MITREtrieval: Retrieving MITRE Techniques From Unstructured Threat Reports by Fusion of Deep Learning and Ontology

Yi-Ting Huang, R. Vaitheeshwari, Meng-Chang Chen, Ying-Dar Lin, *Fellow, IEEE*, Ren-Hung Hwang, *Senior Member, IEEE*, Po-Ching Lin, *Member, IEEE*, Yuan-Cheng Lai, Eric Hsiao-Kuang Wu, *Member, IEEE*, Chung-Hsuan Chen, Zi-Jie Liao, and Chung-Kuan Chen

***Abstract*—Cyber Threat Intelligence (CTI) plays a crucial role in understanding and preemptively defending against emerging threats. Typically disseminated through unstructured reports, CTI encompasses detailed insights into threat actors, their actions, and attack patterns. The MITRE ATT&CK framework offers a comprehensive catalog of adversary tactics, techniques, and procedures (TTPs), serving as a valuable resource for deciphering attacker behavior and enhancing defensive measures. Addressing the challenge of time-consuming manual analysis of MITRE TTPs in unstructured CTI reports, this paper presents MITREtrieval, a novel system that leverages deep learning and ontology to efficiently extract MITRE techniques. This approach mitigates issues related to the implicit nature of TTPs, textual semantic dependencies, and the scarcity of adequately labeled datasets, enabling more effective analysis even with limited sample sizes. Our approach combines a sophisticated sentence-level BERT deep learning model with ontology knowledge to address sparse data challenges, using a voting algorithm to merge outcomes. This results in a more accurate classification of MITRE techniques, capturing contextual nuances effectively. Our evaluation confirms MITREtrieval's effectiveness in identifying techniques, regardless of their representation in training samples. MITREtrieval has surpassed benchmarks, achieving F2 scores of 58%, 62%, and 69% in multi-label technique identification across 113, 46, and 23 CTI reports, respectively, thereby streamlining CTI analysis and improving threat intelligence.**

***Index Terms*—Cybersecurity, deep learning, MITRE ATT&CK, natural language processing, ontology, threat intelligence.**

Manuscript received 21 February 2024; revised 26 April 2024; accepted 9 May 2024. Date of publication 27 May 2024; date of current version 21 August 2024. This research was supported by the Ministry of Science and Technology, Taiwan, under grant number 112-2622-E-008-005. The associate editor coordinating the review of this article and approving it for publication was K. Xue. *(Corresponding author: Eric Hsiao-Kuang Wu.)*

Yi-Ting Huang is with the Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: ythuang@mail.ntust.edu.tw).

R. Vaitheeshwari, Eric Hsiao-Kuang Wu, Chung-Hsuan Chen, and Zi-Jie Liao are with the Department of Computer Science and Information Engineering, National Central University, Taoyuan 320, Taiwan (e-mail: vaitheeshwarir@gmail.com; hsiao@csie.ncu.edu.tw; chunghsuanchen7@gmail.com; a486572148@gmail.com).

Meng-Chang Chen is with the Institute of Information Science, Academia Sinica, Taipei 115, Taiwan (e-mail: mcc@iis.sinica.edu.tw).

Ying-Dar Lin is with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: ydlin@cs.nctu.edu.tw).

Ren-Hung Hwang is with the College of Artificial Intelligence, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: rhhwangg@cs.ccu.edu.tw).

Po-Ching Lin is with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 62102, Taiwan (e-mail: pclin@cs.ccu.edu.tw).

Yuan-Cheng Lai is with the Department of Information Management, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: laiyc@cs.ntust.edu.tw).

Chung-Kuan Chen is with the Research Division, CyCraft Technology, New Taipei City 220, Taiwan (e-mail: ck.chen@cycarrier.com).

Digital Object Identifier 10.1109/TNSM.2024.3401200

I. Introduction

THE RISING number of cyber-attack events has amplified the significance of sharing Cyber Threat Intelligence (CTI) as a strategic countermeasure against prevalent cyber threats, especially formidable adversaries like Advanced Persistent Threats (APT) [1]. CTI is a large-scale data-driven process that systematically analyzes observations, including malware, suspicious events, and other rapidly evolving cybersecurity data. To facilitate CTI usage, many studies [2], [3], [4], [5], [6] focused on collecting, analyzing, and extracting evidence, such as indicators of compromise (IoC) in CTI reports. CTI about emerging threats and threat actors enables the effective development of cybersecurity solutions [7].

CTIs have been divided into two primary categories: structured and unstructured. Both forms of cybersecurity information have played crucial roles in enabling proactive defense mechanisms within CTI capabilities. Structured CTIs, such as STIX [8], TAXII [9], or YARA [10], were the focus of numerous studies [11], [12] that concentrated on sharing structured threat intelligence. Concurrently, another line of research has delved into the mechanisms for extracting threat intelligence from unstructured threat reports [2], [3], [5], [6]. Prominent security organizations and companies, including Microsoft, FireEye, and Symantec, consistently publish CTI reports to share their discoveries and insights on various threat actors and malware.

The primary objective of analyzing CTI reports is to extract threat intelligence about adversaries. This intelligence encompasses information detailing threat actor motives, targets, and attack behaviors, which typically include specifics like IP addresses, domains, and hashes. However, adversaries often modify their IPs or domains to evade detection [13]. Consequently, the focus has increasingly shifted

TABLE I
TTP OF SAMPLE EVENT

Event	In 2019, the USA launched cyber-attacks on Iran
Tactic	To bring down Iran's missile and rocket launch systems
Technique	**Phishing**
Procedure	Identify Iran's weapon launch systems, develop the malware, and infect computers via phishing

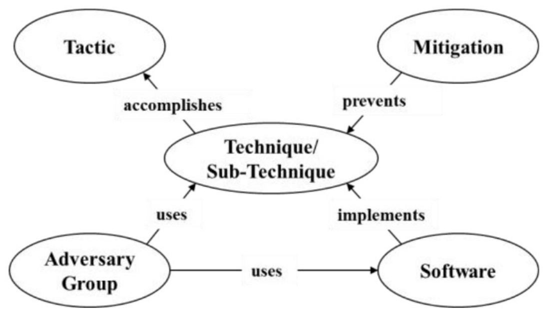

Fig. 1. Object model relationships of MITRE [16].

towards understanding the tactics, techniques, and procedures (TTP) [14], which are essential components of threat intelligence.

These TTPs, although crucial, are often not explicitly stated but rather are implicitly embedded within the CTI reports. 'Tactics' refer to the high-level descriptions of objectives that threat actors aim to accomplish. 'Techniques' provide more detailed insights into the specific behaviors that facilitate the achievement of these tactics. 'Procedures' delve deeper, outlining how threat actors employ various techniques to fulfill their objectives. An illustrative example of TTP related to a specific cybersecurity event can be found in Table I.

Thus, TTPs can fully represent an attack event or threat actor in a clear, readable manner which helps security analysts characterize incoming threats quickly. MITRE ATT&CK, a knowledge base born from real-world observations, has standardized these TTPs and gained significant traction in the security sector [17]. The object model relationship of MITRE ATT&CK is shown in Fig. 1. While several cybersecurity firms now tag their CTI reports based on MITRE ATT&CK to summarize the primary threat actions of adversaries, a substantial number of these reports remain untagged.

However, the manual annotation of CTI reports with MITRE techniques by cybersecurity experts is a daunting task due to the sheer volume and length of these documents. Recent works have put forth great effort into analyzing TTPs in CTI reports [2], [19], [20], [21], [22], [23], [24], [25], [26]. Although pioneering works like TTPDrill [2] and rcATT [19] have employed term frequency to identify techniques within CTI reports, these methods are plagued with shortcomings.

First, a high number of false negatives implies that many techniques in CTI reports are missed because term frequency cannot distinguish synonyms and locate correct information within a lengthy report. For example, synonyms such as "phishing" and "email" represent the same techniques of T1566 and will not be recognized as semantically similar by the term-frequency method. And if the word "email" is only mentioned once in a report with thousands of words, it would be ignored.

Within the cybersecurity industry, minimizing false negatives is often prioritized over reducing false positives [27], [28]. In practical terms, this means that for cybersecurity companies, the risk of missing crucial threat intelligence is considered more detrimental than the challenge of handling an excess of information. Therefore, it becomes critical to focus on semantic dependency to accurately discern the most relevant information for classifying techniques.

In studies addressing TTPs on a smaller scale, many researchers have limited their focus to only 5 or 10 types of techniques [22], [23], [24], [25], [26], predominantly due to the scarcity of adequately labeled data. Such a narrow scope inevitably leads to a restricted range of technique classifications provided by tools or systems, constraining the breadth of intelligence available to security firms. This limitation hampers their ability to gain a comprehensive understanding of adversary strategies and behaviors. Consequently, it is vital to integrate external knowledge sources, particularly for techniques that are underrepresented or have limited sample data, to ensure a more thorough and effective classification process.

In response to the complex challenges prevalent in cybersecurity intelligence, we have developed 'MITREtrieval' - a system specifically designed to extract MITRE ATT&CK techniques from CTI reports. At the core of MITREtrieval is a topic classifier, a tool we introduced to effectively sift through and eliminate irrelevant content from CTI reports, thereby isolating sentences that are explicitly technique-centric. This ensures that our model's focus is directed solely towards pertinent and actionable data.

To capture the intricate semantic relationships within these texts, we implemented the 'Sentence BERT with MITRE Description Embedding' (SBMDE) model. This advanced sentence-level BERT model is adept at bridging semantically related words and sentences, thus enabling a deeper understanding of contextual nuances.

Further enhancing MITREtrieval's capabilities is the integration of COMAT (Cybersecurity Ontology based on MITRE ATT&CK) [29]. The insights derived from COMAT are utilized within a deep neural network framework, combined with a sophisticated level-voting algorithm. This amalgamation is particularly effective in classifying techniques, even in scenarios characterized by limited sample data. The interaction of these components positions MITREtrieval as a standout system in the realm of large-scale technique classification, effectively addressing the nuanced demands of cybersecurity intelligence extraction.

