On the Application of Natural Language Processing for Advanced OSINT Analysis in Cyber Defence

Florian Skopik, Benjamin Akhras, Elisabeth Woisetschläger,
Medina Andresel, Markus Wurzenberger, Max Landauer
firstname.lastname@ait.ac.at
AIT Austrian Institute of Technology
Vienna, Austria

ABSTRACT

Open Source Intelligence (OSINT), in addition to closed military sources, provides timely information on emerging cyber attack techniques, attacker groups, changes in IT products, policy updates, recent events, and much more. Often, dozens of analysts scour hundreds of sources to gather, categorize, cluster, and prioritize news items, delivering the most pertinent information to decision makers. However, the sheer volume of sources and news items is continually expanding, making manual searches increasingly challenging. Moreover, the format and presentation of this information vary widely, with each blog entry, threat report, discussion forum, and mailing list item appearing differently, further complicating parsing and extracting relevant data. The research projects NEWSROOM and EUCINF, under the European Defence Fund (EDF), focus on leveraging Natural Language Processing (NLP) and Artificial Intelligence (AI) to enhance mission-oriented cyber situational awareness. These EDF initiatives are instrumental in advancing Taranis AI, a tool designed to categorize news items using machine learning algorithms and extract pertinent entities like company names, products, CVEs, and attacker groups. This enables the indexing and labeling of content, facilitating the identification of relationships and grouping of news items related to the same events – a crucial step in crafting cohesive "stories." These stories enable human analysts to swiftly capture the most significant current "hot topics", alleviating them from the task of consolidating or filtering redundant information from various sources. Taranis AI further enhances its capabilities by automatically generating summaries of reports and stories, and implementing a collaborative ranking system, among other features. This paper serves as an introduction to Taranis AI, exploring its NLP advancements and their practical applications. Additionally, it discusses lessons learned from its implementation and outlines future directions for research and development.

CCS CONCEPTS

• **Security and privacy**; • **Information systems → Information retrieval**; • **Computing methodologies → Natural language processing**; • **Applied computing → Cyberwarfare**;

This work is licensed under a Creative Commons Attribution International 4.0 License.

ARES 2024, July 30–August 02, 2024, Vienna, Austria
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1718-5/24/07
https://doi.org/10.1145/3664476.3670899

KEYWORDS

OSINT analysis, NLP, cyber defence, situational awareness

ACM Reference Format:
Florian Skopik, Benjamin Akhras, Elisabeth Woisetschläger,, Medina Andresel, Markus Wurzenberger, Max Landauer. 2024. On the Application of Natural Language Processing for Advanced OSINT Analysis in Cyber Defence. In *The 19th International Conference on Availability, Reliability and Security (ARES 2024), July 30–August 02, 2024, Vienna, Austria.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664476.3670899

1 INTRODUCTION

Open-source intelligence (OSINT) involves gathering and analyzing data from publicly available sources to generate actionable intelligence [2]. The literature [16] distinguishes between technical Cyber Threat Intelligence (CTI), as well as tactical and strategic CTI. Technical CTI primarily comprises simple data used to configure security systems, such as indicators for Security Information and Event Management (SIEM) systems, domain names for blocking in name servers or proxies, and execution patterns for blocking in Endpoint Detection and Response (EDR) solutions. In contrast, tactical and strategic CTI is more nuanced, often presented in natural language on various news sites and security platforms. It encompasses a wide range of information, including updates on new threat actors, security product features, breaches, incidents, campaigns, vulnerabilities, patches, mitigations, countermeasures, exploits, post-exploitation activities, policy developments (such as GDPR-related lawsuits), updates on security standards (ISO, BSI, ANSI, CIS, OWASP, etc.), mergers, acquisitions, and failures, among other security-related topics.

In the military context of cyber information warfare (CIW), CTI plays a crucial role in identifying disinformation campaigns and attempts to manipulate political and public opinion, often through the dissemination of fake news, at an early stage. Additionally, CTI may provide indicators of adversarial activities, frequently perpetrated by hostile states and advanced persistent threats (APTs). These indicators, linked to CIW activities, can signal the onset of attacks, such as website defacement, cross-site scripting (XSS), social media account hijacking, and the dissemination of deepfakes. Consequently, CTI serves as a valuable source of information for promptly detecting adversarial CIW activities and identifying threat actors, along with their tactics, techniques, and procedures (TTPs) [2].

Gathering public information on potential security threats, vulnerabilities, trends, attacker TTPs, as well as new risks, is essential for maintaining situational awareness and initiating early countermeasures. The typical Open-Source Intelligence (OSINT) workflow [2], as illustrated in Fig. 1, involves five phases: (i) direction

and planning, (ii) collection, (iii) processing, (iv) analysis, and (v) production and dissemination.

In essence, national authorities and military analysis centers, after determining the requirements and type of relevant insights, conduct broad-scale collection from potentially hundreds of sources, sifting through thousands of articles daily. They then analyze this information for relevant content to create what are referred to as "products", essentially reports tailored for specific constituencies, such as commanders, to support decision-making processes, particularly in military contexts. It's evident that the quality of these reports is highly dependent on the sophistication of the analysis phase.

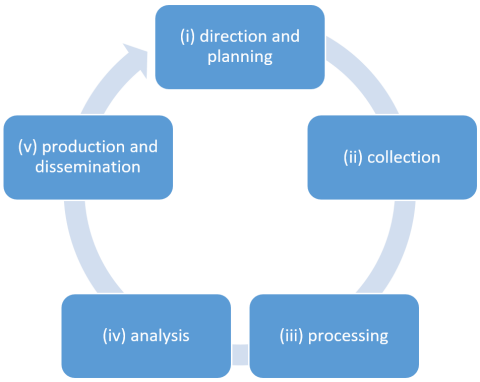

Figure 1: The standard OSINT workflow.

However, the ingestion, analysis, and utilization of semantically richer "soft" Cyber Threat Intelligence (CTI) pose greater challenges compared to well-structured, machine-readable technical CTI. Soft CTI typically manifests as unstructured, free-form text containing high-level, often ambiguous strategic information intended for human consumption. Consequently, within complex analysis workflows, it is primarily consumed by human analysts. This manual process is laborious, resource-intensive, and prone to errors. The human element not only slows down the analysis, but also significantly impedes scalability. With the increasing number of OSINT sources, the wide variety of different formats and appearances, and the frequency of published articles, there is a pressing need for new analysis techniques to keep pace with these developments and ensure that no critical information is overlooked. Fortunately, advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI) have been remarkable in recent years [15].

