Enhancing Cyber Situational Awareness with AI: A Novel Pipeline Approach for Threat Intelligence Analysis and Enrichment

Dzenan Hamzic[1](✉), Florian Skopik[1], Max Landauer[1], Markus Wurzenberger[1], and Andreas Rauber[2]

[1] AIT Austrian Institute of Technology, Vienna, Austria
{dzenan.hamzic,florian.skopik,max.landauer,markus.wurzenberger}@ait.ac.at
[2] Vienna University of Technology, Vienna, Austria
rauber@ifs.tuwien.ac.at

Abstract. Cyber Situational Awareness (CSA) is crucial for understanding and anticipating developments across diverse domains. This paper introduces a novel approach employing advanced Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques to effectively analyze and enrich Cyber Threat Intelligence (CTI) and Open Source Intelligence (OSINT) data. The paper designs an unified CTI and OSINT processing pipeline that integrates named entity recognition (NER), relationship extraction, classification, and summarization, addressing current limitations in CTI analysis. Notably, our evaluation of existing language models revealed significant shortcomings, with general-purpose tokenizers recognizing only 1.62% of specialized MITRE ATT&CK terms. In contrast, our pipeline achieves superior performance, notably surpassing state-of-the-art models in some important aspects. Practical military and civilian scenarios further demonstrate the pipeline's value in generating actionable intelligence, enabling complex reasoning by combining symbolic knowledge graphs and semantic vector search methods. Future developments focus on refining model scalability and enhancing analytical capabilities to increase the effectiveness, efficiency, and applicability of our approach.

Keywords: cyber situational awareness · cyber threat intelligence · nlp · defence industry · cti processing pipeline

1 Introduction

Cyber Situational awareness (CSA) refers to understanding what is happening in a given context and is essential across domains such as aviation, military operations, emergency response, healthcare, and power grid management [7]. More precisely, SA involves perceiving elements in the environment, understanding

B. Coppens et al. (Eds.): ARES 2025 Workshops, LNCS 15995, pp. 44–62, 2025.
https://doi.org/10.1007/978-3-032-00633-2_3

their meaning, and anticipating future developments [13]. It relies on integrating diverse and distributed data sources [6].

Modern systems generate vast amounts of data reflecting internal states and external conditions. With advances in Internet and datalink technologies, data can be gathered from virtually anywhere. The challenge today lies not in data availability but in accessing, filtering, and retrieving relevant information efficiently [6].

Cyber Situational Awareness (CSA) focuses on the digital domain, utilizing data from various cyber sensors, including intrusion detection system (IDS) alerts and Cyber Threat Intelligence (CTI). This information often originates from Open Source Intelligence (OSINT) and is subsequently processed by analytical tools or directly assessed by decision-makers [23].

OSINT tools have emerged to address the task of collecting publicly available information for intelligence purposes. Although many such tools exist, they are often limited in scope-either focusing on a single source or integrating data from multiple platforms in a way that hinders streamlined processing and analysis. This limitation underscores the need for advanced intelligence systems capable of rapidly providing relevant insights and supporting the generation of actionable intelligence [15].

To maintain accurate CSA, it is crucial to correlate, filter, prioritize, and interpret large volumes of complex data. Artificial intelligence (AI) and natural language processing (NLP) offer promising solutions to automate these tasks, thereby supporting analysts and decision-makers in both strategic and operational contexts [3,16,21,23].

To address the challenges of processing unstructured, noisy, and high-volume CTI and OSINT data that posseses inconsistent formats, lack structure, and limit the use of automation, this paper presents a comprehensive AI- and NLP-driven analysis pipeline that enables end users (militaries, security operations centers, national and cross-border cyber hubs, etc.) to perform advanced document analysis through prompt-based interaction. The proposed solution enables efficient aggregation, extraction, and structuring of intelligence to enhance cyber threat detection and response in both military and civilian contexts.

The key contributions of this work are:

- Present a unified AI/NLP pipeline for processing, preparing, and enriching CTI and OSINT data.
- Address current limitations in OSINT/CTI analysis.
- Discuss and describe how the pipeline components can be implemented.
- Propose structured approaches for interpreting unstructured intelligence reports to support analytical workflows.
- Identify areas related to OSINT analysis that require further research.
- Illustrate the proposed approach through fictional but realistic military and civilian cybersecurity scenarios.

The remainder of this paper structures as follows: Sect. 2 reviews related work in CTI processing. Section 3 presents our OSINT analysis and enrichment pipeline. Section 4 demonstrates a practical example of pipeline usage. Section 5 illustrates fictional military and civil scenarios. Section 6 presents the conclusion and outlines future work.

2 Related Work

Recent advancements in CTI have increasingly focused on automating data extraction and analysis from heterogeneous sources to address challenges posed by large volumes of unstructured data. Chang [4] proposed an automated pipeline specifically designed to predict tactics and techniques according to the MITRE ATT&CK framework. This approach emphasized scalable data collection and preprocessing techniques, significantly reducing manual effort. However, it mainly focused on general CTI extraction without explicitly considering comprehensive structured information extraction and advanced relationship identification necessary for complex scenarios faced by various organizations, including military, governmental, and commercial entities.

To enhance CTI analysis further, recent studies have begun integrating AI comprehensively throughout the entire CTI processing lifecycle. Alevizos and Dekker [2] introduced a structured AI-enhanced CTI pipeline, leveraging neural networks for intelligence ingestion, collaborative human-AI analysis, and automated generation of predictive mitigation strategies. While this pipeline provides substantial improvements in automation, speed, and accuracy, it does not explicitly integrate structured knowledge graphs and semantic retrieval mechanisms. Consequently, it lacks the capability for complex multi-hop reasoning, crucial for comprehensive situational awareness and strategic decision-making across diverse organizational contexts.