We have gathered 724 CTI reports from the security community, spanning from 2018 to 2022, and labeled with MITRE ATT&CK techniques, to showcase the effectiveness of MITREtrieval on contemporary CTI reports and its superiority over current state-of-the-art methods. This carefully curated dataset, which we plan to make publicly available, will serve as a valuable resource for future security research. MITREtrieval is designed to identify relevant attack techniques from CTI reports, thus minimizing the reliance on manual analysis. However, it is worth noting that certain preliminary or subsequent stages of the process might involve manual intervention or review. The main contributions are as follows:

- We have successfully merged the detailed, structured information from ontology (COMAT) with the advanced analytical capabilities of a deep neural network (SBMDE). This integration tackles the challenge of multi-label document classification in cybersecurity, where the depth of ontology knowledge enhances the contextual understanding of the deep learning model. A key challenge was ensuring that the ontology's structured data seamlessly informs the deep learning process without imposing rigid constraints. Our solution including a level-voting algorithm that balances the deep learning predictions with the ontology insights, significantly improves the performance of Technique classification at a different range of given sample thresholds compared to all existing systems, especially at Recall and F2 scores.
- We have released a labeled dataset of 700 CTI reports, annotated with 14 tactics and 165 techniques. This dataset significantly enriches the resources available to the security community and supports further cybersecurity research.
- We have developed a system that autonomously detects MITRE ATT&CK techniques in CTI reports, significantly streamlining cybersecurity analysis. It accurately identifies techniques across three thresholds (6, 40, and 80), corresponding to 113, 46, and 23 TTPs respectively. This method reduces the manual effort required to analyze CTI reports, with our fastest predictions taking less than a minute for shorter reports and between 3-10 minutes for longer ones.

The paper is organized as follows: Section II offers background on CTI reports, MITRE ATT&CK, ontology, and related works. Section III details the inputs, outputs, and objectives of our study. Section IV explores our system architecture and its components. Section V assesses MITREtrieval's performance, and Section VI concludes with a summary and future research directions. For the convenience of the reader, an Index of Abbreviations is included in the Appendix.

II. Background

A. CTI Reports

CTI reports share knowledge about threat actors, especially APT, describing their motives, targets, and attack behaviors. A CTI report provides a variable source for understanding attacks [30]. As a result, many enterprises or cybersecurity organizations publish their own CTI reports to share information about threat actors with the community. For example, companies such as Microsoft and Google and cybersecurity companies like FireEye and Trend Micro will release CTI reports based on their findings about specific threat actors.

CTI strives to reduce asymmetrical knowledge between attackers and defenders through the gathering and analysis of defensive counterintelligence on cyber threats [31]. The content of CTI reports can be roughly divided into four parts: (1) Overview, (2) Attack description, (3) Conclusion, and (4) IoC. Background knowledge such as target, industry, name of a threat actor, and history are discussed in the overview. The attack description describes the threat action that the threat actor takes. The conclusion gives a threat actor summary by the author of the CTI report. The IoC is always listed at the bottom of the report.

B. MITRE ATT&CK

MITRE has developed various ATT&CK matrices tailored for specific threat scenarios, outlining threat actor objectives and their methods. These matrices are categorized into three distinct scenarios: enterprise, mobile, and industrial control systems (ICS). This paper concentrates on the enterprise matrix, as enterprise-targeted attacks are the most prevalent and the insights gained are particularly beneficial to cybersecurity firms. The enterprise matrix by MITRE ATT&CK is structured with tactics and techniques: tactics form the rows, while the corresponding techniques for each tactic are listed in the columns. In its 10th version, the enterprise matrix encompasses 14 tactics and 188 techniques.

C. Related Work

Recently, the extraction of TTPs from unstructured documents like blogs and CTI reports has garnered increasing attention in the research community. Three primary approaches have been adopted for this task: (1) utilizing Natural Language Processing (NLP) techniques, (2) applying Ontology-based methods, and (3) deploying Machine Learning models. Within this section, we delve into detailed discussions of works undertaken in these three categories.

1) Natural Language Processing: The field of NLP has been instrumental in enhancing the analysis of CTI reports, with notable methodologies developed to extract key information. Niakanlahiji et al. [24] introduced SECCMiner, an innovative system that utilizes NLP techniques such as stemming and parts of speech (POS) tagging. SECCMiner is adept at identifying noun phrases within CTI reports and aligning them with pre-established noun phrases that correspond to APT techniques. For instance, a sentence like "The malware downloads files from a remote command-and-control server" might not explicitly reference a specific technique like T1005 Ingress Tool Transfer, but SECCMiner's sophisticated NLP approach, enriched with domain-specific knowledge, can effectively deduce such implicit references.

Chen et al. [32] employ a methodology that includes POS tagging but focuses on isolating verbs as potential indicators of threat actions. Their process involves filtering out words unrelated to security by matching them against a curated list of security-relevant keywords. The core of their technique lies in aligning these extracted potential threat actions with known TTP threat actions. While this approach effectively identifies key actions, it is primarily based on word matching, which may not fully capture the richer contextual nuances present in CTI reports.

2) Ontology: Ontology-based approaches in cybersecurity have been pivotal in sharing knowledge about malware and threat actors, yet their association with techniques often remains underexplored. Traditional cybersecurity ontologies, such as the Unified Cybersecurity Ontology (UCO) [33] and Malware Ontology (MALOnt) [34], have primarily concentrated on amalgamating diverse, unstructured data to facilitate malware information sharing. While these ontologies have effectively structured knowledge around malware and threat actors, they tend to overlook the intricacies of evolving threat behaviors. To bridge this gap, Husari et al. [2] introduced TTPDrill, a unique threat-action ontology dedicated to mapping CTI reports to MITRE ATT&CK techniques. TTPDrill involves extracting threat actions, using a dependency parser from both the CTI reports and predefined procedure examples of each technique.

The Best Matching 25 (BM25) function [35], a ranking function used in information retrieval, is then employed to calculate the cosine similarity between these verb-object pairs in reports and techniques. If the similarity exceeds a certain threshold, the corresponding technique is considered present in the report. While TTPDrill showcases the potential of leveraging the rich knowledge within MITRE ATT&CK, its reliance on term-frequency-based features can lead to ambiguities. For example, techniques like T1132 (Data Encoding) and T1027 (Obfuscated Files or Information) might share the same verb-object tuple (verb="encode", object="data") but have distinct goals, potentially resulting in inaccurate mappings. Additionally, dependency parsers may extract general terms like "file" or "use," which are inadequate for precise technique classification.

Li et al. [36] advanced this field with AttacKG, achieving an F1-score of 0.789 using graph template matching. AttacKG constructs entity-dependency graphs for TTPs based on MITRE descriptions and aligns them with similar graphs from threat reports using graph alignment algorithms. However, AttacKG's focus on verbs (threat actions) over adjectives (properties of IOCs) limits its capability to fully represent techniques, especially those characterized by attributes like 'obfuscated malware'.

3) Machine Learning: In the realm of CTI report analysis, machine learning models have been predominantly utilized for technique classification. Notable works include rcATT by Valentine Legoy et al. [19], which employs TF-IDF for feature extraction followed by SVM for model training. Similarly, Ayoade et al. [20] proposed a method akin to rcATT but with an added focus on classifying reports into tactics before techniques using a confidence propagation strategy. Li et al. [23] and Peng et al. [22] both introduced approaches centering on topic extraction from CTI reports, albeit with slight variations in their methodologies. Ghazi et al. [21] ventured into Named Entity Recognition (NER) to extract a wide range of entities, including TTPs, from CTI report sentences, relying on a manually labeled training dataset.

MITRE's contribution, TRAM [25], utilizes a similar TF-IDF technique coupled with a machine-learning model trained on manually labeled sentences. However, these machine learning-based methods are not without their limitations, particularly in processing extensive reports where crucial but infrequently mentioned words might be overlooked. Furthermore, the inability of some models to effectively learn technique-specific features poses another challenge.

CTI reports are often laden with a plethora of words not directly tied to specific techniques, such as generic terms like "use," "target," as well as ambiguous words like "file," "list," or "C2" that may imply multiple techniques. This lexical complexity poses challenges for machine learning models in accurately identifying techniques. Moreover, the scarcity of comprehensive open-source labeled datasets further hampers the efficacy of machine learning models, limiting their scope to a narrow range of techniques. This limitation is evident in some existing works [22], [23], [24], [26], which are constrained to classifying only a small subset of techniques, thereby reducing their utility for cybersecurity experts.

Similarly, Rani et al. [37] introduced TTPHunter, a model utilizing BERT-based linear classifiers to align sentence contexts with relevant TTPs. While effective in extracting threat intelligence and profiling threats, TTPHunter's capabilities are restricted to 50 TTP classes, primarily due to limited data samples. This model also encounters difficulties in document-based analysis, which may not fully capture the complexity of TTP occurrences within CTI reports.

Our approach stands in contrast to these methods in three significant ways. The first one involves the Topic Classification process, where we employ a binary sentence classifier. This classifier is specifically designed to discern sentences in CTI reports that are relevant to MITRE ATT&CK techniques, ensuring a focused analysis. Secondly, we have developed COMAT, an inference-based ontology derived from the MITRE ATT&CK framework. This ontology is strategically queried using group, software, and threat action data extracted from CTI reports, allowing for a more detailed technique inference process. Lastly, we have developed SBMDE, a sentence-based BERT model enriched with MITRE description embeddings. This model captures the nuanced semantics and dependencies within CTI reports, addressing the limitations often seen in term-frequency or topic-modeling methods. The rich, context-aware embeddings generated by SBMDE form the basis for training our deep neural network, ensuring a nuanced understanding of the CTI report. Finally, in the Knowledge Fusion stage, we incorporate insights from the COMAT ontology into our deep learning model. This critical integration enhances our model's ability to accurately identify and understand TTPs within CTI reports.

