becomes a matter of public interest before it is discussed on forums. This example suggests that features of malware described in security reports draw the interest of underground communities and probably inspire the future evolution of other malware: in this case, the evasive behavior of Angler is mentioned in explanation threads [47].

Concurrent discussions. The third scenario is presented with the entity EternalBlue, a computer exploit developed by the National Security Agency that was leaked by a hacker group in April 2017. As shown in Figure 7, occurrences in hacker forums and reports overlap, particularly in the early months of discussion, and continue to be referenced by both sources to this day. In this case, there was no trace of the vulnerability exploited by EternalBlue before it became a public concern due to the WannaCry ransomware attack. This scenario represents events that immediately spark simultaneous discussions among both security experts and enthusiasts on hacker forums, given the scale of the threat [21].

Answer to RQ1: On a global point of view, the majority of cyber threat-related keywords make their first appearance in forum discussions. Therefore, hacker forum discussions have anticipated some of the concepts shown in threat reports. Only a small portion of keywords appear to be quasi-concurrently discussed in both sources.

6.2 Longitudinal Latency Analysis (RQ2)

The analysis of Section 6 shows that keywords mainly appear in forums before cybersecurity reports and newspapers. However, it does not consider how time shifts the distribution of this trend. In

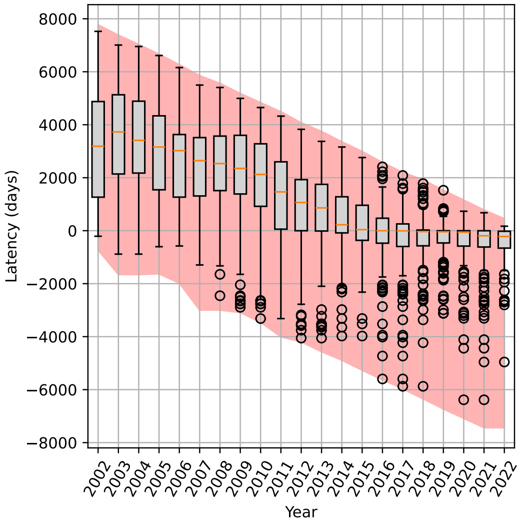


Figure 8: Boxplot of annual distributions of keyword latency calculated on the subset of keywords filtered with the dictionary of known entities. The red area in the background indicates the region of space enclosed between the minimum and maximum latency dates of all keywords (i.e., those not filtered with the dictionary).

recent history, some underground forums have been scrutinized and tackled by authorities [36]. The emergence of social media and private messaging apps might have posed as new communication channels for malware developers to share information, hindering the role of traditional forums. In this experiment, we provide a longitudinal analysis of the keyword latency over the last twenty years to show how time has influenced its distribution. We analyze data in year-long increments, starting from 2002 to 2023, under the assumption that data distributions remain stable within this timeframe. For each year, we examine new keywords identified from forum posts within that year and calculate their latency using the entire dataset of reports. We plot the resulting year-long distributions with boxplots in Figure 8. The boxplots highlight a consistent trend: the latency of newly discovered keywords has been steadily decreasing since 2012. Until 2012, the median latency is bounded between 3189 and 1465 days. This result shows that the great majority of keywords appearing in forums were captured in security technical reports, usually between 8.74 years and 4.01 years later. Despite this result, security reports still capture some keywords in a timely manner. However, after 2012, we can observe a slow descent of the median latency until reaching negative values in 2016. These results clearly suggest that the relevance of CTI extracted from hacker forums has been slowly decreasing over time and that their overall influence over novel malware sample discoveries has reduced. Another reason could be related to larger investments in CTI since 2019 [2], which could have improved the awareness and monitoring capabilities of the security community. In addition, we run independent hypotheses tests over each annual sample of the population and are able to confirm that the mean latency is positive until 2015 and negative after 2017 with confidence level $\alpha = 0.05$. However, we failed to reject the null hypothesis for the samples of 2015 and 2016. This result suggests that the mean latency could already be negative by 2015, despite reaching a positive value close to 0, in particular 166.51 days, whereas in 2016, it could still be positive, even if the mean latency is -160.42 days. Most likely, between 2015 and 2016, stationary data distributions could be captured with time windows shorter than one year. This result is consistent with the fact that in that period of time, the mean latency drops below zero. Lastly, for the subset of keywords that we filter with our manually curated dictionary, we analyze their latency distributions in Figure 9. We observe that all identified categories (e.g., Worm, Trojan, Spyware) undergo a similar trend. However, the Worm category dropped by early 2014, and it disappeared almost completely in 2022. Despite this general result, we observe that some malware categories are still slightly actively developed in the most recent years, such as Ransomware, Trojan, Spyware, and Riskware.

Answer to RQ2: The majority of keywords no longer make their first appearance on these underground platforms since 2015. This means that the relevance of traditional hacking communities has been decreasing, but some novel malware is still being advertised. As a consequence, the CTI community should direct its efforts to more thriving communities.