Addressing domain-specific requirements, Zhou et al. [25] developed the CTI View framework, explicitly targeting Advanced Persistent Threat (APT) intelligence extraction. By employing customized NLP models, particularly a hybrid BERT-BiLSTM-CRF architecture enhanced with GRU layers, CTI View demonstrated superior performance in identifying critical threat entities and Indicators of Compromise (IOCs) from heterogeneous textual reports. Despite these advancements, the approach predominantly addresses entity extraction and did not incorporate advanced relationship extraction or reasoning capabilities, which are pivotal for constructing comprehensive cyber situational awareness frameworks applicable to multiple sectors.

In a similar vein, Alam et al. [1] introduced LADDER, an innovative knowledge extraction framework designed to extract and categorize attack patterns from CTI reports systematically. LADDER effectively addresses the extraction of complex attack patterns, mapping them comprehensively to the MITRE ATT&CK framework. Although LADDER significantly advances pattern extraction, it does not extensively address semantic retrieval or the integration of structured symbolic reasoning via knowledge graphs for enhanced query capabilities.

Li et al. [11] proposed AttacKG, a method for automatically extracting structured attack behavior graphs from CTI reports and constructing Technique

Knowledge Graphs (TKGs). AttacKG aggregates detailed technique-level knowledge across multiple CTI reports, providing valuable insights into the dependencies and interactions among attack entities. However, AttacKG primarily focuses on constructing knowledge graphs at the technique level without deeply integrating semantic retrieval methodologies or leveraging transformer-based NLP models for advanced textual analysis.

Further advancing the extraction and classification of tactics, techniques, and procedures (TTPs), Rani et al. [17] introduced TTPXHunter. This tool significantly enhances cybersecurity threat intelligence by automatically extracting actionable TTPs from cyber threat reports. It employs advanced natural language models fine-tuned on domain-specific data, achieving superior accuracy compared to previous approaches. Nonetheless, while effective in TTP extraction, TTPXHunter does not inherently support complex relationship extraction or multi-hop query capabilities through integrated symbolic knowledge graphs.

Rani et al. [19] also proposed CAPTAIN, a novel APT attribution method leveraging TTP sequences to identify threat actors accurately. CAPTAIN introduces a sophisticated similarity measure for comparing sequences of TTPs to attribute cyber-attacks to specific APT groups. While highly effective for attribution tasks, CAPTAIN primarily focuses on TTP-based attribution without explicitly integrating semantic search or advanced relationship reasoning across broader CTI applications.

Despite significant progress, the reviewed approaches share common limitations: They often lack comprehensive integration of entity extraction, relationship modeling, and semantic reasoning in a unified pipeline. Specifically, most solutions either focus narrowly on entity recognition, TTP extraction, or attack attribution, without supporting the full transformation of unstructured CTI into structured, queryable knowledge. Moreover, they frequently omit multi-hop reasoning capabilities by not combining symbolic knowledge graphs with semantic retrieval mechanisms. In response to these shared gaps, our paper proposes a holistic pipeline that integrates advanced NLP, transformer-based models, and dual-retrieval techniques-combining knowledge graphs for structured reasoning with vector search for semantic enrichment. This enables detailed, explainable, and context-aware cyber situational awareness applicable across organizational contexts, from military to civilian domains.

3 Method for CTI Analysis and Enrichment

The proposed OSINT analysis and enrichment pipeline (see Fig. 1) consists of eight sequential steps designed to enhance CTI texts for efficient storage, retrieval, and advanced analysis. The order of the steps is carefully chosen, and each step produces additional information which supports the model in the next step to perform a certain task. For example, the first step extracts the named entities from a textual report. The second step takes the named entities creates linked triplets, which are structured representations in the form (Subject, Relation, Object) that capture relationships between entities. The content is enriched with additional context and supports the model applied in step 3 to better understand and classify the given text. Also step 4 uses the information produced in steps 1 and 2 to categorize the report. Another example would be the ordered sequence of step 5, which performs the text-to-TTP task using the TTPFShot [9] framework to tag the OSINT texts with TTPs from the Mitre ATT&CK framework. Next step 6 identifies APT groups within the text. By enriching the text with TTP IDs, the model applied in step 6 may perform better at the task of labeling APT groups given the TTP IDs from step 5. Step 7 summarizes OSINT texts and connects them with the triplets generated from step 2, which eventually serve as additional information for the clustering in step 8. All steps serve the purpose of enriching the content, which is ultimately stored in two databases: a Vector DB for semantic similarity-based text retrieval, and a Graph DB for structured relationship reasoning.