Together, these innovations form the backbone of MITREtrieval, a system that transcends traditional content classification by combining deep neural network analysis with the rich ontological insights of COMAT. The important notations used in the MITREtrieval system are mentioned in Table II. For instance, C represents a CTI report, the primary subject of our study, while W_i and S_i refer to the individual words and sentences within these reports, crucial for granular analysis. Whereas, WE_i and SE_i denote the embeddings for sentences and words, respectively, highlighting our approach to understanding the deeper semantic and contextual layers of the CTI reports. Moreover, the table incorporates the MITRE ATT&CK framework into our analysis through notations like $tech_i$ for individual techniques and Des_i for their descriptions. The table also introduces terms related to our deep learning approach, such as DL_{ans} for predictions and various levels of technique matches (L_1, L_2, L_3, $must$, $must_{not}$) in ontology queries. Understanding these notations is crucial for following

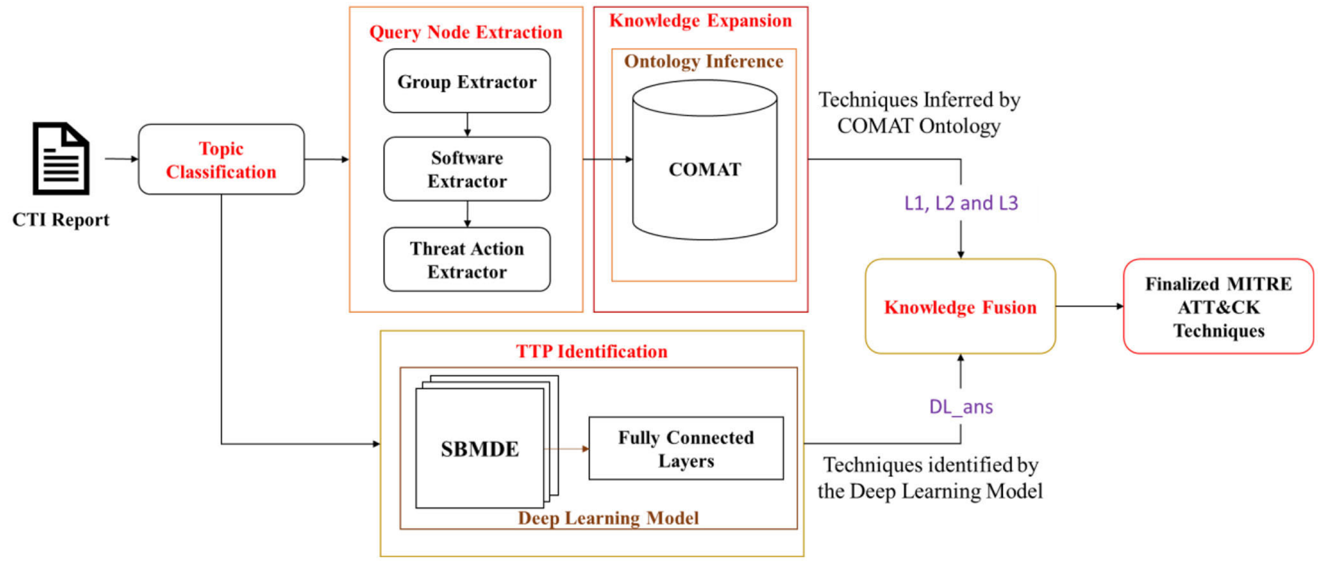

Fig. 2. MITREtreival system architecture.

TABLE II
NOTATIONS AND THEIR DESCRIPTIONS

Notation	Description
C	A CTI report
W_i	The i-th word in CTI report
S_i	The i-th sentence in CTI report
SE_i	The i-th sentence embedding in CTI report
WE_i	The i-th word embedding in CTI report
$Soft$	All software mentioned in CTI report
$Group$	Threat actor group name mentioned in CTI report
$\{V_i, O_i\}$	The i-th verb in CTI report
$tech_i$	The i-th MITRE technique in CTI report
Des_i	The i-th MITRE technique description
DL_{ans}	Prediction of deep learning method
Φ	Infer path set (Ontology)
L_1	Techniques of first level matches in ontology query
L_2	Techniques of second level matches in ontology query
L_3	Techniques of third level matches in ontology query
$must$	Techniques of keyword matches in ontology query
$must_{not}$	Techniques of keyword mismatches in ontology query
$output$	Final output after knowledge fusion

the methodology and findings as they form the backbone of our approach to dissecting and interpreting CTI reports.

III. Problem Formulation

A CTI report C that contains n number of words and sentences represents the input of MITREtrieval. The target output will be $output = \{tech_1,\ tech_2, \ldots.\ tech_m\}$, where m is the total number of MITRE techniques. The MITRE ATT&CK techniques retrieval problem can be modeled as a function that maps input CTI report C to binary vector *output*. The objective of MITREtrieval is maximum F_2 to minimize false negatives.

IV. MITREtreival

A. Overview of Workflow

MITREtrieval is designed to intricately align CTI reports with the nuanced techniques detailed in the MITRE ATT&CK framework. The architecture of MITREtrieval, depicted in Fig. 2, encompasses several integral components.

1) *Topic Classification:* The initial stage involves training a topic classifier to sift through sentences in a CTI report, identifying those that specifically relate to cybersecurity techniques.
2) *Query Node Extraction:* After determining the relevance of a sentence to a particular technique, we extract key 'query nodes'. These nodes include:
   1) Group (denoted as *Group*)
   2) Software (denoted as *Soft*)
   3) Threat actions (represented as { V, O})
3) *Knowledge Expansion:* The identified query nodes are crucial for correlating sentences with corresponding techniques in our COMAT ontology, thus expanding our knowledge repository. To demonstrate this process, we present actual CTI report examples in Fig. 3, showcasing the efficacy of both topic classification and knowledge expansion.
4) *TTP Identification:* At this stage, each sentence from the CTI report is processed through SBMDE model which transforms them into rich, context-aware embeddings. These embeddings serve as the core for training our deep neural network, aligning closely with the specific nuances and context of the CTI report. The SBMDE model is integral to this phase, as it not only enhances the sentence embeddings with detailed semantic understanding but also integrates MITRE-specific

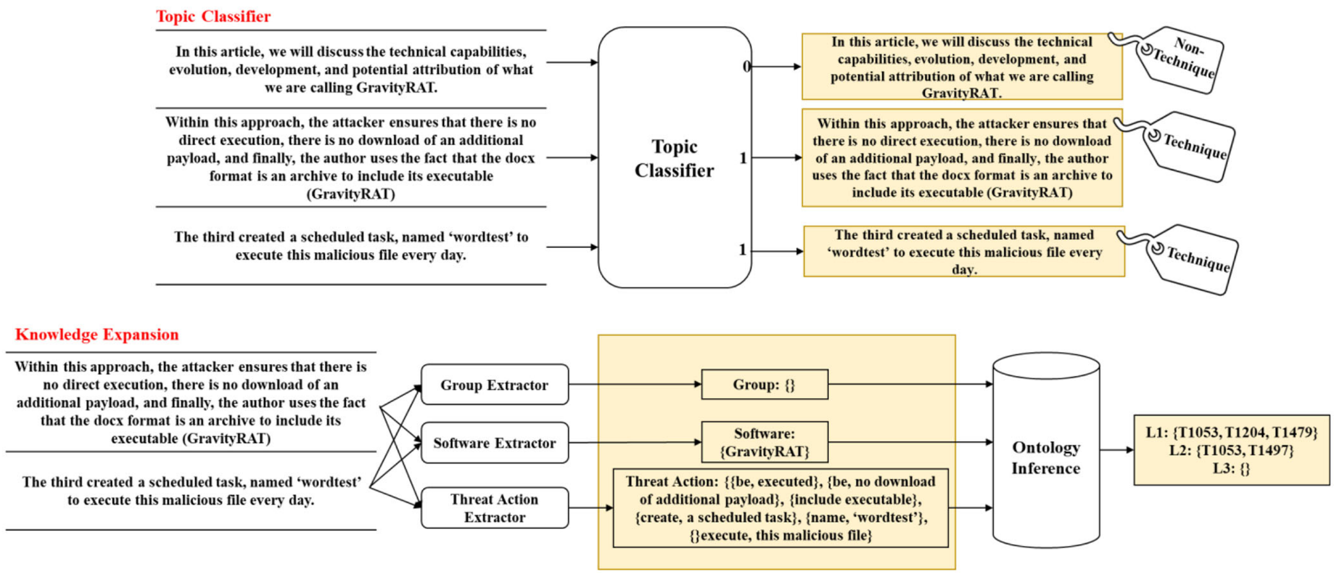

Fig. 3. The workflow of the topic classifier and knowledge expansion. The yellow boxes represent the output of each component.

insights, crucial for precise and comprehensive TTP identification.

5) *Knowledge Fusion:* In the final phase, MITREtrieval integrates COMAT's ontological insights with the SBMDE model's deep learning analysis, utilizing a level-voting algorithm. This fusion combines structured ontological knowledge and context-rich deep learning predictions, enhancing TTP identification accuracy in CTI reports.

B. Topic Classification

The content of CTI reports often includes many sentences without attack descriptions which do not help classify TTPs (e.g., background information about the cybersecurity company). These non-relevant sentences, if not filtered out, could be mistakenly treated as features during the TTP classification process, leading to inaccuracies.

To address this, we developed a binary sentence classifier by fine-tuning the BERT [38] model to discern relevant from irrelevant content. The architecture of this topic classifier is depicted in Fig. 4. The fine-tuned BERT model, as well as the dataset used for its training, is publicly available on GitHub [39] to facilitate research reproducibility.

For the topic classifier, we sourced sentences from the procedure examples listed in the MITRE ATT&CK framework as pertinent data points. These examples typically describe the methods employed by threat actors or malicious software when executing techniques. Conversely, we gathered non-relevant data samples from mainstream news sources such as CNN or BBC, which often detail the background narratives of threat actors without specific attack methodologies. The corpus for fine-tuning comprised 9305 relevant and 9132 irrelevant sentences.