7 DISCUSSION

7.1 Lessons Learned

We believe that our analysis highlights three lessons for the threat intelligence research community.

◇ **Underground communities are a valuable source of reactive CTI.** This is explained by the number of keywords (i.e., malware mentions and names of threat actors) that appeared first in discussions and were subsequently recorded in security reports. From a global point of view, as discussed in Section 6, more than 60% of the keywords appeared first in hacker forum discussions. With our framework, we were able to track the history of more than one malware that has been first discussed in the underground community and then employed in real cyberattacks. As anticipated in Section 1, the history of DarkComet started in 2008, when its author began the development as a challenging side project [7, 53]. With our framework, we were able to track discussions of such malware in 2008, where users shared tutorials for setting it up on target machines, and also track the post with the original release of “Dark Comet 2.0 RC4”. This malware played a key role four years later during the Syrian conflict, where it was used to spy on activists. However, this is not the only case: we were also able to track the history of other malware, such as DroidJack, which was announced in 2014 and employed in attacks targeted to Polish banking users just a few months later [46].

◇ **Hacker forum communities are less relevant, but they are still working on malware development.** In the last ten years, underground forums have been consistently lagging behind cybersecurity reports in terms of novel malware releases. The average timeliness of CTI extracted decreased until reaching negative values between 2015 and 2016, meaning that novel keywords are less likely going to be discussed first in forums. This result suggests that

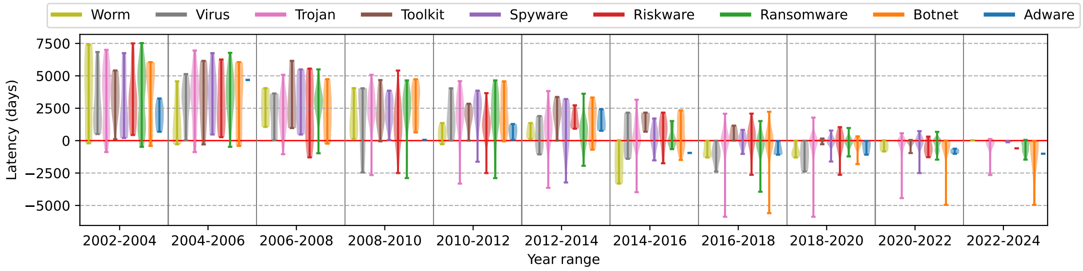


Figure 9: Violin plot of latency distribution for each identified malware category. We grouped samples biennially to reduce the size of the plot.

these long-existing communities have a lower impact on malware development, but we are still able to collect evidence of new malware being developed and advertised on such platforms up to these days. By filtering the matched keywords that have positive latency, we can track posts that advertise novel malware. For example, we find Typhon Reborn, a *"heavily refactored and improved version of the older and unstable Typhon Stealer"* [48], and XWorm V2.2, a Windows RAT with ransomware capabilities [13], both publicly advertised between 2022 and 2023.

◇ **Communities are probably moving to other platforms.** Since the volume of reports has remained relatively constant over the years while the volume of discussions in hacker forums appears to have decreased since the early months of 2022, and assuming that underground communities are still interested in malware development, the decrease of novel keywords extracted from traditional hacker forums could be due to users moving to other platforms, such as Telegram [49]. This motivates the need to monitor underground communities to avoid missing relevant information.

◇ **You should have known it earlier.** Let us take as an example the XRAT malware, also known as Quasar RAT, a publicly available tool for remote administration that is still employed by threat actors for cyberespionage and facilitating cyberattacks [11]. The history of XRAT seems to start in 2014 [11], while Malpedia [38], a publicly available and free malware inventory, tracks this malware from 2016[2]. With our framework, we detected early discussions of XRAT appearing in 2011, when it was advertised with a post that read: *"[RELEASE] Amazing RAT [...] It's not a virus. Download and run in a sandbox. Just run the XRat program, find Create Server and Create it"*. Other discussions followed in 2012, until in 2014, the likely author of the malware wrote the following post: *"[...] I won't give support for xRAT 1.0, this is a thread for xRAT 2.0. ;)"*. This supports our speculation that XRAT was indeed already circulating before 2014 and that, with our framework, we were able to trace its history before the release of XRAT 2.0, the version that drew public interest. Early detection of malware advertisements may allow practitioners to close the gap in the arms race with cyber attackers and reduce the latency in discovering novel developments in the community. We believe that analyses similar to the one proposed in this work should be applied to any emerging source of CTI to estimate the importance of the source and the relevance of the extracted information in terms of timeliness. In other words, if a source seems to anticipate malicious phenomena, it should be monitored more closely. Thus, CTI, besides researching the automatic extractions of relevant entities, should focus on the development of tools for monitoring underground communities and estimating their impact on cyber attacks (in this work, estimated through the timeliness of the information). Finally, we would like to emphasize the importance of avoiding this "error" in the future.