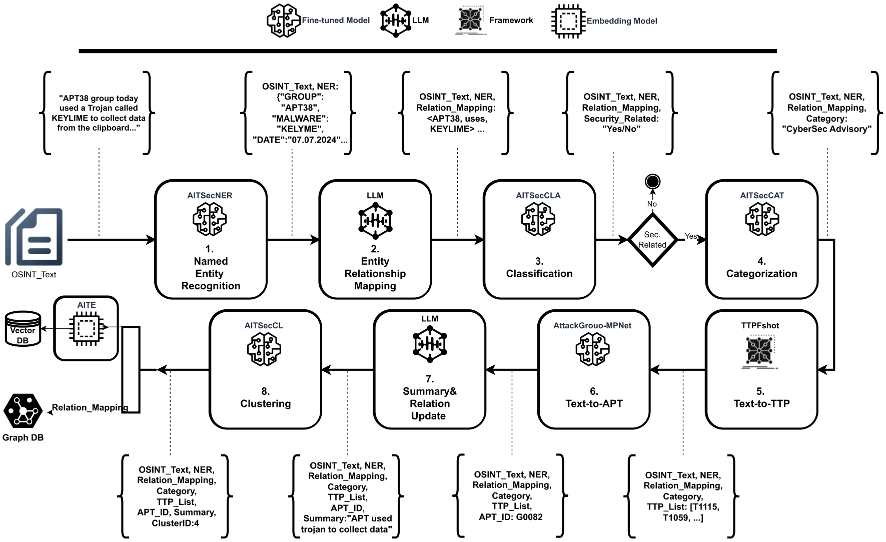

Fine-tuned Model
LLM
Framework
Embedding Model
"APT38 group today used a Trojan called KEYLIME to collect data from the clipboard..."
OSINT_Text, NER: {"GROUP": "APT38", "MALWARE": "KELYME", "DATE":"07.07.2024"...
OSINT_Text, NER, Relation_Mapping: <APT38, uses, KEYLIME> ...
OSINT_Text, NER, Relation_Mapping, Security_Related: "Yes/No"
OSINT_Text, NER, Relation_Mapping, Category: "CyberSec Advisory"
OSINT_Text
AITSecNER
1. Named Entity Recognition
LLM
2. Entity Relationship Mapping
AITSecCLA
3. Classification
No
Sec. Related
Yes
AITSecCAT
4. Categorization
AITE
Vector DB
AITSecCL
8. Clustering
LLM
7. Summary& Relation Update
AttackGrouo-MPNet
6. Text-to-APT
TTPFshot
5. Text-to-TTP
Relation_Mapping
Graph DB
OSINT_Text, NER, Relation_Mapping, Category, TTP_List, APT_ID, Summary, ClusterID:4
OSINT_Text, NER, Relation_Mapping, Category, TTP_List, APT_ID, Summary:"APT used trojan to collect data"
OSINT_Text, NER, Relation_Mapping, Category, TTP_List, APT_ID: G0082
OSINT_Text, NER, Relation_Mapping, Category, TTP_List: [T1115, T1059, ...]

Fig. 1. OSINT analysis and enrichment pipeline.

The textual reports that serve as input to the pipeline are initially gathered and provided by external systems, such as Taranis AI[1], which is an OSINT tool that utilizes AI for information gathering and situational analysis. These reports are in a raw, unstructured text format and typically include a report title, timestamp, and body text.

The remaining sections describe each step in detail. Furthermore, each of the sections provides examples to support comprehension. Additionally, candidate solutions for each step are listed.

[1] https://taranis.ai/.

3.1 Step 1: Named Entity Recognition

The first step performs Named Entity Recognition (NER) on the input text, extracting critical entities such as attackers, victims, sectors, dates, malware names, IP addresses, and other cybersecurity-relevant terms. This step is necessary because the extracted entities are used in subsequent stages for triplet construction, classification and categorization, context enrichment, summarization, clustering, and as chunk metadata in the vector database to enable more precise retrieval.

While several specialized NER solutions for the CTI domain exist, most rely heavily on BERT-based models. These models face two key challenges: (1) limited context window length [5] and (2) poor tokenization of specialized CTI terminology [14]. For example, the term `CVE-2021-44228` may be split into tokens like [“CVE”, “-”, “2021”, “-”, “442”, “28”] by a general-purpose tokenizer, losing its semantic meaning. Such poor tokenization of CTI-specific terms can harm tasks like classification, relation extraction, and embedding generation. Consequently, further research is needed to evaluate existing tools comprehensively and potentially develop tailored NER models with custom tokenizers.

One promising candidate is the AITSecNER[2] model, a GLiNER-based [24] model fine-tuned on the AnnoCTR [10] dataset. However, AITSecNER is limited to a maximum input window of 384 tokens, restricting its usability on large CTI reports. Additionally, the motivation for developing custom NER models and tokenizers arises from our observation (see Table 1) that most commonly used security-focused transformer models do not adequately cover the MITRE ATT&CK dictionary of Groups and Software.

Table 1. Tokenizer Coverage of Mitre ATT&CK Software Dictionary

Tokenizer	Recognized	Unrecognized	Coverage (%)	Missing (%)
SecurityBERT	11	667	1.62%	98.38%
SecurityBERT+	11	667	1.62%	98.38%
SecBERT	225	453	33.19%	66.81%
ATTACKBERT	50	628	7.37%	92.63%

Table 1 presents the tokenization coverage of several prominent security-focused language models when applied to the Mitre ATT&CK dictionary of 687 Software-name entries. The results indicate that most tokenizers fail to represent the majority of these specialized terms as single tokens. Specifically, SecurityBERT [8] and SecurityBERT+[3] cover only 1.62% of the terms, meaning 98.38% of the entries would be split into subwords or unknown tokens. ATTACKBERT[4]

[2] https://huggingface.co/selfconstruct3d/AITSecNER.
[3] https://huggingface.co/ehsanaghaei/SecureBERT_Plus.
[4] https://huggingface.co/basel/ATTACK-BERT.

performs slightly better at 7.37% coverage, while SecBERT [12] achieves the highest coverage of 33.19%. These findings motivate the need for developing domain-adapted tokenizers that can better preserve the semantic integrity of CTI-specific terminology.

One additional motivation for building a custom NER model stems from our evaluation of OpenAI's GPT-4o in a zero-shot setting on cybersecurity-specific entity extraction (Table 2). The evaluation is performed on the AnnoCTR dataset's [10] test split.

Table 2. LLM Performance per Entity Type

Entity Type	GT Entities	Predicted	Precision	Recall	F1-Score
TACTIC	21	2	0.000	0.000	0.000
CON	33	0	0.000	0.000	0.000
MALWARE	50	55	0.395	0.382	0.386
SECTOR	20	23	0.0567	0.0647	0.0589
TECHNIQUE	10	13	0.000	0.000	0.000
ORG	20	26	0.1333	0.1400	0.1350
DATE	45	48	0.380	0.380	0.380
LOC	20	20	0.0400	0.0233	0.0283