Military operations, in particular, stand to benefit from augmenting CTI processes with AI capabilities. This includes functionalities such as mission-oriented vulnerability report filtering, automatic clustering of stories, and notification of relevant military personnel about emerging threats pertinent to their operational areas. Additionally, at both tactical and strategic levels, militaries could leverage AI-supported CTI applications for generating current threat assessments, tracking changes in the threat landscape, and identifying emerging threats at an early stage.

Through stakeholder engagements and workshops with end-users from both civil and military domains, including CERTs (Computer Emergency Response Teams) and ISACs (Information Sharing and Analysis Centers), we have identified five essential key questions that human analysts encounter daily. The objective of Taranis AI is to support these user stories with appropriate technical solutions:

- *User Story 1:* What are the "hot topics" of the last 24 hours?
- *User Story 2:* What do we know about a specific entity? (e.g., a vulnerability, malware, company, product, person, location etc.)
- *User Story 3:* How can I find more related news items after reading this interesting article?
- *User Story 4:* Which news items are pertinent to my mission?
- *User Story 5:* How can I efficiently sum up my findings for my commander or operators of military IT services?

In course of several European Defense Fund projects, especially the EDF projects NEWSROOM[1] and EUCINF[2], as well as the civil CEF project AWAKE[3], we are investigating the integration of NLP and AI methods into Taranis AI, a fork of the well-known OSINT gathering tool Taranis-NG [13], to support these user stories. The contributions of this paper are therefore the discussion of candidate technologies and approaches to optimize the OSINT analysis workflow and the introduction of the new open-source fork Taranis AI [9] in the course of a case study.

The remainder of this paper is organized as follows. Section 2 summarizes important background and outlines related work. Section 3 highlights the use of NLP and AI for OSINT gathering and analysis. Our implementation in Taranis AI is further described in Sect. 4. We discuss the applicability, shortcomings and planned extensions in Sect. 5. Finally, Sect. 6 concludes the paper.

2 BACKGROUND AND RELATED WORK

We explore the current landscape of OSINT analysis using a state-of-the-art tool, Taranis-NG [13], and delve into the underlying principles of natural language processing (NLP) and recommender systems, which serve as the foundational basis for several powerful extensions.

2.1 OSINT Analysis Tools

OSINT stands for "Open Source Intelligence" and refers to the collection and analysis of information originating from publicly available sources. It encompasses a wide range of information, including social media, online news sources, government documents, and other freely accessible resources. While our work on Taranis AI may be best compared with a specialized newsreader, the domain of OSINT analysis tools, which potentially can contribute to and interact with Taranis AI, is much broader.

Maltego[4] specializes in uncovering relationships among people, companies, domains and publicly accessible information on the Internet. It's also known for taking the sometimes enormous amount of discovered information and plotting it all out in easy-to-read charts and graphs.

[1]Funded under the European Defence Fund (EDF) with grant number 101121403.
[2]Funded under the European Defence Fund (EDF) with grant number 101121418.
[3]Funded under the Connecting Europe Facilities (CEF) program with grant number INEA/CEF/ICT/A2020/237311.
[4]https://www.maltego.com/

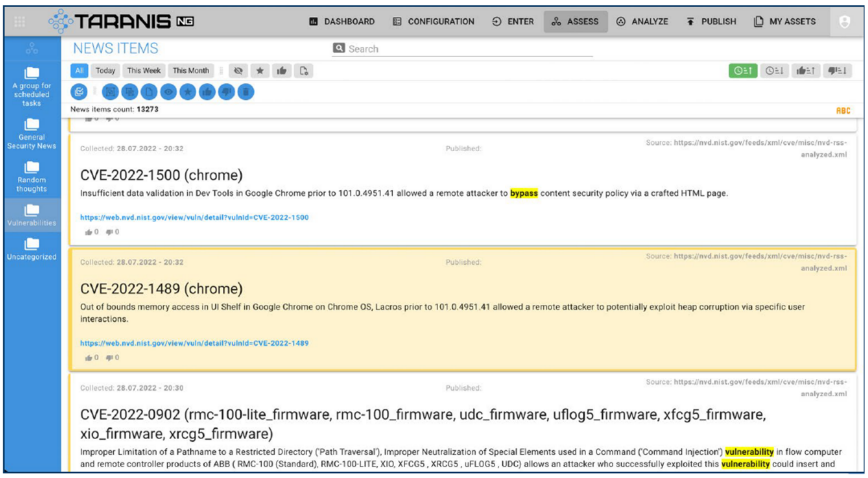

Figure 2: Taranis-NG GUI supporting the analysis process (see tabs at the top).

Several search engines specialize in crawling and collecting open-source network data, such as IP addresses, domains, URLs, hashes, ASNs, Bitcoin wallet addresses, and various indicators of compromise (IOCs). Among common solutions are VirusTotal[5], urlscan.io[6], Censys[7], and Shodan[8]. Third-party software commonly uses these sources, such as Mitaka[9] and Sputnik[10]. Other search engines preserve historic versions of web pages as well as entire leaked data sets that are otherwise removed from the web, such as IntelligenceX[11].

Similarly, tools for red team reconnaissance aim at collecting open-source data about a specific target and collect information from Websites, DNS infrastructure, whois databases and so on. A well-known solution is Spiderfoot[12]; another one Recon-ng[13]. As the name implies, BuiltWith[14] finds out what popular websites are built with, for example, detect whether a website is using WordPress, Joomla, or Drupal as its CMS and provide further details. Metagoofil[15] extracts metadata from public documents and can investigate almost any kind of document that it can reach through public channels.

For the use cases motivating our work, we refer to taranis3[16] and taranis-ng[17] as the starting point. These tools allow gathering news items from potentially hundreds of sources and allow sorting and filter them in the course of an advanced analysis workflow.

2.2 OSINT Analysis Workflow with Taranis-NG

Taranis-NG serves as an OSINT gathering and analysis tool designed for CSIRT teams and organizations [13]. It facilitates OSINT crawling, assessment, and reporting, fosters team-to-team collaboration, and includes a user portal for straightforward self-asset management. In essence, the solution is capable of crawling various pre-configured data sources, such as websites or social media channels, to gather unstructured news items. Analysts then process these items to compile structured report items, which are subsequently utilized to generate products like PDF files for final publication.