Utilizing this data, BERT was fine-tuned to classify sentences based on their relevance to attack behaviors. The output CLS vector from BERT was then directed through

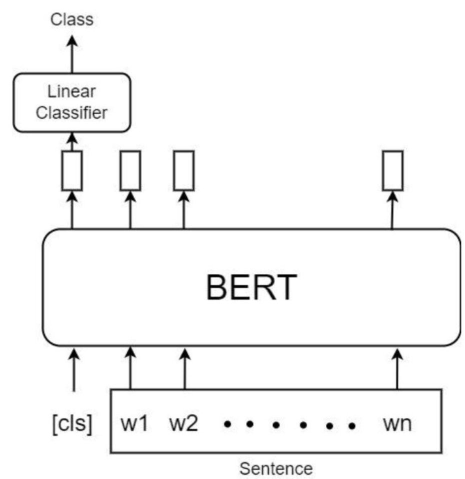

Fig. 4. Topic classifier model architecture.

a fully connected layer to finalize the classification. The implementation of this topic classifier ensures that subsequent reports are purged of non-essential content, thereby focusing solely on sentences that elucidate MITRE ATT&CK techniques.

C. Query Node Extraction

The process of knowledge expansion in our system begins with the extraction of query nodes from CTI reports using NLP techniques. These nodes, which represent key features within the reports, are subsequently utilized to query our COMAT ontology. This allows for the identification of corresponding techniques and enriches the raw data for specific analytical tasks.

CTI reports provide detailed insights into threat groups, software, and their attack behaviors, covering key elements such as group identity, software used, and specific actions. 'Group' refers to the name of the threat actor discussed in the report, while 'software' denotes the specific tools or malware employed by the actor. Threat actions are articulated through a combination of verbs and objects, indicating the specific behaviors exhibited by the threat group or software.

To methodically extract these features, we construct dictionaries for both group names and software as defined within the MITRE ATT&CK framework. These dictionaries play a pivotal role in identifying and matching the names of groups and software mentioned in CTI reports.

For the extraction of threat actions, which are characterized by verb-object pairs, we employ Extractor [40], a Semantic Role Labeler (SRL)-based tool adept at discerning entity relationships within sentences. This tool is particularly adept at identifying the verb-object pairings that signify the threat actions in the text.

In selecting the Extractor tool for relation extraction in our study, we carefully considered its advantages over newer techniques, such as those by Hu et al. [41], Bastos et al. [42], and Zhang et al. [43]. These recent developments in open relation extraction are known for their adaptability across different domains and minimal fine-tuning efforts, as highlighted in the work of Belfadel et al. [44]. However, our selection of the Extractor was based on its specific capability to accurately identify verb-object pairs, crucial for defining threat actions in the cybersecurity context. While the newer techniques offer broader adaptability, the precision and relevance of the Extractor in parsing cybersecurity-specific entities and relationships made it a more suitable choice for our study.

Furthermore, the practicality of implementing the Extractor, given its lower computational demands and focused functionality, solidified its selection for our research objectives. To summarize, our approach utilizes the group name (*Group*), software (*Soft*) names, and several threat actions ($\{V_i,\ O_i\}$) as the query nodes in the given report.

D. Knowledge Expansion - Ontology Inference

COMAT is an inferable ontology we developed for querying information from the MITRE ATT&CK framework. It advances cybersecurity ontology by prioritizing the dynamic aspects of threat behaviors over traditional focuses on malware and threat actors. Unlike [33] and [34], which are instrumental in organizing malware information, COMAT enriches this knowledge base by delving into adversarial actions with an SRL approach for threat action extraction. Moreover, our design of six types of meaningful inference paths, considering the types of CTI topics, sets COMAT apart in its ability to infer technical actions efficiently and accurately.

This methodological advancement, coupled with the integration of knowledge graph techniques from [45] and inference capabilities noted in [2], positions COMAT as a significant enhancement in the realm of cybersecurity knowledge mining and technique identification.

COMAT's ontology structure comprises six primary classes: Tactic, Technique, Group, Software, Mitigation, and the newly introduced Threat Action. These classes are interconnected through four types of relations, forming a comprehensive representation of the cybersecurity domain as mentioned in

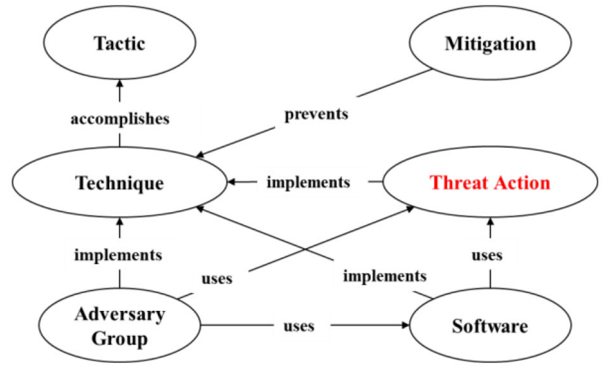

Fig. 5. COMAT ontology classes and relations.

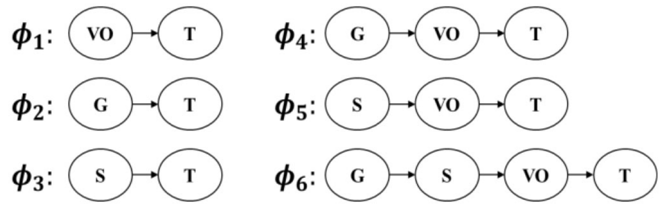

Fig. 6. Inference Path.

Fig. 5. COMAT's ontology is enriched by extending the object-model relationships within MITRE ATT&CK to include the class of Threat Action, thereby offering a novel perspective in representing such actions within the ontology. This enhancement aligns with prior studies asserting the composition of a threat action as a verb-object pairing.

To effectively leverage COMAT's capabilities, we crafted six inference paths to navigate through the ontology and deduce the relevant techniques, as illustrated in Fig. 6.

The $L1$ paths ($\phi_1,\ \phi_2,\ \phi_3$) provide a general search by querying direct links to groups, software, and threat actions, minimizing misprediction risks. In contrast $L2$ paths ($\phi_4,\ \phi_5$) and $L3$ path (ϕ_6) align with the more common themes in CTI reports, focusing on groups and software, respectively.

The ontology inference process begins with leveraging Roberta [47] to generate embeddings for the inputs (query nodes) *Group*, *Soft*, and $\{V_i,\ O_i\}$ along with existing COMAT instances. Subsequently, we employ cosine similarity, a measure of the angular closeness between two vectors, to assess the degree of similarity between the input embeddings and the ontology's instance embeddings. When the similarity measure for an input surpasses a carefully calibrated threshold of 0.75, it signifies a strong correspondence, leading to the input being associated with the most similar instance in the ontology. The threshold value was determined through iterative testing, balancing the performance metrics in mapping inputs to ontology instances.

When inputs surpass the set threshold, they are mapped to corresponding instances in COMAT, transforming raw data into structured entities within the ontology. These 'mapped nodes' encapsulate the input as part of COMAT's interconnected knowledge network. For technique identification, we systematically navigate these nodes using predefined

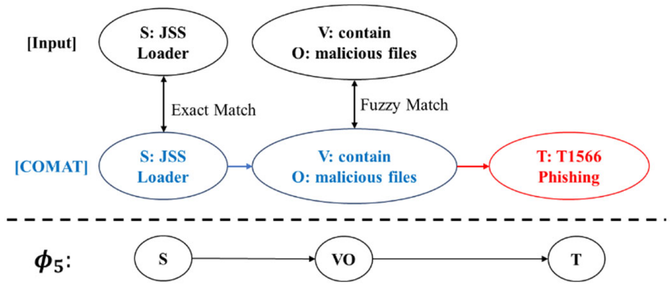

Fig. 7. An example of query mapping module and technique inference by COMAT.

inference paths, reflecting the ontology's complex relationships and interactions for accurate technique determination.

To illustrate this process, consider Fig. 7, which demonstrates the application of inference path ϕ_5. When presented with a query like {JSS Loader, contain, malicious file}, 'JSS Loader' is identified as software (*Soft*) through an exact match. Concurrently, the phrase 'contain malicious Microsoft Office' undergoes alignment through a fuzzy match function. This involves embedding comparison and similarity scoring, leading to the accurate inference of Technique T1566 - Phishing. The COMAT programs are available on GitHub [39] for community use and further research.

E. TTP Identification

While the COMAT framework effectively predicts techniques using its various inference paths by primarily leveraging historical and static relationships within the ontology, CTI reports often articulate complex scenarios and novel tactics that extend beyond the scope of predefined ontological mappings. To address the evolving complexity of threat reports, we've crafted a deep learning model that utilizes sentence transformers (such as BERT) to analyze CTI reports. This model features a fully connected (dense) layer that classifies nuanced sentence embeddings, linking them to specific MITRE techniques.

We chose BERT for document representation, given its renowned prowess in achieving state-of-the-art results across a multitude of natural language processing tasks and its adaptability to diverse challenges. For our specific task, we fine-tuned a BERT model tailored for document classification. While our approach is rooted in the conventional BERT architecture, there are some distinctions. Notably, when inputting a series of words, we appended a fully connected layer atop the BERT CLS vector. This enhancement serves as our classifier, optimizing the BERT model for text classification, as depicted by the topic classifier in Fig. 4. Although fine-tuning BERT achieved great performance on many text classification tasks, such as sentiment analysis [48] and fake news classification [49], it still has some problems. When text is very long, BERT cannot achieve proper results because of a limited sequence length of 512 tokens [50], leading to further information loss. Thus, Lu et al. proposed a sentence-level BERT [51]. It first leveraged pre-trained Roberta [47] to turn each sentence into sentence embedding and regarded the sentence as a basic input unit in BERT instead of a token. We refer to the architecture proposed by the paper [51] because CTI reports always contain hundreds of words and techniques are usually described by one or two sentences. As a result, we applied a sentence-level BERT model as CTI report representation.