7.2 Threats to Validity

Sampling Bias. Our experimental results are mostly based on the information contained in the CrimeBB dataset [36] and threat reports collected with our crawlers (see Section 4). Despite being the largest collection of hacker forum posts currently available, CrimeBB may not cover the whole underground community. Some forums may have been overlooked by the original authors of the dataset, or their crawler may not have collected all the available information on the original platforms. Similar considerations can be made with respect to our CTI reports dataset. For example, the presence of concurrent discussions (see Section 6.1) may be impacted by sampling bias. Indeed, there may be prior discussions or reports mentioning a certain keyword that our framework could have failed to capture or could be completely missing in the analyzed data, making the matching entity fall under one of the first two categories (first appearance in forums, or first appearance in reports). However, we believe that these datasets are representative enough to estimate the relevance of CTI on such platforms.

Report and Forum Information Correctness. Some cybersecurity reports could provide incorrect information [56]. This issue could introduce noise in the keywords extracted by our framework from such reports (*e.g.,* reportedly incorrect malware names or threat actors). On the other hand, the information contained in underground forums could be considered unreliable. Users of these forums are often malicious individuals, and the published information may have been purposely manipulated (*e.g.,* advertising backdoored malware with a different name). Unfortunately, to the best of our knowledge, there is no way to assess the correctness of each threat report and forum post. Overall, this could impact the validity of the cyber-threat keywords extracted, introducing noise in our analysis. In general, we assume the correctness of the contained information coherently with other research works[44, 56].

English-only Forums. Given the additional challenges of processing non-English languages [55] (*e.g.,* Chinese, Russian), we reduced

[2] https://twitter.com/malwrhunterteam/status/789153556255342596

the scope of our analysis to English-speaking forums (see Section 4). Thus, our analysis does not consider whether the identified trends also apply to the missing communities. However, we focus our analysis on the most popular underground communities and prevalent language, showing that we are able to track the release of popular malware (see Section 7).

False Positives. Let us note that our system may produce false positive matches. These are represented by keywords matching between forums and reports that are not associated with any security-related threat. To reduce the number of false matches, we filter the matched keywords with a dictionary of known entities (see Section 5.3). Our results show that when filtering matches by known keywords, the final results still show similar trends. However, the number of matches is reduced. We attribute this phenomenon to false matches but also to keywords that may be missing from our dictionary while still being related to cyber threats. Unfortunately, to the best of our knowledge, there are no ground truth dictionaries that track all the known cyber threats.

Discussion Filtering. The huge volume of discussions makes an experimental analysis of the textual content challenging. To reduce the scope of our analysis to relevant content, we employ a BERT-based model trained on a custom built dataset (see Section 5.1). However, such a model could overlook relevant forum posts containing threat-related keywords, decreasing the number of total matches, or include irrelevant content, increasing the number of false positives. However, considering the high performances achieved by the model used in this work, we believe that the impact of this problem is limited.

8 RELATED WORKS

Monitoring and Alert Generation. A noteworthy area of CTI research is related to alert generation: large-scale attacks can be anticipated with proactive analysis of the platforms where security experts discuss novel CTI-related topics. Sapienza et al. [42] propose a framework and its subsequent extension [43] based on monitoring Twitter (recently renamed as "X") and hacker forums aiming at generating alerts for impending cyber threats. DISCOVER [43] involves collecting social media feeds from prominent figures in the cybersecurity sector, searching for content related to exploits and vulnerabilities, and applying text-mining techniques to keep important terms and eliminate irrelevant ones. Subsequently, the system checks if the terms identified during the filtering stage have been previously employed in hacking forums. Then, it reports the frequency of mentions and the content of posts, generating an alert. Differently from our approach, hacker forums are used to confirm whether certain keywords – identified from other sources – have been discussed. Sapienza et al. [42], similarly to DISCOVER [43], detects novel terminology and raises alarms based on a set of rules, but the input data comes from both hacker forums and Twitter communications. While sharing some similarities with both works [42, 43], our study has a completely different goal. The output of our framework is not alarms but essentially the complete list of keywords and their date of appearance in forums and reports, from which we conduct our longitudinal analysis. We analyze the relation between underground communities and historical cyber attack records, studying the role of underground communities and providing insights that are helpful to CTI practitioners and the research community. Our results confirm the utility of proactive approaches [42, 43] and promote the advancement of CTI research on this matter while studying emerging communities, as traditional ones seem to fall behind. In addition, while these works [42, 43] analyze short time windows of data that come from such sources, around 4 months of data, we analyze a much larger one, with around 20 years of data. Adewopo et al. [4] collect and filter tweets based on a fixed set of keywords and extract data from darknet marketplaces. They use a single processing pipeline conveying information from both deep and surface web data sources to generate a word embedding matrix. Then, a ML classifier marks data as either related to cyber threats or not, and in cases where it is relevant, the tool generates an alert. However, the authors consider only darknet marketplaces and not platforms meant for hosting discussions. Dionísio et al. [17] present a pipeline to process data extracted from Twitter and identify tweets containing information related to cybersecurity. Then, a deep learning model extracts named entities to generate a warning or detect intrusion indicators.