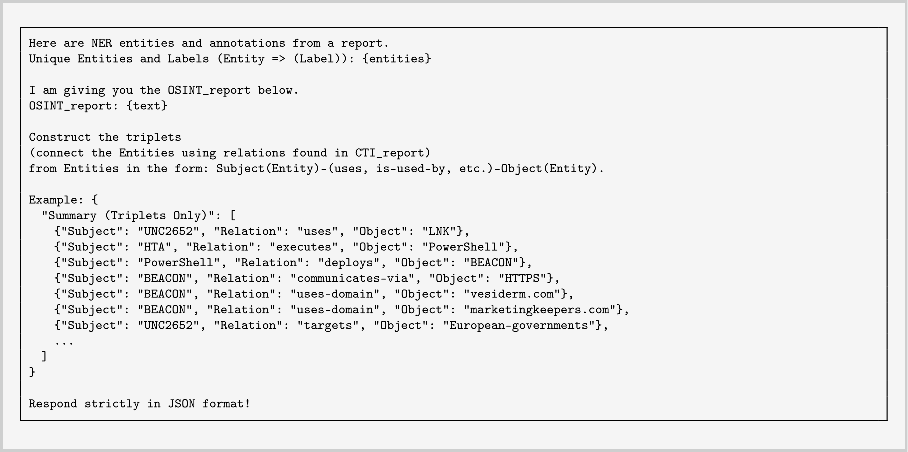
As Table 2 demonstrates, the zero-shot GPT-4o model struggles with extracting specialized CTI entities. Critical entity types such as TACTIC, TECHNIQUE, and CON (Cybersecurity Concepts) were either not identified or achieved zero F1-scores. Even common entity types like SECTOR, ORG, and LOC exhibit very low precision and recall, indicating frequent false positives or missed extractions.

Notably, the model performs slightly better for MALWARE and DATE entities, reaching F1-scores of 0.386 and 0.380 respectively. However, these results remain insufficient for operational CTI tasks where precise extraction of entities like TTPs, Groups, and Tools is crucial.

This evaluation confirms that general-purpose LLMs, even at the GPT-4o level, fail to capture the domain-specific semantics of cybersecurity texts. They lack the granularity needed to identify Mitre ATT&CK concepts, malware families, or organization names in threat reports. These findings reinforce the need for developing a dedicated CTI NER model with domain-adapted tokenizers and training on cybersecurity corpora to ensure reliable and accurate entity extraction for downstream analysis tasks.

3.2 Step 2: Entity Relationship Mapping

The second step proceeds with the output of step 1 and employs a LLM to construct relationship triplets in the format Subject-Relation-Object. For example, from the sentence *"APT29 used the SUNBURST malware to compromise SolarWinds."*, the LLM extracts the triplets: `(APT29, used, SUNBURST malware)` and `(SUNBURST malware, targeted, SolarWinds)`. These triplets capture key relationships between entities and form the basis for knowledge graph construction in the Graph DB. These triplets utilize NER-extracted entities and capture the relationships articulated within the OSINT text. This stage is crucial for clearly representing attacker actions and allows comparison with the Mitre ATT&CK-derived triplets to identify novel information. We validated our approach using a preliminary proof-of-concept. Figure 2 illustrates the prompt which we used in our premilary proof-of-concept for constructing triplets from entities and their relations provided in the original OSINT text.



```
Here are NER entities and annotations from a report.
Unique Entities and Labels (Entity => (Label)): {entities}

I am giving you the OSINT_report below.
OSINT_report: {text}

Construct the triplets
(connect the Entities using relations found in CTI_report)
from Entities in the form: Subject(Entity)-(uses, is-used-by, etc.)-Object(Entity).

Example: {
  "Summary (Triplets Only)": [
    {"Subject": "UNC2652", "Relation": "uses", "Object": "LNK"},
    {"Subject": "HTA", "Relation": "executes", "Object": "PowerShell"},
    {"Subject": "PowerShell", "Relation": "deploys", "Object": "BEACON"},
    {"Subject": "BEACON", "Relation": "communicates-via", "Object": "HTTPS"},
    {"Subject": "BEACON", "Relation": "uses-domain", "Object": "vesiderm.com"},
    {"Subject": "BEACON", "Relation": "uses-domain", "Object": "marketingkeepers.com"},
    {"Subject": "UNC2652", "Relation": "targets", "Object": "European-governments"},
    ...
  ]
}

Respond strictly in JSON format!
```

Fig. 2. Prompt template for relation extraction.

The extraction of relationship triplets represents a crucial step in transforming unstructured text into structured data. By leveraging a LLM to infer and generate Subject-Relation-Object triplets, this process effectively captures adversary tactics and operational linkages that might otherwise remain obscured in textual reports. The ability to align these extracted relationships with MITRE ATT&CK triplets allows for the detection of novel tactics and deviations from known attack patterns.

Our preliminary proof-of-concept demonstrates that this approach enhances situational awareness by structuring cybersecurity information in a machine-readable format. As Fig. 2 shows, our prompting strategy facilitates consistent and accurate extraction of triplets from CTI reports. However, further evaluation and revision is required to refine the prompt design, improve relation inference accuracy, and minimize hallucinated or spurious connections. Future work will focus on optimizing the triplet extraction process by incorporating domain-specific constraints and fine-tuning LLMs for improved precision in CTI tasks.

3.3 Step 3: Classification

The augmented report texts, enriched with extracted entities and their corresponding relationship triplets, proceed to the classification (Step 3). This stage serves as a critical filter to determine whether the processed report is relevant to cybersecurity and thus warrants further analysis within the pipeline. Given the heterogeneity of threat intelligence sources - ranging from generic news articles to detailed technical advisories - this filtering step is essential to avoid polluting downstream processes with irrelevant or off-topic content.

Currently, no comprehensive out-of-the-box solution exists that can reliably distinguish cybersecurity-relevant texts from non-relevant content in diverse OSINT datasets. To address this, our preliminary implementation leverages a zero-shot classification approach using the *bart-large-mnli* model[5]. The model is prompted with candidate categories (e.g., "cybersecurity report", "non-cybersecurity report") and infers the most probable class without requiring task-specific fine-tuning.

While this zero-shot approach enables rapid prototyping and preliminary filtering, our evaluation reveals (see Table 3) that general-purpose models such as *bart-large-mnli* struggle with the domain-specific nuances of CTI text. Reports containing ambiguous language, mixed cyber and non-cyber content, or implicit threat descriptions often result in misclassifications. This limitation highlights the need for a specialized classification model tailored to the CTI domain.