Figure 2 illustrates the Taranis-NG graphical user interface (GUI), showcasing recently gathered OSINT items. The solution aligns with the standard OSINT workflow outlined in Fig. 1 by offering essential features of the individual phases through dedicated assess, analyze, and publish tabs (as denoted by the tab names at the top). Following the determination of information requirements from "customers", an OSINT analyst navigates through these tabs to create a report tailored to the recipients' needs.

2.3 NLP, Recommender Systems and AI

Natural Language Processing (NLP) has emerged as an own field of AI that enables computers to understand, interpret, and interact with human language. Over the years, NLP has witnessed significant advancements, with breakthroughs in deep learning and transformer-based models, such as BERT and GPT-3 [17]. These advancements have opened up new possibilities in various sectors, including cyber situational awareness (CSA). With NLP it is possible to ingest and process any kind of freeform text, ranging from formal documents and news reports to informal messages from ad-hoc chats and e-mail.

From the literature, mainly transformer-based models stick out of the masses. They have overcome the limitations of recurrent neural networks (RNNs) and convolutional neural networks (CNNs) by introducing self-attention mechanisms [17]. This way, these models are able to process entire sequences of text simultaneously, making them much more efficient than RNNs and CNNs. Furthermore, the semantic information of text sequences are taken into account. Bidirectional encoder representations from transformers (BERTs) were introduced by Google in 2018 [17]. They are pre-trained language models that use a bidirectional approach to capture the context

[5]https://www.virustotal.com
[6]https://urlscan.io/
[7]https://censys.com/
[8]https://www.shodan.io
[9]https://chromewebstore.google.com/detail/mitaka/
[10]https://github.com/mitchmoser/sputnik
[11]https://intelx.io/
[12]https://github.com/smicallef/spiderfoot
[13]https://github.com/lanmaster53/recon-ng
[14]https://builtwith.com/
[15]https://github.com/laramies/metagoofil
[16]https://github.com/NCSC-NL/taranis3
[17]https://github.com/SK-CERT/Taranis-NG

from both sides of a word. These kinds of models are often first choice for common NLP tasks, including sentiment analysis and text classification. Generative pre-trained transformers (GPTs), such as GPT-3 and GPT-4 released by OpenAI, are massive language models and among the largest language models today. GPT's ability to generate human-like text has tremendous impact on numerous applications and is also well applicable in the area of cyber situational awareness.

Besides these models, it is important to understand the implications of some basic model concepts, such as transfer learning [12] and multimodal NLP [8]. Transfer learning in the field of NLP usually means fine-tuning pre-trained models like BERT and GPT for specific tasks. Thus, models are literally transferred from one domain to another application domain. This reduces the effort and amount of time required to make the model fit for purpose tremendously compared to training a model from scratch. Multimodal NLP means to extend the focus of NLP from basic text to further sorts of data, including images, video, and speech. This is an essential capability for situational awareness to, for instance, ingest audio via speech-to-text transcriptions.

Besides the application of NLP to enrich news items and extract important entities, techniques of recommender systems are of great help to support the OSINT analysis too. For instance, collaborative filtering [3] is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (hence, the collaborative aspect). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person. This approach is well applicable to recommend articles of interests to analysts. Similarly, based on information about which news articles have been considered valuable in the past (e.g., because they have been added to reports or were upvoted by an analyst), new articles that have similar properties, e.g., being from the same source, author or writing style, may be ranked higher and recommended to an analyst for further investigation [11].

3 NATURAL LANGUAGE PROCESSING AND AI-ASSISTED ANALYSIS

All OSINT sources deliver articles which are collected through bots in raw format. After that, a complex NLP and AI pipeline is triggered to process each single article, also called news item, and to make it easier digestible for human analysts.

3.1 Named Entity Recognition

Named Entity Recognition (NER) aims to identify and classify named entities mentioned in unstructured text into predefined categories like person names, organizations, and locations [5]. NER goes beyond simple tagging or keyword extraction by associating each named entity with a specific type, providing a significant advantage when searching for relevant articles. For example, a term recognized by NLP as a location, such as "China", may be less specific from a cyber security perspective compared to a term representing a particular Advanced Persistent Threat (APT) group. Similarly, a widely known product like "Microsoft Windows" may hold less value than an Indicator of Compromise (IoC), such as a SHA1 sum, which refers to a specific attack technique or malware. Product names may appear in articles in various contexts, making them less reliable indicators.

Currently, numerous pre-trained language models are available that promise reliable extraction of common named entity types from text. Some of these models are more complex, while others prioritize speed. The key is to select a model that delivers satisfactory results and can complete its task within a manageable timeframe. In Taranis AI, we utilize the Flair NER-multi model[18] for both German and English articles [1]. These models have demonstrated their effectiveness in extracting generally applicable concepts such as persons and locations.

Topic modeling is the process of creating models that recognize domain-specific concepts, as outlined in [14]. In the context of cyber security, this includes identifying APT groups, IT products, vendors, and other relevant entities. In addition to topic modeling, we leverage a variety of automatically generated lists[19] for entities that typically remain static and have a limited set size. An example of this is countries, which remain relatively constant over time. However, this approach may not be suitable for entities like persons, which are more dynamic. Furthermore, we employ regular expressions to identify CVE numbers and Indicators of Compromise (IoCs), such as hash sums in various forms, file paths, IP addresses, and file names. For IoCs specifically, we utilize IoC Finder, as demonstrated in [7].

NER plays a crucial role as a prerequisite for User Story 3, which involves identifying related articles reporting on the same entity.

3.2 Advanced Search and Filtering

Combining full-text search with entity types identified by NER enhances the power of full-text search. This integration allows for more precise and targeted searches, enabling users to distinguish between different meanings of terms. For example, understanding that "Proton" refers to a product of interest rather than a subatomic particle, or that "Windows" pertains to Microsoft's operating system, illustrates the utility of this feature. Similarly, disambiguating person names with different meanings enhances search accuracy. Named entities significantly contribute to the search and filtering of news items and facilitate grouping articles that cover the same topic [5]. Standard features include additional sorting, filtering, and ranking parameters, such as publication date or article ratings.