Analysis of Assets in Hacker Forums. Other strategies focus on classifying data from hacker forums, aiming to categorize the types of malware attached and model the specific topics addressed in discussions. Some works [5, 19] develop frameworks for the automatic collection and categorization of attachments in forum threads. Deliu et al. [15] develop machine learning models to classify posts extracted from hacker forums as CTI-relevant or irrelevant, and map relevant posts to different topics.

Relevance of CTI sources. Several studies have assessed CTI sources using both qualitative and quantitative approaches, notably focusing on threat intelligence feeds [18, 27, 31, 45] and platforms [50]. Li et al. [27] evaluated threat intelligence feeds, revealing variations in data types and suggesting larger feeds offer more reliable information. They also assessed the accuracy of Indicators of Compromise (IoCs), forensic artifacts such as malicious IPs and malware signatures, and observed consistent metric relationships over different periods. Schaberreiter et al. [45] proposed a methodology to gauge the trustworthiness of threat intelligence sources, emphasizing the need for metric reassessment with new threat information. However, their study primarily focused on methodology development rather than practical applications. Griffioen et al. [18] evaluated the quality of indicators from various open-source feeds, finding low overlap but different performances in terms of timeliness, originality, and sensitivity. Their analysis was limited to IP addresses. Mavzer et al. [31] introduced a "Trust and Quality Tool" to define the quality of threat intelligence data on a public sharing platform, highlighting potential improvements in reliability and information sharing maturity. Tundis et al. [50] chose Twitter as a CTI source for its relevance in security discussions, using a relevancy score and regression models to predict the timeliness of threat intelligence. This approach, however, was restricted to a single source and a single metric. First, these studies predominantly focus on IoCs, which, as pointed out by other research works [28], fail to provide a sufficient understanding of the cyber threat landscape. Importantly, none of these studies specifically focused on CTI information from underground hacker forums, which contrasts with platforms specifically designed for sharing structured intelligence or social networks meant for hosting generic discussions.

CrimeBB Analyses. Other works have analyzed the CrimeBB dataset [36], which, to the best of our knowledge, represents the largest dataset of hacker forum threads. Some application areas of the dataset include the analysis of key actors involved in cybercrime activities and the creation of graphs and social networks to study these underground communities [35, 37]. CrimeBB is also employed in studying the ecosystem and evolution of darknet markets [51]. Other studies [14, 20, 22] focus on the detection of novel information. However, our work focuses on the longitudinal examination of hacker forums, aiming to provide a chronological perspective on the relevance of such platforms in the context of CTI. We employ NLP techniques that enable us to trace the evolution of malicious entities embedded within the posts present in the dataset.

9 LIMITATIONS AND FUTURE WORKS

Besides the threats to validity reported in Section 7.2, our research study is affected by several limitations. The main issue is related to the absence of ground truth: a certain degree of false negatives and false positives in the matched topics cannot be avoided. As a consequence, the quality of the data selected for building our experimental dataset, both hacker forum discussions and CTI reports and news, impacts the final results of our work. The performances of our models inherently depend on the quality of NER dataset APT-NER [52] and the hacker forum dataset that we manually labeled. To the best of our knowledge, there are no benchmark datasets for NER models in the context of CTI, and one of the challenges of CTI research is related to the scarcity of labeled datasets. This analysis could have also been conducted with LLMs recently introduced by OpenAI, such as GPT-3.5 and GPT-4. However, the economic costs introduced could be non-negligible due to the size of the dataset, and they would require further analysis of hallucinations that may represent an additional threat to validity. In our analysis, we do not explore possible correlations among the different data sources. Specifically, there may be correlations in the matches between forums and specific threat report vendors (e.g., HF14 and Ahnlabs). We have not explored such correlations experimentally. However, we believe that the empirical results shown are noteworthy and that this aspect could be tackled in future works. Then, our analysis does not take into consideration the relevance of the CTI extracted from underground communities in terms of IoCs. Our focus is set on higher-level intelligence, *e.g.*, malware, and threat actor names, which permit us to gather insight into the actual trends of the underground communities. Finally, matching topics only by the presence of keywords may fail to capture the depth of certain topics and contextual meanings. As a possible extension to our framework, we plan to adopt more sophisticated techniques, such as word embedding, to analyze the CTI texts and better capture the contextual meaning of the discussions. Then, we could match similar topics by measuring the distance of their representation in latent space and repeat the analysis of the trends. This could allow us to conduct a more thorough analysis for the identification of recurrent themes, malware development lifecycles, and the patterns behind their diffusion. In addition, a deeper analysis, along with additional data provided by external services that track the diffusion of malware [3], could help in the quantification of the impact on cyber attacks of the tracked malware. A formalization of an impact metric to evaluate the *impact* of a source would be extremely valuable and a starting point for future work. Another interesting development could consist in connecting our analysis of the relevance of hacker forums with other popular communication channels, such as social networks (*e.g.*, Twitter/X) or messaging platforms (*e.g.*, Telegram).