Table 3. Classificator Tokenizer Coverage of Mitre ATT&CK Software Dictionary

Tokenizer	Recognized	Unrecognized	Coverage (%)	Missing (%)
bart-large-mnli	10	668	1.47%	98.53%

To address this gap, we plan to develop AITSecCLA, a dedicated cybersecurity text classifier. AITSecCLA will be fine-tuned on a curated dataset of labeled CTI reports, ensuring better understanding of technical language, attack frameworks, malware families, and other domain-specific concepts. The goal is to significantly improve classification accuracy over general-purpose models, particularly in distinguishing subtle indicators of cyber threats embedded within complex narratives.

3.4 Step 4: Categorization

Step 4 extends the classification of step 3 and categorizes cybersecurity-relevant CTI texts to distinguish between different report types, with a particular focus

[5] https://huggingface.co/facebook/bart-large-mnli.

on identifying “Alert”-type reports. This step is crucial, as alerts represent high-priority, time-sensitive intelligence that often demands immediate attention from cybersecurity analysts and incident response teams.

To the best of our knowledge, there exist currently no out-of-the-box models available for this type of CTI-text categorization. Therefore, in our proof-of-concept (AITSecCAT), we consider using few-shot or zero-shot classification with the same model applied in the classification step, namely *bart-large-mnli*. Additionally, a large language model or a sentence transformer could be fine-tuned to perform this categorization task.

Accurately distinguishing alerts at this stage enables the pipeline to prioritize urgent intelligence for immediate action, while directing more detailed reports into deeper analysis paths-ultimately enhancing the pipeline’s responsiveness and operational relevance.

3.5 Step 5: Text-to-TTP

Extracting TTPs is a crucial capability in cyber threat intelligence processing, as it enables direct mapping of observed adversary behaviors to the widely adopted MITRE ATT&CK framework, facilitating structured reasoning and comparative analysis.

Several existing solutions aim to extract TTPs from text, including TRAM[6], TTPHunter [18], and the current state-of-the-art, TTPXHunter [20]. However, both TRAM and TTPHunter are limited to mapping text to only the 50 most common TTPs. TTPXHunter improves on this by supporting up to 193 TTPs. Despite this advancement, none of these models supports the extraction of TTP subtechniques, and all require either resource-intensive model fine-tuning or time-consuming dataset augmentation during the initial setup.

An approach in this pipeline step is TTPFShot [9], which is a retrieval-augmented few-shot learning model specifically designed to extract Mitre ATT&CK TTPs, including subtechniques, from CTI text segments. This retrieval-based few-shot approach significantly reduces the need for large annotated datasets while maintaining high extraction accuracy, making it suitable for dynamic and evolving CTI datasets.

Preliminary evaluations (see Fig. 3) show that TTPFShot outperforms baseline models, including TTPXHunter, in terms of precision, recall, and F1-score when evaluated on the TTPHunter dataset[7]. Its ability to generalize from limited examples while maintaining strong alignment with the MITRE ATT&CK framework positions it as a state-of-the-art solution for TTP extraction.

[6] https://github.com/center-for-threat-informed-defense/tram.

[7] https://github.com/nanda-rani/TTPHunter-Automated-Extraction-of-Actionable-Intelligence-as-TTPs-from-Narrative-Threat-Reports/blob/main/Dataset/TTPHunter_dataset.csv.

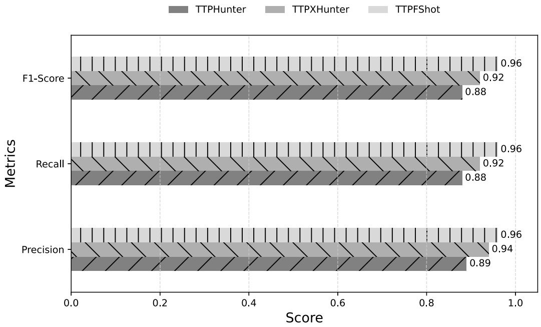


Fig. 3. TTPFShot performance comparison.

Table 4. Summary Evaluation Metrics

Metric	Our approach	BART CNN
ROUGE-1	0.0518	0.0283
ROUGE-2	0.0295	0.0278
ROUGE-L	0.0363	0.0283
METEOR Score	0.0166	0.0099

3.6 Step 6: Text-to-APT

Step 6 involves assigning Mitre ATT&CK APT Group IDs to text chunks. This step is essential for enriching chunk metadata, enabling retrieval based on APT Group IDs once the text chunks are stored in a database. Due to the lack of existing solutions for this specific task, we are actively developing a dedicated method. Our initial approach, the AttackGroup-MPNET model[8], is a variant of MPNet [22] fine-tuned on Mitre ATT&CK procedure descriptions, with an extended tokenizer tailored to the CTI domain. Alternatively, CAPTAIN [19] can be used at this stage as well, assuming TTP extraction has already been completed in Step 5.

3.7 Step 7: Summary and Relation Update

Using newly identified TTPs and APT IDs, Step 7 employs an LLM to update and expand the relation mapping with insights from earlier steps. Additionally, this step generates concise executive summaries derived directly from updated relations and identified entities. Based on initial experiments, this graph-enriched summarization offers superior quality compared to general-purpose summarization models.

To evaluate the quality of the generated summaries, we computed standard metrics such as ROUGE and METEOR. As shown in Table 4, our approach - which uses entity and relationship triplets within an LLM prompt guided by

[8] https://huggingface.co/selfconstruct3d/AttackGroup-MPNET.

the instruction "Generate summary using only the triplets"—outperforms the general-purpose summarization model BART CNN[9] across all evaluation metrics. These results demonstrate that leveraging structured knowledge in the form of entities and relationships may lead to more informative and contextually accurate summaries, especially for CTI reports.