A novel aspect of Taranis AI is its support for collaborative search and filtering [3]. Multiple analysts typically access the same database through the web-based front-end, either serving different stakeholders or collaboratively compiling reports for decision-makers. In the latter scenario, analysts may share search queries or receive recommendations based on previous searches. Furthermore, articles rated as less helpful by one analyst can be filtered out for others, if this feature is enabled. This collaborative approach efficiently processes the vast information space, preventing analysts from re-reading articles that have already been analyzed by others.

Together with NER, full-text search and filtering form the foundation for implementing User Story 2.

[18] https://huggingface.co/flair/ner-multi
[19] https://github.com/taranis-ai/wordlists

3.3 Relevance Ranking

In information science and information retrieval, relevance refers to how well a retrieved document or set of documents aligns with the information need of the user [3]. In our context, the goal is to ensure that the system presents the most useful news items to a human analyst, optimizing their analysis time. This concept is closely tied to advanced search and filtering, as discussed earlier, but extends further with mechanisms from recommender systems [11].

In Taranis AI, we track user activities to adjust relevance ranking. For instance, we monitor which news items have been incorporated into a product or associated with a story, allowing us to understand what types of articles have been deemed useful in the past. Our solution endeavors to estimate relevant properties, including content-related factors such as specific named entities (e.g., products or APT groups) and meta-properties like article length, source type, and author names, to recommend future articles. This predictive approach enables the system to anticipate the relevance of forthcoming articles for an analyst.

In addition to implicit user actions, explicit user feedback is also considered. An upvote/downvote mechanism enables analysts to provide feedback on articles, thereby rating their quality. Consequently, news items with similar content, style, or from the same source may be considered less relevant in the future. The overarching idea is that over time, the system learns user preferences, identifies common properties of "good" articles, and potentially rates the relevance not only of articles but also of the sources that provide them. Ultimately, relevance ranking assists in identifying "interesting items" based on their general importance, when combined with collaborative filtering, and personal preferences derived from historic actions.

By implementing relevance ranking and leveraging mechanisms from recommender systems, we effectively address User Story 4.

3.4 Summary and AI-assisted Report Creation

One of the most time-consuming issues of human OSINT analysis is to read through lengthy texts only to find out that the content is irrelevant. Therefore, meaningful summaries of longer texts are key to speed up the process. Such summaries can effectively be created using NLP models and help to condense lengthy texts to their essential parts for quick decisions on their relevance. Not only single news items might be subject to summarization, but also collection of news items, so called stories or whole discussion forum threads are much more digestible this way.

In addition to summarizing content from input sources, summaries of created reports enhance information sharing. If an analyst has curated a report on a specific topic of interest (e.g., the activities of an APT group or details about a specific malware), main findings and commonalities among selected news items serve as excellent input for automatically generated executive summaries. While the feature to map data to predefined report templates is currently under development and not fully integrated into Taranis AI, AI has the potential to assist in learning suitable parameters such as appropriate length, wording, or content [15].

In the future, AI may also recommend articles to include in reports based on past selections by using concepts from recommender systems research [3]. This capability could streamline the report creation process by suggesting articles that align with previously included content.

Overall, building summaries of larger bodies of texts and employing AI-assisted report creation lays the groundwork for addressing User Story 5.

3.5 Topic and Story Clustering

Large-scale events and news with potentially high impact are usually provided not just by one, but multiple sources at approximately the same time with mostly similar content. Similar to newspapers, all the common OSINT sources report then about the same events or same topics. This is a huge burden for analysts as it means that they get the same information from multiple sources that use slightly different text. Figure 3 shows an example of a critical and already exploited vulnerability for the popular WS-FTP server. Here, multiple sources report about that, for instance, bleepingcomputer.com, csirt.divd.nl, feeds.feedbruner.com, and theregister.co.uk.

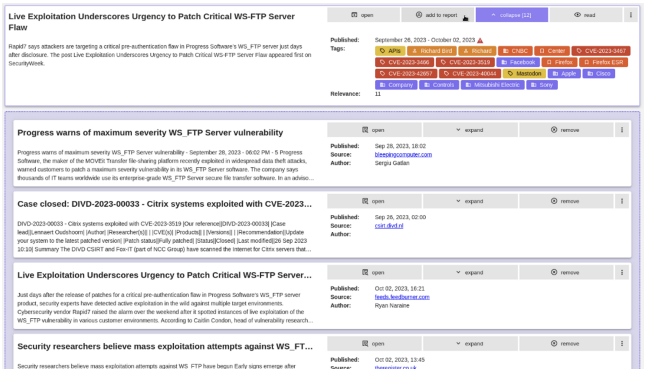

Figure 3: Story clustering example.

We adapted the story clustering solution presented in [6] to our requirements. In particular, in our approach we use weights to construct the keywords co-occurrence graph, and Louvain community detection algorithm. Furthermore, in order to detect if two news articles report about the same event, unlike the original approach we use a pre-trained multi-lingual transformer model to compute cosine similarity between the texts. This enables our solution, Taranis AI, to automatically cluster multiple news items that report on the same story, despite variations in title and writing style. The top box presents the summary of this story along with identified tags using NER. It also highlights that relevant articles were published from September 26, 2023, to October 2, 2023. Below the summary, the relevant news items comprising this story are displayed.

Clustering related articles and generating stories significantly reduces the number of articles an analyst must review. In our tests, we consistently reduced the number of daily news articles by approximately 40-80% using this approach. Additionally, Taranis AI offers the optional feature of visualizing the cluster's development over time. This feature illustrates the frequency of articles published on a particular topic, which is particularly valuable for broader, less specific topics. Analysts can leverage this visualization to uncover publishing trends related to specific technologies, vendors, countries, or individuals. Thus, we effectively address User Story 1.

4 THE TARANIS AI PROJECT: A DEMONSTRATION OF CAPABILITIES

In order to demonstrate the implementation of the discussed features we are going to walk through a short case study in which an OSINT analyst from a military CERT ingests all sorts of news from more than 100 OSINT channels with the aim to compile a report about new and severe vulnerabilities that are relevant for operators of military IT systems.