10 ETHICAL ISSUES

The research presented in this paper demonstrates the potential of hacker forums as a resource for CTI, suggesting a more proactive approach to cybersecurity. However, the ethical and privacy concerns associated with accessing and analyzing data from these forums, where sensitive and potentially illicit information may appear, cannot be overlooked [6, 36]. To mitigate potential ethical issues, in the analysis presented in this work, we considered the suggestions made by Pastrana et al. [36], who discussed the ethical issues when analyzing data from underground forums. First, we limit the scope to security entities (i.e., keywords), anonymizing individuals and platforms and avoiding the disclosure of sensitive and personal data (e.g., posts can include private messages, e-mail addresses, and IP addresses). In addition, we take precautions against downloading malicious content. Secondly, the presentation of the results is objective, focusing on the data's implications for cybersecurity rather than commenting on the behaviors observed within these forums. Ultimately, our study not only advances the field of cybersecurity by promoting a shift toward more proactive strategies but also takes into consideration ethics and privacy issues.

11 CONCLUSION

In this paper, we discussed the historical relevance of hacker forums for anticipating potential trends in the cyber threat landscape. We extracted from over 88 million hacker forum posts around 2 million relevant posts and analyzed them along with 75,000 threat reports written in natural language, covering a period of around 20 years, from 2002 up to the early months of 2023. We defined a framework to conduct exploratory analysis, filter forum data, and apply optimal NLP techniques commonly employed in the CTI for identifying named entities from natural language text. We validated our approach through experimental evaluation, resulting in a pipeline for processing both hacker forum discussions and security reports with high accuracy. Our results showed that many topics discussed in forums are later reported in security literature. We believe that such results signal that hacker forums are essential sources for automatized proactive CTI approaches. Overall, our study lays the groundwork for the development of machine learning models for automated threat analysis.

ACKNOWLEDGMENTS

We thank our anonymous reviewers and shepherd for their useful comments and Cambridge Cybercrime Centre for granting us access to the CrimeBB dataset. Tommaso Paladini acknowledges support from TIM S.p.A. through the Ph.D. scholarship.

REFERENCES

[1] [n. d.]. Industrial-strength natural language processing. https://spacy.io/

[2] [n. d.]. Threat Intelligence Market Size, Share, Growth & Trends [2030] — fortunebusinessinsights.com. https://www.fortunebusinessinsights.com/threat-intelligence-market-102984. [Accessed 18-03-2024].

[3] abuse.ch. [n. d.]. URLhaus | Malware URL exchange. https://urlhaus.abuse.ch/

[4] Victor Adewopo, Bilal Gonen, and Festus Adewopo. 2020. Exploring Open Source Information for Cyber Threat Intelligence. In *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, December 10-13, 2020*, Xintao Wu, Chris Jermaine, Li Xiong, Xiaohua Hu, Olivera Kotevska, Siyuan Lu, Weija Xu, Srinivas Aluru, Chengxiang Zhai, Eyhab Al-Masri, Zhiyuan Chen, and Jeff Saltz (Eds.). IEEE, 2232–2241. https://doi.org/10.1109/BIGDATA50022.2020.9378220

[5] 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, ISI 2020, Arlington, VA, USA, November 9-10, 2020*. IEEE, 1–6. https://doi.org/10.1109/ISI49825.2020.9280548

[6] Randa Basheer and Bassel Alkhatib. 2021. Threats from the Dark: A Review over Dark Web Investigation Research for Cyber Threat Intelligence. *J. Comput. Networks Commun.* 2021 (2021), 1302999:1–1302999:21. https://doi.org/10.1155/2021/1302999

[7] Bitdefender. [n. d.]. Technical Advisory: Various Threat Actors Targeting ManageEngine Exploit CVE-2022-47966. https://www.bitdefender.com/blog/businessinsights/tech-advisory-manageengine-cve-2022-47966/

[8] Blackberry. [n. d.]. Agent Tesla Malware. https://www.blackberry.com/us/en/solutions/endpoint-security/ransomware-protection/agent-tesla

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

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

[11] CISA. 2018. *Quasar Open-Source Remote Administration Tool.* https://www.cisa.gov/news-events/analysis-reports/ar18-352a

[12] Richard Colbaugh and Kristin Glass. 2011. Proactive defense for evolving cyber threats. In *2011 IEEE International Conference on Intelligence and Security Informatics, ISI 2011, Beijing, China, 10-12 July, 2011*. IEEE, 125–130. https://doi.org/10.1109/ISI.2011.5984062

[13] Cyble. [n. d.]. EvilCoder Project Selling Multiple Dangerous Tools Online. https://cyble.com/blog/evilcoder-project-selling-multiple-dangerous-tools-online/

[14] Nathan Deguara, Junaid Arshad, Anum Paracha, and Muhammad Ajmal Azad. 2022. Threat Miner - A Text Analysis Engine for Threat Identification Using Dark Web Data. In *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*, Shusaku Tsumoto, Yukio Ohsawa, Lei Chen, Dirk Van den Poel, Xiaohua Hu, Yoichi Motomura, Takuya Takagi, Lingfei Wu, Ying Xie, Akihiro Abe, and Vijay Raghavan (Eds.). IEEE, 3043–3052. https://doi.org/10.1109/BIGDATA55660.2022.10020397