To effectively process the content of CTI reports, we designed the SBMDE which capitalizes on the sentence-level structure and integrates MITRE knowledge to enhance the identification of TTPs. Sentences are first tokenized using the nltk package [52] and then transformed into embeddings with pre-trained Roberta, calculating the average word embeddings for each sentence as shown in Equation (1). These sentence embeddings are stored as features of the SBMDE model (2).

$$SE_i = \frac{\sum_{i=1}^{m} Roberta(W_i)}{m} \quad (1)$$

$$C = (SE_1,\ SE_2, \ldots\ldots,\ SE_n) \quad (2)$$

Given that CTI reports often describe multiple techniques, our task falls into the realm of multi-label classification, where a single report may be associated with several classes (techniques). Addressing this complexity requires a model that not only identifies various techniques but also accurately classifies each one within the context of the report. To meet this challenge, we designed the SBMDE model, depicted in Fig. 8, which is adept at handling multi-label scenarios by considering the diverse range of techniques a CTI report might encompass.

In our model, every sentence from a CTI report is transformed into an embedding via Roberta, effectively capturing the unique contextual nuances of each sentence. These sentence embeddings are then combined with embeddings of technique descriptions, enabling the model to associate specific sentences with relevant techniques. This concatenation of context-aware sentence vectors with technique descriptions is crucial for accurately classifying each sentence under multiple potential technique labels.

To discern the intricate relationships between sentences in a CTI report and their corresponding MITRE techniques, we've integrated a multi-head self-attention layer. This layer analyzes various sentence positions within the report, synthesizing them into a coherent representation of the sequence. The culmination of this deep learning process is manifested in the fully connected layer, which yields a multi-label output (DL_{ans}). This output is a comprehensive list representing the probabilities of various techniques as they are mentioned or implied in the CTI report.

For instance, consider a hypothetical CTI report detailing a cyber-attack. The COMAT identifies techniques such as 'T1059 - Command and Scripting Interpreter' and 'T1204 - User Execution,' on its inference paths that link threat actions, groups, and software identified in the report. However, SBMDE, by analyzing the nuances of the report's content, identifies these techniques and additionally predicts 'T1566 - Phishing,' a technique that COMAT did not capture due to its reliance on pre-defined ontological relationships that may not include the latest tactics or emerging threats.

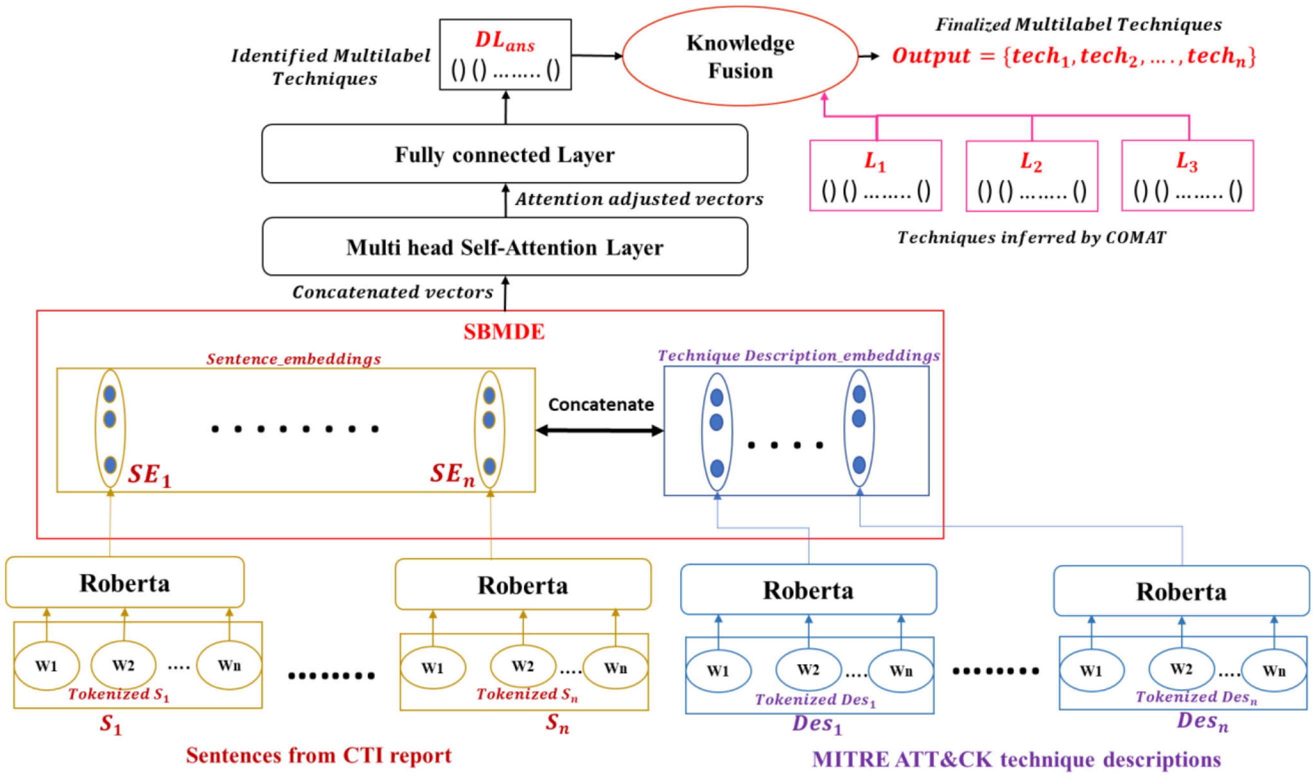

Fig. 8. Deep learning model architecture with Knowledge Fusion process.

Algorithm 1 Level-Voting Algorithm

Input: DL_{ans}, $L1$, $L2$, $L3$
Output: $output$

procedure Level-Voting (DL_{ans}, $L1$, $L2$, $L3$)
  $i \leftarrow 1$
  output $\leftarrow DL_{ans}$
  for $i \leq len(output)$ **do**
  if $DL_{ans}[i] = 1$ and $L1[i] = 1$ **then**
    $output[i] = 1$
  if $L2[i] = 1$ or $L3[i] = 1$ **then**
    $output[i] = 1$

To reconcile these differences and ensure a comprehensive analysis, we combine the outputs from both models—COMAT's L1, L2, and L3 results, and the SBMDE deep learning model's DL_{ans} into the level voting algorithm described in the following section. This process enables us to finalize the set of techniques present in the CTI report, ensuring that both historical data and current context are considered in our analysis.

F. Knowledge Fusion

In our methodology, we enhance the technique prediction process by integrating knowledge from the MITRE ATT&CK framework with the content of CTI reports using a level-voting algorithm as mentioned in Algorithm 1. It consolidates the findings from both the COMAT ontology and the SBMDE deep learning model, adjudicating the final output based on a consensus across multiple levels of inference. The inputs for this level-voting procedure are the technique predictions from SBMDE deep learning model (DL_{ans}) and the inference results from COMAT's first, second, and third levels ($L1$, $L2$, and $L3$ respectively). The level-voting algorithm functions as follows: It begins with the initial assumption that the output is equivalent to the predictions made by the SBMDE model (DL_{ans}).

It then iteratively checks each technique. If both the SBMDE model and the $L1$ inference suggest the presence of a technique, the algorithm confirms this prediction by assigning a '1' to the corresponding technique in the output. This is based on the rationale that $L1$ captures techniques previously utilized by the group or software, or those linked to threat actions mentioned in the CTI report—information which the SBMDE model corroborates.

Furthermore, if either $L2$ or $L3$ infers the technique's presence, a '1' is similarly assigned to that technique in the output. This is because $L2$ and $L3$ encompasses techniques historically associated with the threat actions undertaken by the group or software documented in the report, thus carrying a high likelihood of being mentioned again.

V. Evaluation

A. Data Collection

In our study, we meticulously gathered a dataset from diverse sources to evaluate the effectiveness of MITREtrieval. We sourced reports from the rcATT database, which provides a rich, manually labeled collection, excluding non-English entries. Additionally, we tapped into a comprehensive GitHub

TABLE III
DATASET STATISTICS

Sources	Number	Mean Technique per report (SD)	Mean report per technique (SD)
rcATT[19]	613	7.98 (6.90)	32.63 (45.51)
APT cybercriminal [53]	66	14.44 (9.55)	6.86 (6.90)
Self-Collected	45	12.02 (8.03)	4.61 (4.59)

TABLE IV
TRAINING AND TESTING SET NUMBER

Threshold	Training	Testing
6 (113 ttps)	584	100
40 (46 ttps)	560	74
80 (23 ttps)	550	62

directory that archives APT reports dated from 2006 to 2022. To further enrich our dataset, we handpicked recent reports from 2020 to 2022 labeled with MITRE ATT&CK techniques from various cybersecurity entities such as Cyble and Fortinet. This dataset amalgamation resulted in a diverse array of reports, with our dataset statistics revealing an imbalanced distribution of MITRE techniques, characterized by a high standard deviation in the number of reports per technique as shown in Table III. With most techniques represented by less than 50 samples, the challenge was evident in creating a robust system for large-scale multi-label classification.

To assess MITREtrieval's performance across different scales, we set three technique thresholds—113, 46, and 23 TTPs—and divided our dataset accordingly, allocating 90% for training and the remaining 10% for testing. This strategic distribution ensured a balanced representation of techniques in both the training and testing sets, as summarized in our dataset statistics table (Table IV).

B. Evaluation Metric

In our study, we emphasize the importance of recall over precision due to the critical nature of accurately identifying TTPs in CTI reports. A high recall rate is vital to minimize the risk of overlooking potential threats, which is crucial in cybersecurity contexts. To reflect this priority, we adopt the F2 score as our primary evaluation metric. The F2 score places a greater emphasis on recall, weighing it twice as heavily as precision.