4.1 Deploy Taranis AI

The default way to deploy Taranis AI is via docker compose[20]. Deploying it via alternative container engines like podman[21] works as well. The listing in Fig. 4 describes the standard way to get an instance up and running. Notice, configure settings in .env if needed, especially JWT secrets and DB passwords should be regenerated. The default credentials are user / user and admin / admin.

```
# Deployment -- clone via git
git clone --depth 1 https://github.com/taranis-ai/taranis-ai
cd Taranis AI/docker/
## Configuration
cp env.sample .env
## Startup & Usage
docker compose up -d
## Use the application
http://<URL>:<TARANIS_PORT>/login
```

Figure 4: Deployment of Taranis AI.

4.2 Setting Things up: Configuring Taranis AI

After logging in with the admin credentials, the first step is to define "sources" from which OSINT news items shall be gathered from (cf. Figure 5). Each source is defined by providing at least a name, an url, and a collector to be used. In case of RSS feeds this is rather simple. For other sources, such as e-mail, ticket system (e.g., Request Tracker[22] and JIRA[23]) or Web Site collection using Selenium[24], a more fine-grained configuration may be required. Several common sources are already defined by default.

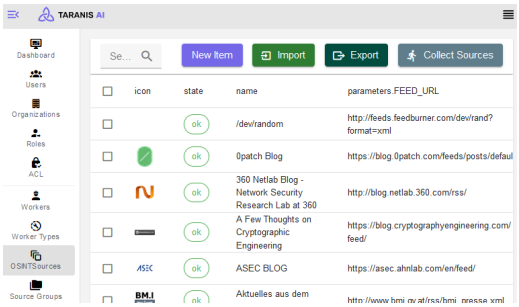

Figure 5: Configuration of OSINT sources in the Admin panel.

[20] https://docs.docker.com/compose/
[21] https://podman.io/
[22] https://github.com/bestpractical/rt
[23] https://www.atlassian.com/software/jira
[24] https://www.selenium.dev/

Additionally, sources can be clustered into source groups. This is particularly helpful when applying bots to pre-process items, such as a tagging bot or a summary bot. Here differently configured bots may be applied to different groups of sources. An example would be the application of bots that extract IoCs (such as sha1sums, paths, file names etc.) which makes sense for rather technical sources but is less useful if not disturbing for general news items from the daily press.

Bots are a means to transform ingested news items or to enrich them. Several bots are available to extract IoCs, tag news items using keyword matching with predefined lists or named entity recognition (NER), create summaries of lengthy articles, or create whole stories by grouping related articles as described before. Bots are applied to all sources within a source group. Moreover, the order of bots can be configured too. For instance, it makes sense to run keyword extraction using regular expressions or matching with predefined lists first, and in case not enough relevant keywords were detected, apply additional more resource-intensive NLP algorithms to determine useful keywords. Figure 6 shows the details.

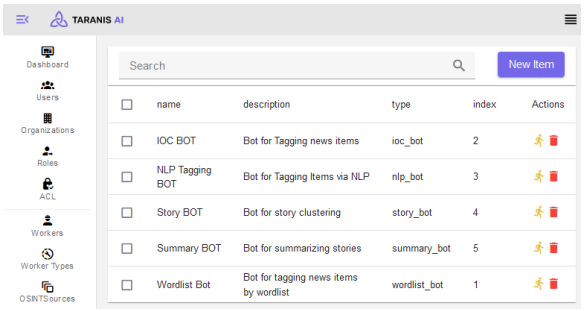

Figure 6: Configuration of bots.

Taranis AI comes with some predefined word lists, which are the basis for the simplest form of tagging. Predefined lists for APT groups (see Figure 7), IT products extracted from the CVE database, Vendors extracted from the CVE database, countries, large organizations etc. exist. Further lists can easily be added, e.g., municipalities of a country. This way, we can ensure that whenever one of these manually predefined keywords appears in a news item, Taranis AI tags the article accordingly. For some highly volatile categories, such as person names, it usually makes more sense to apply the NLP bot.

The goal of Taranis AI is to support analysts with creating products which are published via different channels. Therefore, Taranis AI allows the creation of different types of products that are rendered using standard formats, e.g., html, pdf, or text. It can even push reports to MISP[25]. Additionally, reports may be created for different constituencies in a certain format and structure, which is efficiently supported by a templating mechanism that determines how news items are integrated and rendered in a product type. Figure 8 shows an example of an html product of type "CERT Daily Report" with a certain structure and mechanisms to map selected stories to the actual report. Several product types are configured by default, further can be created on demand. Eventually, these

[25] https://www.misp-project.org/

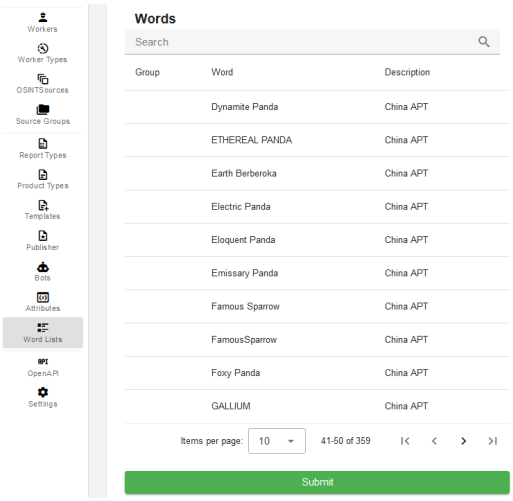

Figure 7: A WordLists example: APT Groups.

products are pushed to publishers which put them on Web servers, send them by mail, or put them on a MISP server.

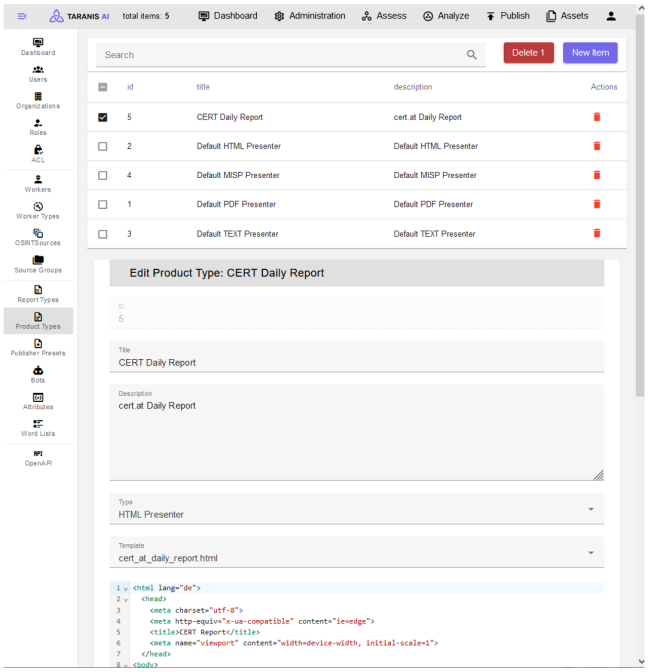

Figure 8: Product type "CERT Daily Report".