3.8 Step 8: Clustering

Finally, Step 8 clusters the processed texts using a custom-trained embedding model, AITSecCL, which is currently being fine-tuned on cybersecurity-specific dictionaries and threat intelligence corpora[10]. This model builds on MPNet [22], adapted to the CTI domain. Once finalized, it will be used for clustering, assigning each text segment a cluster ID that identifies reports covering similar topics or related threat activities.

Clustering serves two primary purposes: (1) grouping intelligence related to the same incident or campaign, and (2) detecting emerging events or anomalies that deviate from existing patterns. This mechanism enhances situational awareness by enabling rapid correlation of new reports with known threat campaigns and by surfacing novel or previously unseen threats. The assigned cluster ID is carried forward and stored alongside each text chunk in the database, supporting efficient retrieval and downstream analysis.

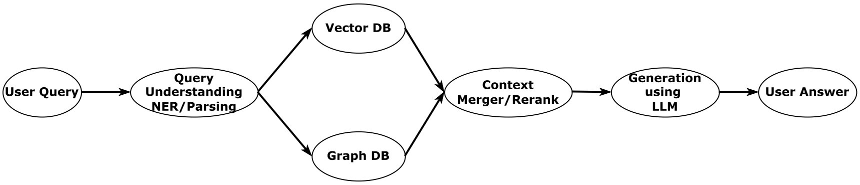


Fig. 4. Pipeline application example.

After the 8 steps of the pipeline, the data persistence occurs across two specialized databases. Text segments are chunked and embedded using our custom embedding model from the clustering step (see Sect. 3.8), optimized for cybersecurity-specific terminology, enriched by metadata from previous steps, and stored in a vector database. Concurrently, updated relation mappings are stored in a graph database. Both databases also retain metadata such as timestamps and source details, ensuring comprehensive data contextualization and traceability.

[9] https://huggingface.co/facebook/bart-large-cnn.

[10] https://huggingface.co/selfconstruct3d/mpnet-classification-finetuned-cyber-groups.

4 Pipline Application Example

The designed pipeline enables advanced information retrieval and reasoning by combining symbolic knowledge graphs and semantic vector searches. This dual retrieval mechanism significantly enhances the system's capability to answer complex, multi-hop,i.e., queries that require reasoning across multiple pieces of information-often spanning different documents or entities-and context rich queries that typical retrieval pipelines struggle to address due to their reliance on single-pass, shallow (semantic similarity only) retrieval and limited reasoning over disconnected facts.

As illustrated in Fig. 4, the process begins with query understanding, where entities, time frames, tactics, and targeted infrastructure are extracted using our custom NER model. The system then simultaneously queries:

- **Graph Database:** For structured relationship reasoning between actors, techniques, and assets based on pre-processed triplets.
- **Vector Database:** For semantically similar text segments enriched with metadata such as TTPs, infrastructure, clusters, and campaign context.

Both streams of information are merged and optionally reranked, balancing symbolic accuracy with semantic relevance. The merged context is passed to the LLM, which generates a detailed, context-aware answer.

4.1 Example Prompt and Detailed Pipeline Execution

To demonstrate the pipeline's practical utility, we present an example query (prompt) and generated answers, highlighting the system's ability to merge graph-based reasoning with semantically retrieved context for complex CTI tasks.

Example Prompt 1: "Which APT groups have used clipboard scraping techniques in the past year?"

Pipeline Execution:

- **NER and Parsing:** The system detects entities like "APT groups" and techniques such as "clipboard scraping".
- **Graph Query:** Traverses known relationships to find APT nodes linked to TTP nodes related to clipboard scraping (e.g., MITRE ATT&CK T1115).
- **Vector Search:** Retrieves semantically relevant report chunks mentioning clipboard scraping, APT operations, and similar cases.
- **Merging and Reranking:** Prioritizes graph hits with strong TTP-APT links and combines them with top vector matches.
- **LLM Answer Generation:** *"Based on available intelligence, APT38 and APT29 have both deployed clipboard scraping techniques over the past year. APT38 leveraged the KEYLIME trojan, while APT29 used CHOPSTICK malware, targeting clipboard memory to collect sensitive data."*

This example demonstrates how the hybrid pipeline:

- Handles multi-hop queries involving both cyber and physical components.
- Leverages symbolic knowledge graphs for explainable reasoning paths.
- Enriches factual retrieval with semantic context from vector-based similarity search.
- Produces LLM-generated answers grounded in structured and unstructured data sources.

Such capabilities are critical for advanced cyber threat intelligence scenarios, supporting analyst workflows in attribution, impact assessment, and proactive defense planning.

5 Military and Civil Use Cases

To demonstrate the practical potential of our proposed pipeline, we present two fictional but realistic use cases-one from the military domain and one from the civil sector where our hybrid retrieval pipeline (Fig. 4) supports situational awareness and decision-making. Both scenarios are illustrative mockups based on representative technologies and imaginable threat situations, showcasing how the system supports cybersecurity operations across different contexts.

5.1 Red and Blue Background

In this fictional scenario, the conflict involves two opposing sides: Red and Blue. Red is a hybrid threat actor combining a paramilitary group, the Heroic Brigades (HB), and an APT group, Teasing Mosquito (TM). Their goal is to destabilize Blue politically, economically, and socially, using influence operations, disinformation, and cyber-physical attacks—especially against critical energy infrastructure. Blue is a state-led coalition defending against Red's multi-domain hybrid threats. Blue aims to ensure societal stability, protect infrastructure, and counter disinformation and election interference. The confrontation spans several months, with Red's actions aligned to MITRE ATT&CK tactics and techniques. Red also collaborates with neutral hacker groups to purchase system access.

5.2 Scenario 1: Surveillance System Compromise

The framed content outlines how the pipeline components could work together to support situational awareness and threat analysis. A framed layout is used to clearly separate this illustrative example from the rest of the text. All data and actors are hypothetical.

Description: Red actors exploit leaked WiFi credentials for surveillance cameras to monitor Blue's mission activities.