[15] 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 BigData 2017), Boston, MA, USA, December 11-14, 2017*, Jian-Yun Nie, Zoran Obradovic, Toyotaro Suzumura, Rumi Ghosh, Raghunath Nambiar, Chonggang Wang, Hui Zang, Ricardo Baeza-Yates, Xiaohua Hu, Jeremy Kepner, Alfredo Cuzzocrea, Jian Tang, and Masashi Toyoda (Eds.). IEEE Computer Society, 3648–3656. https://doi.org/10.1109/BIGDATA.2017.8258359

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/V1/N19-1423

[17] Nuno Dionísio, Fernando Alves, Pedro Miguel Ferreira, and Alysson Bessani. 2019. Cyberthreat Detection from Twitter using Deep Neural Networks. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*. IEEE, 1–8. https://doi.org/10.1109/IJCNN.2019.8852475

[18] Harm Griffioen, Tim Booij, and Christian Doerr. 2020. Quality evaluation of cyber threat intelligence feeds. In *Applied Cryptography and Network Security: 18th International Conference, ACNS 2020, Rome, Italy, October 19–22, 2020, Proceedings, Part II 18*. Springer, 277–296.

[19] John Grisham, Sagar Samtani, Mark W. Patton, and Hsinchun Chen. 2017. Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. In *2017 IEEE International Conference on Intelligence and Security Informatics, ISI 2017, Beijing, China, July 22-24, 2017*. IEEE, 13–18. https://doi.org/10.1109/ISI.2017.8004867

[20] Jack Hughes, Seth Aycock, Andrew Caines, Paula Buttery, and Alice Hutchings. 2020. Detecting Trending Terms in Cybersecurity Forum Discussions. In *Proceedings of the Sixth Workshop on Noisy User-generated Text, W-NUT@EMNLP 2020 Online, November 19, 2020*, Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi (Eds.). Association for Computational Linguistics, 107–115. https://doi.org/10.18653/V1/2020.WNUT-1.15

[21] HYPR. [n. d.]. EternalBlue. https://www.hypr.com/security-encyclopedia/eternalblue

[22] Risul Islam, Md Omar Faruk Rokon, Evangelos E. Papalexakis, and Michalis Faloutsos. 2020. TenFor: A Tensor-Based Tool to Extract Interesting Events from Security Forums. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2020, The Hague, Netherlands, December 7-10, 2020*, Martin Atzmüller, Michele Coscia, and Rokia Missaoui (Eds.). IEEE, 515–522. https://doi.org/10.1109/ASONAM49781.2020.9381356

[23] Youngjin Jin, Eugene Jang, Jian Cui, Jin-Woo Chung, Yongjae Lee, and Seungwon Shin. 2023. DarkBERT: A Language Model for the Dark Side of the Internet. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 7515–7533. https://doi.org/10.18653/V1/2023.ACL-LONG.415

[24] Hyeonseong Jo, Yongjae Lee, and Seungwon Shin. 2022. Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text. *Comput. Secur.* 120 (2022), 102763. https://doi.org/10.1016/J.COSE.2022.102763

[25] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality Enhanced Language Models for Open-Ended Text Generation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/df438caa36714f69277daa92d608dd63-Abstract-Conference.html

[26] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* 34, 1 (2022), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

[27] Vector Guo Li, Matthew Dunn, Paul Pearce, Damon McCoy, Geoffrey M. Voelker, and Stefan Savage. 2019. Reading the Tea leaves: A Comparative Analysis of Threat Intelligence. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, Santa Clara, CA, 851–867. https://www.usenix.org/conference/usenixsecurity19/presentation/li

[28] Zhenyuan Li, Jun Zeng, Yan Chen, and Zhenkai Liang. 2022. AttacKG: Constructing Technique Knowledge Graph from Cyber Threat Intelligence Reports. In *Computer Security - ESORICS 2022 - 27th European Symposium on Research in Computer Security, Copenhagen, Denmark, September 26-30, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13554)*, Vijayalakshmi Atluri, Roberto Di Pietro, Christian Damsgaard Jensen, and Weizhi Meng (Eds.). Springer, 589–609. https://doi.org/10.1007/978-3-031-17140-6_29

[29] Matteo Liberato. 2022. *SecBERT: Analyzing reports using BERT-like models.* Master's thesis. University of Twente.

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[31] Kadir Burak Mavzer, Ewa Konieczna, Henrique Alves, Cagatay Yucel, Ioannis Chalkias, Dimitrios Mallis, Deniz Cetinkaya, and Luis Angel Galindo Sanchez. 2021. Trust and quality computation for cyber threat intelligence sharing platforms. In *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 360–365.

[32] Rob McMillan. 2013. Definition: threat intelligence. *Gartner. com* 5 (2013).