4.3 OSINT Analysis, Report Generation and Sharing Workflow: A Case Study

The capabilities of Taranis AI from the analyst's perspective are best understood through a practical demonstration. This case study focuses on the workflow of Taranis AI in a typical scenario:

- *Efficient Navigation via the Dashboard:* After logging into Taranis AI's user interface, the user is shown the Dashboard, which efficiently categorizes and displays the most critical tags by their respective categories. This design ensures that users are immediately presented with a prioritized view of cyber intelligence, allowing for quick identification of emerging threats and trends.
- *Seamless Transition to Detailed Analysis:* Upon selecting a specific tag, the user interface intuitively transitions to the 'Assess' view. Here, the user is presented with a collection of stories related to the chosen tag. This functionality exemplifies the system's ability to facilitate deeper dives into specific topics without overwhelming the user with information.
- *Enhanced Filtering and Sorting Capabilities:* To refine the analysis, Taranis AI offers advanced filtering options. Users can apply additional filters, such as time frames (e.g., filtering by week) and sorting parameters (e.g., sorting by relevance). This feature is crucial in sifting through vast amounts of data to extract the most pertinent information.
- *The 'Analyze' View – Integration of Stories into Reports:* An essential feature of Taranis AI is the ability to add selected stories to a report with a simple button click. This streamlined process aids analysts in compiling comprehensive reports, ensuring that key insights and data are efficiently captured and documented for decision-making purposes.
- *Report Dissemination:* Here, the effectiveness of the previous steps converges, allowing for the directed dissemination of detailed, insightful, and actionable intelligence reports. Dissemination takes place via pre-defined channels, such as MISP server, a Web server, email and the like. Important is that stories and dissemination targets can carry classification labels to avoid the unintentional publication of classified data according to the common Bell-LaPadula model.

With the start of its shift, the analyst logs into Taranis AI and is presented a dashboard (see Figure 9) that shows the most frequently hit keywords per category, e.g., the top organizations, persons, products, CVEs, vendors etc. of recently gathered OSINT.

Since the analyst's goal is to create a report about the most pressing new vulnerabilities (that many public platforms are reporting about), s/he sifts through the most interesting keywords, and thus clicks "SQL injection" – an important attack technique that was recognized through a custom word list. This causes Taranis AI to switch over to the Assess tab.

The Assess tab shows all news items related to (and tagged with) "SQL injection", clustered into stories. The analyst discovers a cluster of news item (i.e., a story) regarding the APT "Lazarus Group" (recognized through a custom word list – compare Figure 7), after clicking "open" s/he can inspect tags and related news items (Figure 10).

By having a glimpse at the tags, the analyst finds out that all the items seem to affect Spanish aerospace only, a quick cross check with the created summary of this story confirms that. So, it is not relevant for the use case at hand. The analyst therefore resets the filters to receive all collected items from the last 72 hours, which are also clustered into stories. Stories of larger size automatically have higher relevance. So, the analyst ranks items according to their

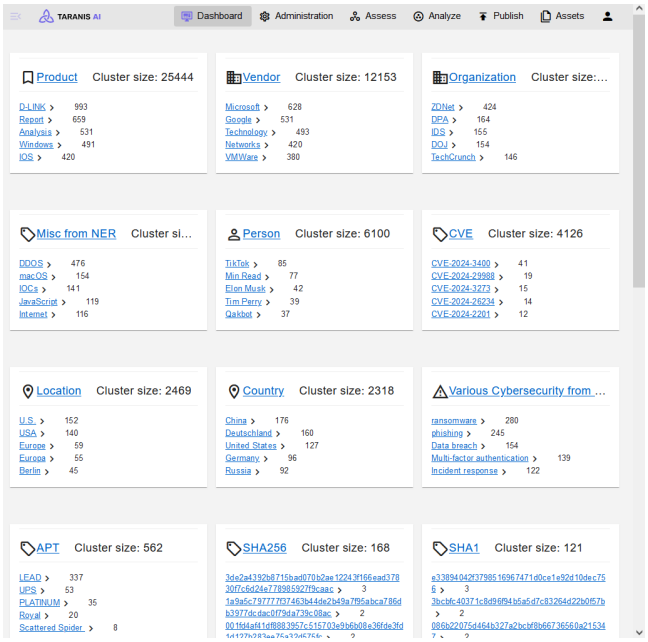

Figure 9: Dashboard, presented after Login.

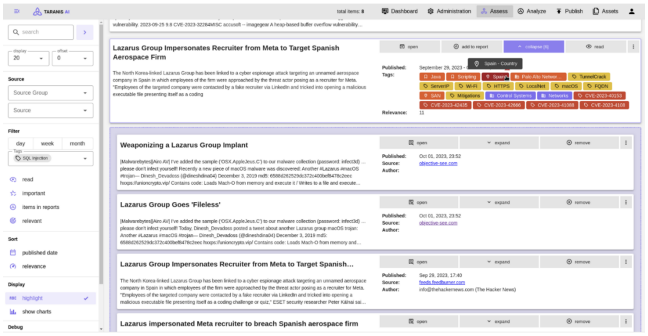

Figure 10: Lazarus Group being active in Spain.

relevance to get to know the most important stories (see Figure 11) and reviews whether they contain valuable information about new vulnerabilities.

Figure 11: A story of a new vulnerability in WS-FTP.

The analyst has now the chance to deeper dive into the story by investigating the single articles that constitute the story, following the links to the original articles, checking from which sources they came and when they have been published. Since the story seems relevant for the constituency, the analyst decided to add it to a report by clicking the "add to report" button. This provides the functionality to choose a predefined report (in our case "Vulnerability Report for military systems" – see Figure 12).