This metric is used for both micro-averaging and macro-averaging analyses, ensuring our evaluation strategy is aligned with our primary objective: to capture a comprehensive range of relevant techniques with minimal omissions. The adoption of the F2 score underscores our commitment to a thorough approach in technique identification within CTI reports. The calculation for micro-average precision, recall, and F2 score is shown in equation (3), (4) and (5) respectively.

$$P_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{3}$$

$$R_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{4}$$

$$F2_{micro} = \frac{5 * \ P_{micro} * R_{micro}}{4 * \ P_{micro} + R_{micro}} \tag{5}$$

In addition to micro-averaging, we employ macro-averaging to calculate precision, recall, and the F2 score as shown in equations (6), (7) and (8) respectively. The advantage of using macro-averaging lies in its ability to treat each class (technique) equally, regardless of the number of instances (support values) it has in the dataset. This approach is particularly beneficial when evaluating techniques that are less frequently represented in the samples. By giving equal weight to all classes, macro-averaging ensures that the performance of these fewer common techniques significantly influences the overall metric, providing a balanced view of the model's effectiveness across all techniques, not just the ones that occur frequently.

$$P_{macro} = \frac{\sum_{i=1}^{n} \frac{TP_i}{TP_i + \ FP_i}}{n} \tag{6}$$

$$R_{macro} = \frac{\sum_{i=1}^{n} \frac{TP_i}{TP_i + \ FN_i}}{n} \tag{7}$$

$$F2_{macro} = \frac{\sum_{i=1}^{n} \frac{5TP_i}{4TP_i + \ FP_i + FN_i}}{n} \tag{8}$$

C. Overall Performance

We compared the performance of MITREtrieval with three existing state-of-the-art methods: rcATT [19], TTPDrill [2], and AttacKG [36], along with a related deep neural network model [54] and TRAM, a system introduced by MITRE that trains a sentence classifier by manually labeling sentences. These comparisons were conducted at three different thresholds on the same dataset.

The results, summarized in Table V, demonstrate that MITREtrieval consistently outperforms other models across all thresholds in terms of F2 Score. Particularly at the 23 TTP threshold, MITREtrieval achieves an impressive 69% and 71% F2 Score for micro and macro averages, respectively, showcasing its effectiveness. AttacKG [36], at 113 TTP threshold, shows a remarkable macro-average recall of 51%, outstripping its micro-average recall of 24%. This suggests that AttacKG excels at capturing a wide array of TTPs, particularly in the macro sense, although it does not consistently do so across individual classes as reflected by the micro-average.

As the threshold decreases (i.e., moving from 113 TTPs to 23 TTPs), the performance of most models tends to improve. This could indicate that as the complexity of the task decreases, models become better at classification. However, there is a trade-off between precision and recall in several models. For instance, while MITREtrieval demonstrates a balance between these metrics, AttacKG [36] stands out with its consistently high precision rate in macro average, especially significant at the 23 TTP threshold. This indicates AttacKG's strength in capturing a broad spectrum of true positives. In contrast, TTPdrill [2] shows a tendency for higher recall at the cost of precision at the 113 TTP threshold, leading to more misclassifications.

TABLE V
PERFORMANCE MEASURE THROUGH DIFFERENT THRESHOLDS

Threshold	**6(113 TTPs)**					
Paper	**Micro Average**			**Macro Average**		
	Recall	**Precision**	**F2 Score**	**Recall**	**Precision**	**F2 Score**
HAN [54]	30%	6%	21%	35%	9%	22%
AttacKG [36]	24%	12%	20%	51%	18%	37%
rcATT[19]	25%	42%	34%	19%	28%	20%
TRAM [25]	48%	28%	42%	50%	26%	42%
TTPdrill [2]	64%	14%	37%	65%	13%	36%
MITREtrieval	**74%**	31%	**58%**	**69%**	30%	**55%**
Threshold	**40(46 TTPs)**					
HAN [54]	38%	18%	31%	24%	13%	21%
AttacKG [36]	31%	24%	29%	50%	49%	50%
rcATT[19]	39%	47%	40%	34%	43%	35%
TRAM [25]	62%	30%	51%	63%	28%	50%
TTPdrill [2]	68%	19%	45%	73%	18%	45%
MITREtrieval	**78%**	35%	**62%**	**76%**	35%	**62%**
Threshold	**80(23 TTPs)**					
HAN [54]	29%	25%	28%	29%	12%	23%
AttacKG [36]	67%	51%	63%	51%	50%	51%
rcATT[19]	50%	52%	50%	44%	47%	45%
TRAM [25]	73%	31%	57%	73%	29%	56%
TTPdrill [2]	67%	24%	49%	72%	23%	50%
MITREtrieval	**85%**	39%	**69%**	**84%**	43%	**71%**

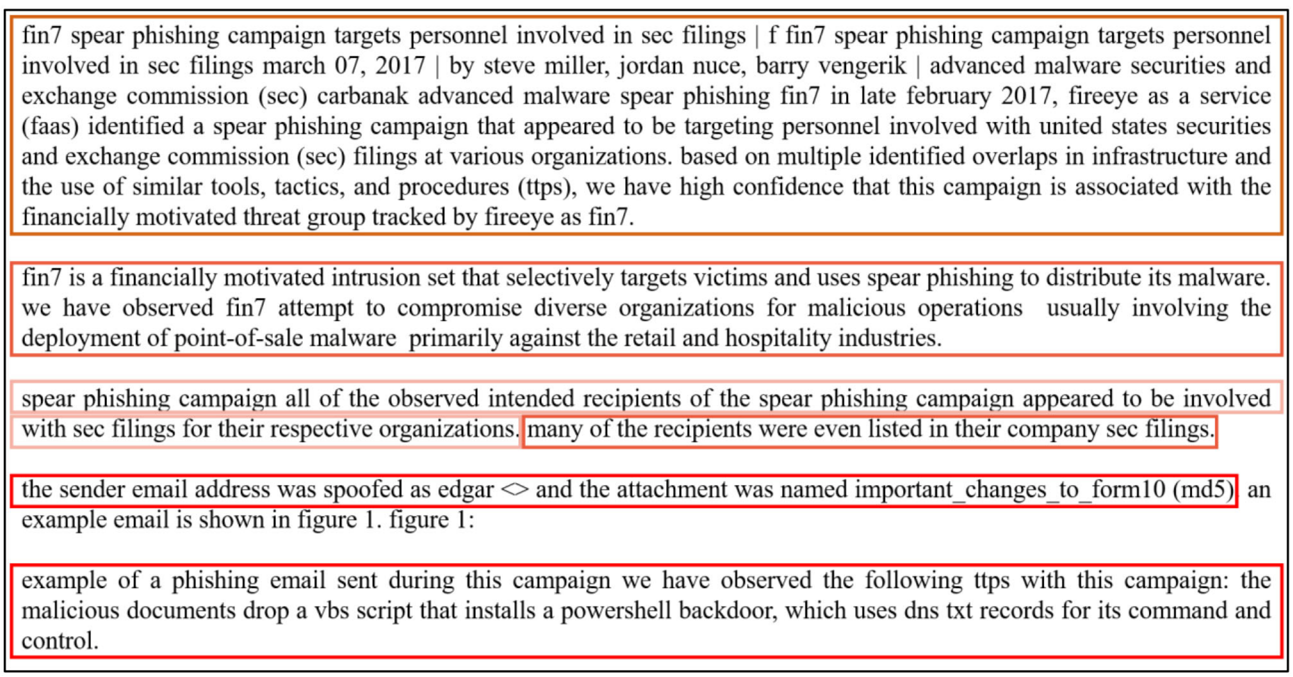

Fig. 9. Attention visualization of SBMDE deep learning model.

There's a noticeable discrepancy between micro and macro averages for some models. For instance, rcATT [19] at 113 TTPs has a significantly higher precision in the micro average than in the macro average, which might suggest it performs disproportionately well in some classes while struggling with others.

To further ilustrate our model's accuracy in identifying TTPs within CTI reports, we visualize the last layer's attention weights in Fig. 9. Darker red boundary boxes in the figure indicate sentences of greater importance to the neural network. As seen, our model focuses more intensely on technique-related sentences,

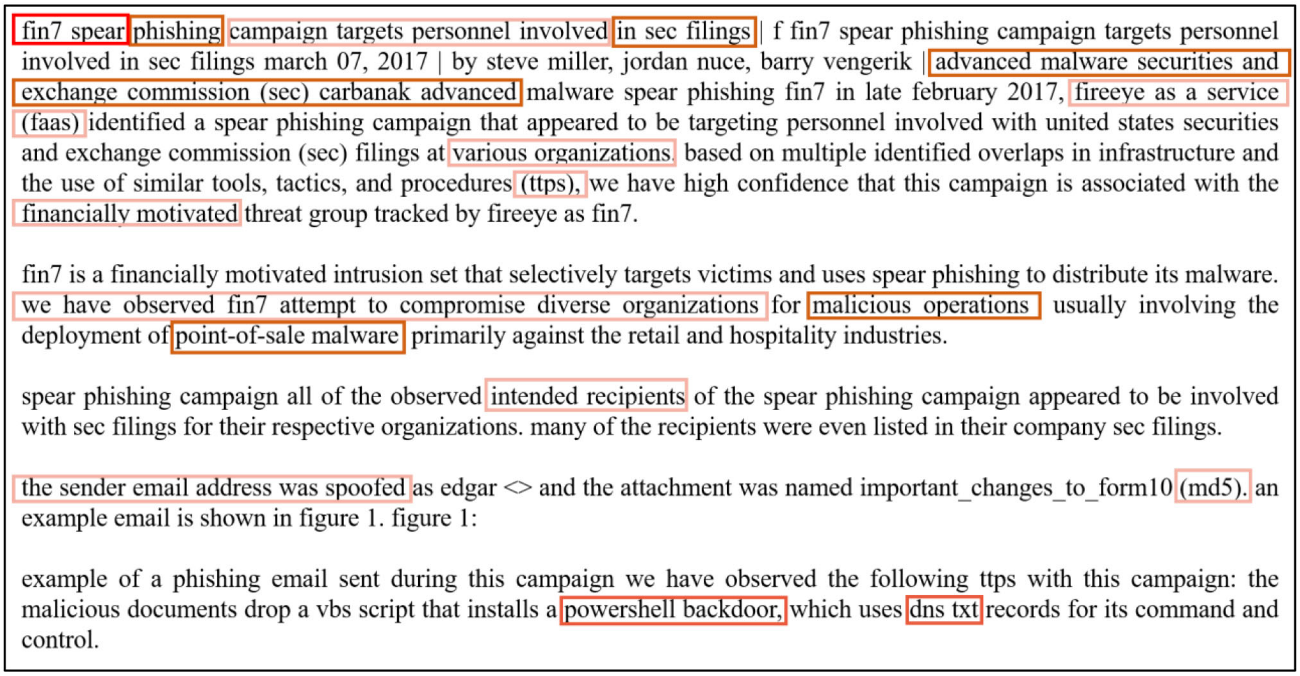

Fig. 10. rcATT [19] visualization.