[33] Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named Entity Recognition without Gazetteers. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*. The Association for Computer Linguistics, 1–8. https://aclanthology.org/E99-1001/

[34] OpenAI. 2024. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. https://openai.com/index/gpt-4/

[35] 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 - 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 11050)*, Michael D. Bailey, Thorsten

Holz, Manolis Stamatogiannakis, and Sotiris Ioannidis (Eds.). Springer, 207–227. https://doi.org/10.1007/978-3-030-00470-5_10

[36] Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. 2018. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 1845–1854. https://doi.org/10.1145/3178876.3186178

[37] Ildiko Pete, Jack Hughes, Yi Ting Chua, and Maria Bada. 2020. A Social Network Analysis and Comparison of Six Dark Web Forums. In *IEEE European Symposium on Security and Privacy Workshops, EuroS&P Workshops 2020, Genoa, Italy, September 7-11, 2020.* IEEE, 484–493. https://doi.org/10.1109/EUROSPW51379.2020.00071

[38] Daniel Plohmann, Martin Clauss, Steffen Enders, and Elmar Padilla. 2017. Malpedia: a collaborative effort to inventorize the malware landscape. *Proceedings of the Botconf* (2017).

[39] Md. Rayhanur Rahman, Rezvan Mahdavi-Hezaveh, and Laurie A. Williams. 2023. What Are the Attackers Doing Now? Automating Cyberthreat Intelligence Extraction from Text on Pace with the Changing Threat Landscape: A Survey. *ACM Comput. Surv.* 55, 12 (2023), 241:1–241:36. https://doi.org/10.1145/3571726

[40] Sagar Samtani, Kory Chinn, Cathy Larson, and Hsinchun Chen. 2016. AZSecure Hacker Assets Portal: Cyber threat intelligence and malware analysis. In *IEEE Conference on Intelligence and Security Informatics, ISI 2016, Tucson, AZ, USA, September 28-30, 2016.* IEEE, 19–24. https://doi.org/10.1109/ISI.2016.7745437

[41] Sangfor Technologies. 2023. Comparing Proactive vs. Reactive Cybersecurity in 2023. https://www.sangfor.com/blog/cybersecurity/proactive-vs-reactive-cybersecurity-2023

[42] Anna Sapienza, Alessandro Bessi, Saranya Damodaran, Paulo Shakarian, Kristina Lerman, and Emilio Ferrara. 2018. Early Warnings of Cyber Threats in Online Discussions. *CoRR* abs/1801.09781 (2018). arXiv:1801.09781 http://arxiv.org/abs/1801.09781

[43] Anna Sapienza, Sindhu Kiranmai Ernala, Alessandro Bessi, Kristina Lerman, and Emilio Ferrara. 2018. DISCOVER: Mining Online Chatter for Emerging Cyber Threats. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 983–990. https://doi.org/10.1145/3184558.3191528

[44] Kiavash Satvat, Rigel Gjomemo, and V. N. Venkatakrishnan. 2021. Extractor: Extracting Attack Behavior from Threat Reports. In *IEEE European Symposium on Security and Privacy, EuroS&P 2021, Vienna, Austria, September 6-10, 2021.* IEEE, 598–615. https://doi.org/10.1109/EUROSP51992.2021.00046

[45] Thomas Schaberreiter, Veronika Kupfersberger, Konstantinos Rantos, Arnolnt Spyros, Alexandros Papanikolaou, Christos Ilioudis, and Gerald Quirchmayr. 2019. A Quantitative Evaluation of Trust in the Quality of Cyber Threat Intelligence Sources. In *Proceedings of the 14th International Conference on Availability, Reliability and Security* (Canterbury, CA, United Kingdom) *(ARES '19)*. Association for Computing Machinery, New York, NY, USA, Article 83, 10 pages. https://doi.org/10.1145/3339252.3342112

[46] SecurityWeek. [n. d.]. Developers of Android RAT DroidJack Traced to India. https://www.securityweek.com/developers-android-rat-droidjack-traced-india/

[47] Sophos News. [n. d.]. Is the Angler exploit kit dead? https://news.sophos.com/en-us/2016/06/16/is-angler-exploit-kit-dead/

[48] Talos Intelligence. [n. d.]. Typhon Reborn V2: Updated stealer features enhanced anti-analysis and evasion capabilities. https://blog.talosintelligence.com/typhon-reborn-v2-features-enhanced-anti-analysis/

[49] The Hacker News. [n. d.]. Researchers Warn of "Eternity Project" Malware Service Being Sold via Telegram. https://thehackernews.com/2022/05/researchers-warn-of-eternity-project.html

[50] Andrea Tundis, Samuel Ruppert, and Max Mühlhäuser. 2020. On the automated assessment of open-source cyber threat intelligence sources. In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part II 20.* Springer, 453–467.