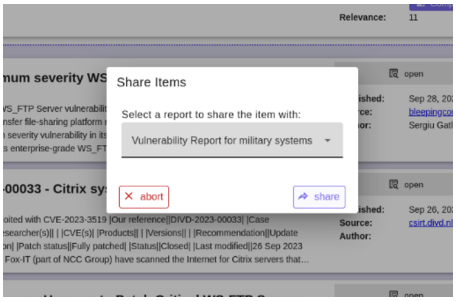

Figure 12: Sharing the story of a vulnerability in WS-FTP.

Switching to the Analyze tab, the analyst sees all reports, which is particularly convenient if multiple constituencies are served simultaneously. As an example, Figure 13 shows the created "Vulnerability Report for military systems" which comprises seven stories so far – the one that was just added, and six others added by other analysts. Now, further stories can be added the same way until an analyst decides to finalize the report.

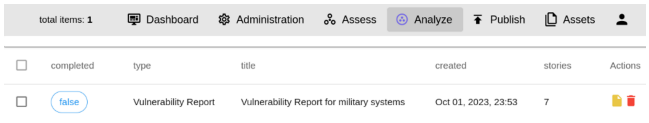

Figure 13: Reports being created and processed.

Next, the analyst switches to the Publish tab, where s/he picks a method to render a report, e.g., convert the collected stories to a pdf, html page, simple text or even a JSON object to push it to a MISP server (see Figure 14).

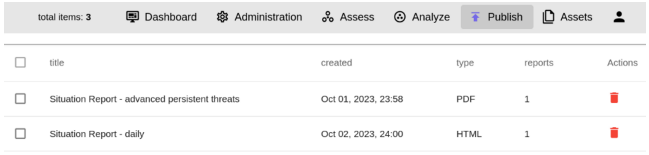

Figure 14: Render products in various formats.

The finally modeled product in Figure 15 shows a simple JSON object on the right hand-side created from the picked WS-FTP vulnerability story. This could also be rendered as simple text, pdf or html report; and pushed out using predefined channels, such as e-mail, Web server or MISP server. In the illustrated case, the JSON

object, representing carefully filtered and assessed information, is pushed to a MISP server and rendered there (see Figure 16) – waiting to be consumed by the final receiver.

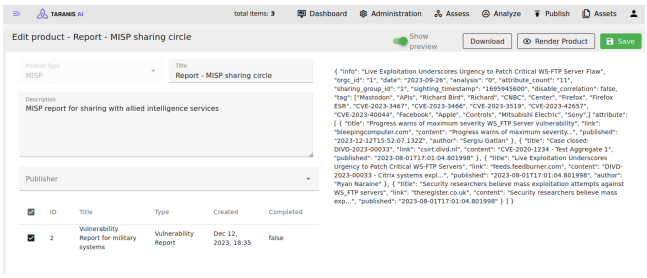

Figure 15: Resulting report to be pushed to MISP via API.

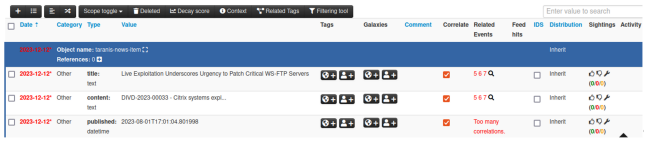

Figure 16: Finally rendered story in MISP.

5 DISCUSSION OF APPLICABILITY

We tested our approach together with our national CERT and NIS authorities to collect qualitative feedback from subject matter experts and prospective users.

5.1 NLP Challenges

The clustering of heterogeneous news items poses significant challenges that encompass various dimensions, including linguistic diversity, thematic breadth, and the need for source-specific handling. To address linguistic diversity, we employ a multi-lingual pre-trained model capable of clustering articles in different languages effectively. This approach not only streamlines the clustering process but also ensures that language barriers do not hinder the analysis of relevant information across diverse linguistic contexts.

Furthermore, the clustering of news items must contend with the inherent variability in topics, sources, and formats. While clustering algorithms can effectively group articles based on thematic similarities, the inclusion of unrelated sources or topics can dilute the quality of clusters. To mitigate this issue, our system allows for the exclusion of specific sources, particularly archive sources, which may not align with the current analysis objectives. This capability ensures that clustering results remain focused and relevant to the user's needs. On the other side, near duplicates, e.g., articles that have a very high similarity and contain verbatim copies of texts, are sorted out as well, but ranking the newer item higher.

Special news items, such as daily summaries, present additional challenges in clustering due to their unique characteristics and context. These items may require specialized handling to ensure accurate clustering results and meaningful analysis. Our system incorporates mechanisms to handle such special news items effectively, ensuring that they are either excluded from clustering, or appropriately split and clustered alongside other relevant content.

In terms of cluster quality, initial feedback from users has been positive, indicating that the system effectively meets their needs for information retrieval and analysis. To continuously improve cluster quality, we employ several strategies, including optimizing the trade-off between speed and accuracy in language model application (e.g., by just using a subset of the lines of text for clustering), calculating cluster quality metrics, and adapting NLP techniques for new source types. Additionally, we address specific challenges related to cyber security and OSINT, such as the distinction between cyber security-relevant items and those with general IT news, the balance between cluster aggressiveness and analyst expectations, and the prioritization of workflow speed while minimizing errors. These efforts contribute to the development of a robust and effective platform for news item clustering in cyber security and OSINT applications, addressing the diverse needs and challenges inherent in heterogeneous data environments.

5.2 Lessons Learned from Application

The need for high-quality gathering: The data gathering process is paramount in any analytical process. Several challenges arise in this endeavor, including the handling of daily digests, the presence of polluting text such as language menus and header/footer content, and the identification and management of updated articles and near duplicates. Addressing these challenges requires robust mechanisms for data preprocessing, de-duplication, and filtering to ensure that only high-quality, relevant information is included in subsequent analysis.