Stage 1 – User Query:

```
Which APT groups have compromised our surveillance systems using WiFi credential leaks?
      Provide past examples and mitigation advice.
```

Stage 2 – Query Understanding (NER/Parsing):

– **Extracted Entities:** Surveillance Systems, APT Groups, WiFi Credential Leaks
– **Focused TTP:** Initial Access (MITRE ATT&CK TA0001)

Stage 3 – Graph Query:

```
MATCH (camera)-[r:COMPROMISED_BY]->(attacker:Actor)
RETURN camera, attacker
```

Example Result:

```
(Surveillance_Camera_A)-[COMPROMISED_BY]->(Teasing Mosquito APT)
```

Stage 4 – Vector DB Retrieval (Chunks + Metadata):
Similarity search: `"Surveillance WiFi compromise initial access"`
Retrieved Chunk:

```
"In 2023, attackers exploited leaked WiFi credentials to access surveillance feeds from
      cameras placed in forward operating bases."
```

Stage 5 – Context Merger and Reranking:

– **Graph Result:** Surveillance_Camera_A compromised by Teasing Mosquito APT
– **Vector Context:** Similar APT tactics observed in 2023
– Reranked for confirmed APT involvement

Stage 6 – LLM Generation (Merged Prompt):

```
Given that Surveillance_Camera_A was compromised by Teasing Mosquito APT, and historic
     reports show attackers using leaked WiFi credentials to access surveillance feeds
     in 2023, summarize the threat, intent, and suggest mitigation.
```

Stage 7 – LLM Answer (Final Assessment):

– **Threat Summary:** Teasing Mosquito APT compromised Surveillance_Camera_A via WiFi leaks.
– **Intent:** Gather intelligence on troop movements and operational patterns.
– **Mitigation Recommendations:**

- Rotate WiFi credentials
- Disable compromised devices
- Deploy traffic monitoring
- Review remote access controls

5.3 Civil Scenario: Vulnerability Prioritization for Patch Management

In this mockup scenario, we walk through how a Security Operations Center (SOC) analyst might use our system to prioritize patching efforts after a newly reported CVE. The framed content shows step-by-step how the pipeline supports decision-making, from entity recognition to final recommendation generation.

Description: A SOC analyst needs to prioritize patching efforts after a new vulnerability (CVE) affecting their web infrastructure is reported.

Stage 1 – User Query:

```
Which active threat actors are currently exploiting CVE-2024-12345, and what is the
     impact? Should we prioritize patching our web servers?
```

Stage 2 – Query Understanding (NER/Parsing):

– **Extracted Entities:** CVE-2024-12345, Web Servers, Threat Actors
– **Focused TTP:** Exploitation for Initial Access (MITRE ATT&CK T1203)

Stage 3 – Graph Query:

```
MATCH (cve:CVE {id:'CVE-2024-12345'})<-[:EXPLOITS]-(group:APT)
RETURN group
```

Example Result:

```
(Scarlet Leopard APT exploits CVE-2024-12345)
```

Stage 4 – Vector DB Retrieval (Chunks + Metadata):
Similarity search: `"Exploitation of CVE-2024-12345 by threat actors"`
Retrieved Chunk:

```
"In March 2025, Scarlet Leopard APT exploited CVE-2024-12345 targeting Apache servers,
     leading to remote code execution in several financial organizations."
```

Stage 5 – Context Merging and Reranking:

– **Graph Result**: APT group confirmed exploiting CVE-2024-12345
– **Vector Context**: Active exploitation reported in the financial sector

Stage 6 – LLM Answer Generation (Merged Prompt):

```
Summarize the risk associated with CVE-2024-12345 based on APT activity and
     exploitation reports. Provide patching recommendations.
```

Stage 7 – LLM Answer (Final Assessment):

– **Threat Summary**: Scarlet Leopard APT is actively exploiting CVE-2024-12345 to achieve remote code execution on Apache web servers.
– **Impact**: Successful exploitation may lead to full system compromise and data exfiltration. Financial organizations have been targeted.
– **Mitigation Recommendations**:

- Apply the latest security patch for CVE-2024-12345 immediately.
- Audit Apache server configurations.
- Monitor for signs of post-exploitation activities linked to Scarlet Leopard APT.

6 Conclusion

In this paper, we presented a comprehensive CSA framework specifically tailored for military and civilian cybersecurity applications, leveraging advanced AI and NLP techniques. Our unified pipeline significantly enhanced the analysis of CTI by improving named entity extraction, relationship mapping, identification of TTPs, attribution of APT groups, summarization quality, and clustering of threat reports.

The proposed framework integrates symbolic knowledge graphs with semantic retrieval methods to provide detailed, explainable, and contextually rich responses to complex cybersecurity queries. Through illustrative military and civilian scenarios, we demonstrated the practical utility and operational value of the pipeline, highlighting its ability to support real-time and strategic decision-making processes.

Pipeline improvements include the continuous refinement of entity extraction techniques, development of specialized cybersecurity classifiers for enhanced content relevance, and advancements in summarization methodologies through structured relationship extraction. Additionally, we emphasized the potential for improved clustering methods, enabling the identification of emerging threats and anomalies more effectively.

Future work will concentrate on enhancing the pipeline's scalability and robustness, refining AI model performance through extensive training on specialized datasets, and integrating additional cybersecurity data sources to further strengthen the situational awareness capabilities of our framework.

Acknowledgments. Funded by the European Union under the European Defence Fund (GA no. 101121403 - NEWSROOM and GA no. 101121418 - EUCINF). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. This work is co-funded by the Austrian FFG Kiras project ASOC (GA no. FO999905301).