[51] Anh V. Vu, Jack Hughes, Ildiko Pete, Ben Collier, Yi Ting Chua, Ilia Shumailov, and Alice Hutchings. 2020. Turning Up the Dial: the Evolution of a Cybercrime Market Through Set-up, Stable, and Covid-19 Eras. In *IMC '20: ACM Internet Measurement Conference, Virtual Event, USA, October 27-29, 2020.* ACM, 551–566. https://doi.org/10.1145/3419394.3423636

[52] Xuren Wang, Songheng He, Zihan Xiong, Xinxin Wei, Zhengwei Jiang, Sihan Chen, and Jun Jiang. 2022. APTNER: A Specific Dataset for NER Missions in Cyber Threat Intelligence Field. In *25th IEEE International Conference on Computer Supported Cooperative Work in Design, CSCWD 2022, Hangzhou, China, May 4-6, 2022.* IEEE, 1233–1238. https://doi.org/10.1109/CSCWD54268.2022.9776031

[53] Wired. [n. d.]. How the Boy Next Door Accidentally Built a Syrian Spy Tool. https://www.wired.com/2012/07/dark-comet-syrian-spy-tool/

[54] Nianwen Xue. 2011. Steven Bird, Evan Klein and Edward Loper. *Natural Language Processing with Python.* O'Reilly Media, Inc 2009. ISBN: 978-0-596-51649-9. *Nat. Lang. Eng.* 17, 3 (2011), 419–424. https://doi.org/10.1017/S1351324910000306

[55] Min Zhang. 2010. Introduction to Chinese Natural Language Processing Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang (Chinese University of Hong Kong, Hong Kong Polytechnic University, City University of Hong Kong, and San Diego State University) Princeton, NJ: Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 4), 2010, x+148 pp; paperbound, ISBN 978-1-59829-932-8, $40.00; e-book, ISBN 978-1-59829-933-5, $30.00 or by subscription. *Comput. Linguistics* 36, 4 (2010), 777–780. https://doi.org/10.1162/COLI_R_00024

[56] Ziyun Zhu and Tudor Dumitras. 2018. ChainSmith: Automatically Learning the Semantics of Malicious Campaigns by Mining Threat Intelligence Reports. In *2018 IEEE European Symposium on Security and Privacy, EuroS&P 2018, London, United Kingdom, April 24-26, 2018.* IEEE, 458–472. https://doi.org/10.1109/EUROSP.2018.00039

A ABLATION STUDY

In this section, we provide the tables showing the complete results of our ablation study. In our study, we greedily remove the steps that have a negative impact on the performances of the selected model. For the model that processes hacker forum posts, as shown in Table 2, the *Misspelling Correction* has a negative impact on the performances of the model. Therefore, we remove it from the pipeline and observe the performances on the test set in Table 4a. The results show that an additional removal is required, this time for the Synonym Homogenization step. Without such a step, the model achieves a final F1-score of 82.15% on our test set (see Table 4b). The model that processes cybersecurity reports, instead, does not require other steps to be removed from the pipeline, as shown in Table 5.

Table 4: Second and third iteration of our ablation study of the NLP pipeline on hacker forum data.

(a) Second iteration of our ablation study of the NLP pipeline without the *Misspelling Correction* step for hacker forums data. According to the recorded weighted average F1-scores on the test set, removing the *Synonym Homogenization* step improves the performances of the DarkBERT model.

Model	Tokenization	NLP pipeline without Misspelling Correction	NLP pipeline without Misspelling Correction and:						
			Unrelated content removal	Passive/active conversion	Synonym homogenization	Stopwords removal	Internet slang removal	Pronouns and subject ellipsis resolution	Aliases handling
DarkBERT [23]	Cased	79.95%	79.24%	79.50%	**82.15%**	78.04%	79.15%	79.16%	79.47%

(b) Third iteration of our ablation study for hacker forums data, without *Misspelling Correction* and *Synonym Homogenization.* According to the recorded weighted average F1-scores on the test set, the best model is DarkBERT [23] without the *Misspelling correction* and *Synonym Homogenization* steps.

Model	Tokenization	NLP pipeline without Misspelling Correction and Synonym Homogenization	NLP pipeline without Misspelling Correction, Synonym Homogenization, and:					
			Unrelated content removal	Passive/active conversion	Stopwords removal	Internet slang removal	Pronouns and subject ellipsis resolution	Aliases handling
DarkBERT [23]	Cased	**82.15%**	81.82%	81.62%	81.46%	79.06%	80.44%	80.79%

Table 5: Second iteration of our ablation study of the NLP pipeline without the *Misspelling Correction* step for cybersecurity reports. According to the recorded weighted average F1-scores on the test set, the best model is DarkBERT [23] without *Misspelling Correction.*

Model	Tokenization	NLP pipeline without Misspelling Correction	NLP pipeline without Misspelling Correction and:						
			Unrelated content removal	Passive/active conversion	Synonym homogenization	Stopwords removal	Internet slang removal	Pronouns and subject ellipsis resolution	Aliases handling
DarkBERT [23]	Cased	**86.33%**	85.50%	86.16%	85.49%	85.99%	85.83%	86.16%	82.33%