TABLE VI
PERFORMANCE MEASURE BEFORE AND AFTER INTEGRATING TOPIC CLASSIFIER

Threshold	Micro Average						Macro Average					
	Recall		Precision		F2 Score		Recall		Precision		F2 Score	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
6 (113 TTPs)	64%	**65%**	26%	**28%**	50%	**51%**	56%	**59%**	25%	**28%**	45%	**48%**
40 (46 TTPs)	71%	67%	26%	**32%**	53%	**55%**	60%	**62%**	22%	**31%**	45%	**52%**
80 (23 TTPs)	71%	**78%**	35%	35%	61%	**63%**	64%	**74%**	36%	36%	51%	**61%**

highlighting its ability to identify key information in CTI reports.

In addition, we visualized the weight distribution in a term-frequency-based model, rcATT [19], as shown in Fig. 10. For this visualization, boundary boxes with varying shades of red are employed to denote word significance. Words that the model deems highly important are highlighted with darker red boundary boxes, while less significant words are enclosed in lighter shades.

Such a pattern is indicative of the limitations inherent in TF-IDF methods, particularly in sparse datasets where key information might be omitted or keywords may not be accurately prioritized. This often leads to a high false negative rate, a significant challenge that our MITREtrieval effectively addresses. In summary, these visualizations and comparative analyses underscore the advanced capability of MITREtrieval in accurately classifying TTPs. The consistent outperformance of MITREtrieval, especially in terms of F2 Score which emphasizes recall, confirms its suitability for tasks where minimizing false negatives is critical.

D. The Effect of Topic Classifier

The introduction of a topic classifier within the MITREtrieval system marks a critical advancement in our technique classification process. This classifier plays a vital role in sifting through the extensive text of CTI reports to identify sentences likely to contain relevant information about the techniques.

For the development of this binary sentence classifier, we compiled a comprehensive dataset of 9305 sentences from MITRE and 9132 from various news sources. The objective was to train the classifier to distinguish between sentences related to MITRE ATT&CK techniques and those that are not. We used 90% of this dataset for training and 10% for testing.

Upon integrating the topic classifier into our SBMDE model, we observed a substantial improvement in the model's performance. The classifier has demonstrated remarkable precision and recall, with scores of approximately 99.5% and 99.7%, leading to an F2 Score of 99.6%. This accuracy highlights the classifier's ability to effectively filter out irrelevant content from the CTI report.

The integration of the topic classifier into the MITREtrieval system has demonstrably enhanced its performance, as detailed in Table VI. Key improvements post-classifier implementation are evident in both micro and macro averaging metrics. For instance, at the 113 TTP threshold, the micro average F2 Score improved from 50% to 51%, and the macro average precision rose from 25% to 28%. Notably, at the more stringent 23 TTP

TABLE VII
PERFORMANCE MEASURE WITH (W) AND WITHOUT (WO) FUSION THROUGH DIFFERENT THRESHOLDS

Threshold	Micro Average						Macro Average					
	Recall		Precision		F2 Score		Recall		Precision		F2 Score	
	WO	W	WO	W	WO	W	WO	W	WO	W	WO	W
6 (113 TTPs)	65%	**74%**	28%	**31%**	51%	**58%**	62%	**69%**	31%	30%	48%	**55%**
40 (46 TTPs)	67%	**78%**	32%	**35%**	55%	**62%**	62%	**76%**	31%	**35%**	52%	**62%**
80 (23 TTPs)	78%	**85%**	35%	**39%**	63%	**69%**	74%	**84%**	36%	**43%**	61%	**71%**

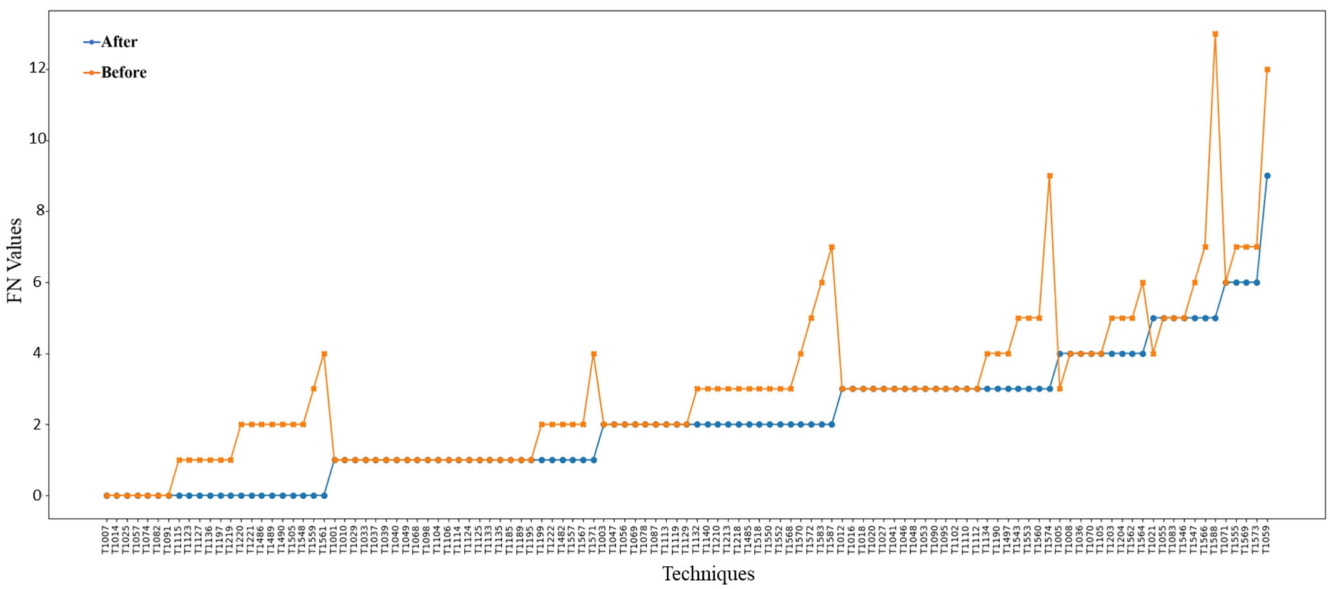

Fig. 11. False Negative values before and after fusion.

threshold, there was a significant increase in micro average recall from 71% to 78%, leading to an F2 Score boost from 61% to 63%. These enhancements across various thresholds underscore the classifier's role in elevating the precision and overall efficacy of the MITREtrieval system.

E. The Effect of Fusion

MITREtrieval, a system that integrates the SBMDE model with COMAT's ontology knowledge from MITRE ATT&CK, aims to enhance the discovery of TTPs in CTI reports. To quantify the impact of this fusion, we assessed MITREtrieval's performance across various thresholds, both with and without the incorporation of ontology knowledge. Table VII details these findings, showing a marked improvement in performance due to the fusion: approximately 7% in F_{micro} and 10% in F_{macro}.

This enhancement is further illustrated in Figures 11 and 12, where we plotted the false negatives (FN) and false positives (FP) for each technique before and after integrating ontology knowledge. Fig. 11 demonstrates a significant decrease in FNs, while Fig. 12 indicates a minimal or negligible increase in FPs. These changes suggest that the fusion effectively addresses gaps in the deep learning model by leveraging the rich, structured knowledge from COMAT.

The substantial decrease in FNs, along with a controlled increase in FPs, highlights the synergy between SBMDE's deep learning capabilities and COMAT's ontology knowledge. This fusion not only improves technique classification accuracy but also ensures a comprehensive understanding of CTI reports, particularly where standalone training data might be insufficient. Building upon these insights, the following section will delve into a detailed comparison of the individual contributions of SBMDE and COMAT, as well as their combined strength in the context of MITREtrieval.

F. Comparative Analysis of SBMDE and COMAT

This section presents a comparative analysis of the individual and combined performances of the SBMDE (the DNN model) and COMAT highlighting their fusion.

Our objective is to assess the distinct contributions of each approach and the enhanced capability achieved through their integration. In Table VIII, performance at the 113 TTP threshold is compared for COMAT, SBMDE, and their fusion. COMAT records micro and macro F2 scores of 43% and 34%, with recall at 54% and 42%, and precision at 24% and 20%, respectively. Conversely, SBMDE exhibits higher micro and macro F2 scores of 51% and 48%, showcasing superior contextual analysis in CTI reports.