The need for flexible and adaptable workflows: Flexibility and adaptability are essential attributes of any workflow designed to support complex analytic tasks. In cyber security and OSINT, this need is magnified by the diverse requirements and preferences of analysts operating within different organizational contexts. Our approach emphasizes the development of flexible workflows that accommodate various analytical methodologies while preserving the core principles of the well-adopted Taranis workflow. This includes the integration of generic mechanisms for flagging articles, assigning different statuses, and facilitating collaboration across analysts, roles, and instances, thereby enabling seamless coordination and information sharing among team members

The need for manually corrected AI-produced results: Despite advancements in artificial intelligence and machine learning, the human element remains indispensable in the analytic process. Manual correction of AI-produced results not only ensures accuracy but also serves as valuable feedback to improve the performance of automated systems over time. Our system facilitates manual interventions such as removing or revising tags, suppressing irrelevant topics, splitting clusters, and providing editable summaries and annotations. By incorporating human expertise and feedback into the learning loop, we enhance the adaptability and effectiveness of AI-driven analysis while acknowledging the nuanced and context-dependent nature of information interpretation.

The need for carefully tuned recommender systems: The reliance on feedback and recommender systems introduces certain limitations that must be carefully considered. The opacity of learned models may lead to the inadvertent filtering out of relevant articles, highlighting the importance of transparency and explainability

in model development. Moreover, continuous monitoring is necessary to detect and mitigate biases and concept drifts that may affect the reliability of automated recommendations. Additionally, permanently dismissed stories may regain relevance in the future, underscoring the need for dynamic and adaptive systems capable of reevaluating past decisions in light of changing circumstances. Addressing these limitations requires a holistic approach that balances the benefits of automation with the inherent complexities of human decision-making in the analytical process.

5.3 Future Extensions for Advanced Cyber Situational Awareness

The conventional models of cyber situational awareness (CSA), as outlined in Pahi's et al. analysis [10], typically involve several sequential phases or levels aimed at establishing a solid foundation for decision-making processes. Following Endsley's well-known model [4], Taranis-AI enhances perception of relevant elements (first level) to facilitate comprehension of the current situation (second level). By incorporating additional features, Taranis AI extends into the realm of projection (third level), enabling informed decisions during operations by anticipating their near-future impact.

In particular, several extensions are planned: (i) Enabling self-asset management to align and match findings from CTI with own assets, thus prioritizing and ranking news items. (ii) Utilizing Large Language Models (LLMs) [18] not only for clustering and categorizing items but also for formulating situational awareness reports tailored to specific stakeholders in their preferred language. This customization is crucial, as conveying the same information to a technical expert versus a commanding officer require vastly different report styles. (iii) Incorporating mission-specific information, such as available or created mission models based on mission engineering, to provide targeted and easily understandable CTI that supports efficient decision-making in high-pressure operational scenarios. (iv) Implementing automatic notification systems for emerging threats, such as when a story on a specific topic experiences sudden growth. (v) Linking automatically filtered free-text CTI with structured CTI, such as attack techniques from MITRE Attack[26] or courses of action from MITRE Defend[27], to automate decision-making processes and advance projection, the highest level of CSA. This could involve presenting operators with countermeasures implemented by other organizations or allied nations when facing similar threats, as well as projecting future adversary tactics based on historical data.

6 CONCLUSION

This paper reported on the current developments of Taranis AI – a Taranis-NG fork with extended NLP and AI capabilities applicable for national authorities and military stakeholders.

Military operations and strategic planning benefit from the application of AI for intelligence analysis. Taranis AI automates the process of ingesting and analyzing CTI, allows tracking of threat trends, and sharing of mission-specific information in a secure way. Therefore, Taranis AI optimizes cyber defence processes, reduces the required personnel resources for cyber defence operations, and lowers the level of required human expertise for carrying out cyber defence activities.

[26] https://attack.mitre.org/
[27] https://d3fend.mitre.org/

ACKNOWLEDGMENTS

The work in this paper was mainly funded by the Connecting Europe Facility (CEF) program in course of the project AWAKE (2020-AT-IA-0254). The work in this paper further has received funding from the European Union - European Defence Fund under GA no. 101121418 (EUCINF) and 101121403 (NEWSROOM). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them. Taranis AI is an open source project under the EUPL license, and is available at https://taranis.ai/

REFERENCES

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*. 54–59.
[2] Isabelle Böhm and Samuel Lolagar. 2021. Open source intelligence: Introduction, legal, and ethical considerations. *International Cybersecurity Law Review* 2, 2 (2021), 317–337.
[3] Mehdi Elahi, Francesco Ricci, and Neil Rubens. 2016. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review* 20 (2016), 29–50.
[4] Mica R Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human factors* 37, 1 (1995), 32–64.
[5] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering* 34, 1 (2020), 50–70.
[6] Bang Liu, Fred X Han, Di Niu, Linglong Kong, Kunfeng Lai, and Yu Xu. 2020. Story forest: Extracting events and telling stories from breaking news. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14, 3 (2020), 1–28.
[7] Francesco Marchiori, Mauro Conti, and Nino Vincenzo Verde. 2023. Stixnet: A novel and modular solution for extracting all stix objects in cti reports. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*. 1–11.
[8] Wongyung Nam and Beakcheol Jang. 2023. A survey on multimodal bidirectional machine learning translation of image and natural language processing. *Expert Systems with Applications* (2023), 121168.
[9] Austrian Institute of Technology. 2024. Taranis AI. https://taranis.ai/.
[10] Timea Pahi, Maria Leitner, and Florian Skopik. 2017. Analysis and assessment of situational awareness models for national cyber security centers. In *International Conference on Information Systems Security and Privacy*, Vol. 2. 334–345.
[11] Shaina Raza and Chen Ding. 2022. News recommender system: a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review* (2022), 1–52.
[12] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*. 15–18.
[13] SK-CERT. 2022. Taranis-NG. https://github.com/SK-CERT/Taranis-NG.
[14] Jennifer Sleeman, Tim Finin, and Milton Halem. 2021. Understanding cybersecurity threat trends through dynamic topic modeling. *Frontiers in big Data* 4 (2021), 601529.
[15] Nan Sun, Ming Ding, Jiaojiao Jiang, Weikang Xu, Xiaoxing Mo, Yonghang Tai, and Jun Zhang. 2023. Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives. *IEEE Communications Surveys & Tutorials* (2023).
[16] Wiem Tounsi and Helmi Rais. 2018. A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Computers & security* 72 (2018), 212–233.
[17] Immanuel Trummer. 2023. From BERT to GPT-3 codex: harnessing the potential of very large language models for data management. *arXiv preprint arXiv:2306.09339* (2023).
[18] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).