References

1. Alam, M.T., Bhusal, D., Park, Y., Rastogi, N.: Looking beyond IoCs: automatically extracting attack patterns from external CTI. In: Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '23). ACM (2023). https://doi.org/10.1145/3607199.3607208
2. Alevizos, L., Dekker, M.: Towards an AI-enhanced cyber threat intelligence processing pipeline. Electronics **13**(11), 2021 (2024). https://doi.org/10.3390/electronics13112021
3. Arazzi, M., et al.: NLP-Based Techniques for Cyber Threat Intelligence. arXiv preprint arXiv:2311.08807 (2023). https://arxiv.org/abs/2311.08807
4. Chang, C.H.: Cyber Threat Intelligence: A Pipeline to Classify Cyber Threats from Disparate Data Sources. Honours thesis, AiLECS Lab, Monash University (March 2022)
5. Ding, M., Zhou, C., Yang, H., Tang, J.: Cogltx: applying bert to long texts. Adv. Neural. Inf. Process. Syst. **33**, 12792–12804 (2020)
6. Endsley, M.R.: Theoretical underpinnings of situation awareness: a critical review. In: Endsley, M.R., Garland, D.J. (eds.) Situation Awareness Analysis and Measurement. Lawrence Erlbaum Associates, Mahwah, NJ (2000). https://www.researchgate.net/publication/292771806_Situation_awareness_analysis_and_measurement_chapter_theoretical_underpinnings_of_situation_awareness, accessed: 2025-04-07
7. Endsley, M.R.: Situation awareness. In: Salvendy, G., Karwowski, W. (eds.) Handbook of Human Factors and Ergonomics, chap. 17. Wiley (2021). https://doi.org/10.1002/9781119636113.ch17, https://doi.org/10.1002/9781119636113.ch17
8. Ferrag, M.A., Ndhlovu, M., Tihanyi, N., Cordeiro, L.C., Debbah, M., Lestable, T., Thandi, N.S.: Revolutionizing cyber threat detection with large language models: a privacy-preserving bert-based lightweight model for iot/iiot devices. IEEe Access **12**, 23733–23750 (2024)

9. Hamzic, D., Skopik, F., Landauer, M., Wurzenberger, M., Rauber, A.: Ttp classification with minimal labeled data: A retrieval-based few-shot learning approach (2025), to appear at the 20th International Conference on Availability, Reliability and Security (ARES 2025), August 11-14, 2025, Ghent, Belgium. Springer (2025)
10. Lange, L., Müller, M., Torbati, G.H., Milchevski, D., Grau, P., Pujari, S., Friedrich, A.: Annoctr: a dataset for detecting and linking entities, tactics, and techniques in cyber threat reports (2024). https://arxiv.org/abs/2404.07765
11. Li, Z., Zeng, J., Chen, Y., Liang, Z.: AttacKG: constructing technique knowledge graph from cyber threat intelligence reports. arXiv preprint arXiv:2111.07093 (2022). https://arxiv.org/abs/2111.07093
12. Liberato, M.: Secbert : analyzing reports using bert-like models, December 2022. http://essay.utwente.nl/93906/
13. Munir, A., Aved, A., Blasch, E.: Situational awareness: Techniques, challenges, and prospects. AI **3**(1), 55–77 (2022). https://doi.org/10.3390/ai3010005
14. Nayak, A., Timmapathini, H., Ponnalagu, K., Venkoparao, V.G.: Domain adaptation challenges of bert in tokenization and sub-word representations of out-of-vocabulary words. In: Proceedings of the first workshop on insights from negative results in NLP, pp. 1–5 (2020)
15. Pieterse, H., Van't Wout, C., Khan, Z., Serfontein, C.: Specialised media monitoring tool to observe situational awareness. In: Proceedings of the 17th International Conference on Information Warfare and Security, p. 244 (2022)
16. Rahman, M.R., Mahdavi-Hezaveh, R., Williams, L.: What are the attackers doing now? Automating cyber threat intelligence extraction from text on pace with the changing threat landscape: A survey. arXiv preprint arXiv:2109.06808 (2021), https://arxiv.org/abs/2109.06808
17. Rani, N., Saha, B., Maurya, V., Shukla, S.K.: Ttpxhunter: Actionable threat intelligence extraction as ttps from finished cyber threat reports. arXiv (2024). https://arxiv.org/abs/2403.03267
18. Rani, N., Saha, B., Maurya, V., Shukla, S.K.: Ttphunter: Automated extraction of actionable intelligence as ttps from narrative threat reports. In: Proceedings of the 2023 Australasian Computer Science Week, pp. 126–134 (2023)
19. Rani, N., Saha, B., Maurya, V., Shukla, S.K.: Chasing the Shadows: TTPs in Action to Attribute Advanced Persistent Threats. arXiv preprint arXiv:2409.16400 (2024). https://arxiv.org/abs/2409.16400
20. Rani, N., Saha, B., Maurya, V., Shukla, S.K.: Ttpxhunter: Actionable threat intelligence extraction as ttps from finished cyber threat reports. Digital Threats: Resand Practice **5**(4), 1–19 (2024)
21. Samtani, S., Li, W., Benjamin, V., Chen, H.: Informing cyber threat intelligence through dark web situational awareness: The azsecure hacker assets portal. Digital Threats **2**(4) (Oct 2021). https://doi.org/10.1145/3450972. https://doi.org/10.1145/3450972
22. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mpnet: Masked and permuted pre-training for language understanding. Adv. Neural. Inf. Process. Syst. **33**, 16857–16867 (2020)
23. Wurzenberger, M., et al.: NEWSROOM: Towards automating cyber situational awareness processes and tools for cyber defence. In: Proceedings of the 19th International Conference on Availability, Reliability and Security (ARES 2024). Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664476.3670914, https://dl.acm.org/doi/10.1145/3664476.3670914

24. Zaratiana, U., Tomeh, N., Holat, P., Charnois, T.: Gliner: Generalist model for named entity recognition using bidirectional transformer (2023). https://arxiv.org/abs/2311.08526
25. Zhou, Y., Tang, Y., Yi, M., Xi, C., Lu, H.: CTI view: APT threat intelligence analysis system. Secur. Commun. Networks **2022**, 1–15 (2022). https://doi.org/10.1155/2022/9875199