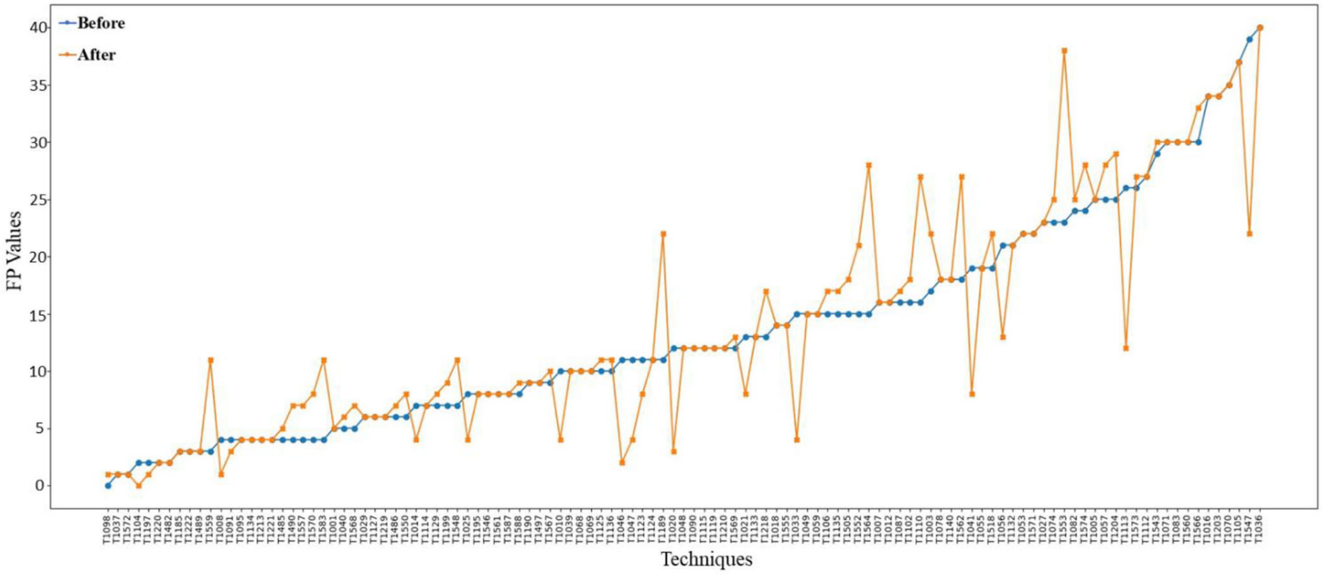

Fig. 12. False Positive values before and after fusion.

TABLE VIII
COMPARATIVE PERFORMANCE METRICS OF COMAT AND SBMDE DEEP LEARNING MODEL

Model	Micro Averaging			Macro Averaging		
	R	P	F2	R	P	F2
COMAT	54%	24%	43%	42%	20%	34%
SBMDE Deep Learning	65%	28%	51%	62%	31%	48%
Fusion (SBMDE + COMAT)	**74%**	**31%**	**58%**	**69%**	30%	**55%**

However, the fusion of SBMDE and COMAT leads to a noteworthy enhancement in performance, with micro and macro F2 scores increasing to 58% and 55%, respectively. This improvement underscores the complementary strengths of the two models. While COMAT's ontology-based approach excels in identifying techniques based on established relationships, it may not fully capture contextually rich or emerging technique descriptions.

Conversely, SBMDE's deep learning methodology, despite its contextual depth, could miss certain techniques due to training data imbalances or nuanced representations. The combined use of SBMDE and COMAT effectively mitigates these individual limitations. The fusion allows SBMDE's contextual accuracy to be augmented by COMAT's extensive ontological knowledge, resulting in a significant reduction in false negatives and a controlled increase in false positives.

VI. CONCLUSION

This paper introduces MITREtrieval, a sophisticated system designed for the efficient extraction of large-scale MITRE ATT&CK techniques from unstructured CTI reports. At its core, MITREtrieval integrates the transformer-based Sentence BERT with the MITRE Description Embedding (SBMDE) model and the MITRE ATT&CK COMAT ontology, capitalizing on their strengths in contextual analysis and knowledge representation, respectively.

Our evaluation demonstrates that MITREtrieval surpasses existing state-of-the-art methods by 20% in terms of the F2 score across various technique thresholds. This significant improvement underscores the synergy between deep learning and ontology-based approaches, especially in scenarios characterized by data imbalance and scarcity.

MITREtrieval marks a substantial advancement in threat intelligence gathering, offering a highly effective tool for classifying a wide range of MITRE ATT&CK techniques within CTI reports. This capability is crucial for proactively defending against evolving cyber threats. However, it is important to note that the system still requires oversight from cybersecurity experts to address potential mispredictions, predominantly due to the system's inherent tendency towards a higher false positive rate. Despite this, MITREtrieval significantly reduces the time required for TTP retrieval – achieving up to a fourfold decrease compared to manual processes. This efficiency, coupled with the ease of correcting mispredictions relative to missed TTPs, positions MITREtrieval as a valuable asset in the domain of cybersecurity analysis.

APPENDIX
INDEX OF ABBREVIATIONS

APT	Advanced Persistent Threat
BERT	Bidirectional Encoder Representations from Transformers
BM25	Best Matching 25
CLS	Classification vector from BERT
COMAT	Cybersecurity Ontology based on MITRE ATT CK
CTI	Cyber Threat Intelligence
DNN	Deep Neural Network

FN	False Negatives
FP	False Positives
ICS	Industrial Control Systems
IoC	Indicator of Compromise
MALOnt	Malware Ontology
MITRE ATT CK	MITRE Adversarial Tactics, Techniques, and Common Knowledge
NER	Named Entity Recognition
NLP	Natural Language Processing
POS	Parts of Speech
Roberta	A robustly optimized BERT approach
SBMDE	Sentence BERT with MITRE Description Embedding
SRL	Semantic Role Labeler
STIX	Structured Threat Information eXpression
TAXII	Trusted Automated Exchange of Intelligence Information
TRAM	Threat Report ATT CK Mapping
TTP	Tactics, Techniques, and Procedures
UCO	Unified Cybersecurity Ontology
YARA	Yet Another Recursive Acronym (a malware identification tool)
rcATT	report classifier based on ATT CK tactics and techniques

REFERENCES

[1] N. Virvilis and D. Gritzalis, "The big four-what we did wrong in advanced persistent threat detection?" in *Proc. Int. Conf. Availabil., Reliabil. Security*, 2013, pp. 248–254.

[2] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and X. Niu, "Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of CTI sources," in *Proc. 33rd Annu. Comput. Secur. Appl. Conf.*, 2017, pp. 103–115.

[3] G. Husari, X. Niu, B. Chu, and E. Al-Shaer, "Using entropy and mutual information to extract threat actions from cyber threat intelligence," in *Proc. IEEE Int. Conf. Intell. Security Informat. (ISI)*, 2018, pp. 1–6.

[4] S. Zhou, Z. Long, L. Tan, and H. Guo, "Automatic identification of indicators of compromise using neural-based sequence labelling," 2018, *arXiv:1810.10156*.

[5] Z. Zhu and T. Dumitras, "Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports," in *Proc. IEEE Eur. Symp. Security Privacy (EuroS&P)*, 2018, pp. 458–472.

[6] Y. Gao, L. Xiaoyong, P. Hao, B. Fang, and P. Yu, "HinCTI: A cyber threat intelligence modeling and identification system based on heterogeneous information network," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 708–722, Feb. 2022.

[7] W. Tounsi and H. Rais, "A survey on technical threat intelligence in the age of sophisticated cyber-attacks," *Comput. Secur.*, vol. 72, pp. 212–233, Jan. 2018.

[8] S. Barnum, *Standardizing Cyber Threat Intelligence Information With the Structured Threat Information Expression (STIX), Volume 11*, Mitre Corp., McLean, VA, USA, pp. 1–22, 2012.

[9] J. Connolly, *Mark Davidson, and Charles Schmidt the Trusted Automated Exchange of Indicator Information (TAXII)*. Mitre Corp., McLean, VA, USA, pp. 1–20, 2014.

[10] "The pattern matching knife, VirusTotal." Yara. Accessed: May 19, 2022. [Online]. Available: https://virustotal.github.io/yara/

[11] S. Qamar, Z. Anwar, M. A. Rahman, E. Al-Shaer, and B.-T. Chu, "Data-driven analytics for cyber-threat intelligence and information sharing," *Comput. Secur.*, vol. 67, pp. 35–58, Jun. 2017.

[12] T. D. Wagner, K. Mahbub, E. Palomar, and A. E. Abdallah, "Cyber threat intelligence sharing: Survey and research directions," *Comput. Secur.*, vol. 87, Nov. 2019, Art. no. 101589.

[13] P. Gao et al., "Enabling efficient cyber threat hunting with cyber threat intelligence," in *Proc. IEEE 37th Int. Conf. Data Eng. (ICDE)*, 2021, pp. 193–204.

[14] F. Maym, R. Bixler, R. Jones, and S. Lathrop, "Towards a definition of cyberspace tactics, techniques and procedures," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2017, pp. 4674–4679.

[15] "TTPs Within Cyber Threat Intelligence," Accessed: Apr. 26, 2022. [Online]. Available: https://www.optiv.com/explore-optiv-insights/blog/tactics-techniques-and-procedures-ttps-within-cyber-threat-intelligence

[16] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, and C. B. Thomas, "Mitre att&ck: Design and philosophy," Mitre Corp., McLean, VA, USA, Rep. MP180360R1, 2018.

[17] "Adversarial tactics, techniques and common knowledge." MITRE. Accessed: Feb. 16, 2022. [Online]. Available: https://attack.mitre.org/

[18] "TTP Cyber security." Accessed: May 3, 2022. [Online]. Available: https://trustnetinc.com/ttp-cyber-security/

[19] V. Legoy, M. Caselli, C. Seifert, and A. Peter, "Automated retrieval of ATT&CK tactics and techniques for cyber threat reports," 2020, *arXiv:2004.14322*.

[20] G. Ayoade, S. Chandra, L. Khan, K. Hamlen, and B. Thuraisingham, "Automated threat report classification over multi-source data," in *Proc. IEEE 4th Int. Conf. Collab. Internet Comput. (CIC)*, 2018, pp. 236–245.

[21] Y. Ghazi, Z. Anwar, R. Mumtaz, S. Saleem, and A. Tahir, "A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, 2018, pp. 129–134.

[22] S. Peng, A. Zhou, S. Liao, and L. Liu, "A threat actions extraction method based on the conditional co-occurrence degree," in *Proc. 7th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, 2020, pp. 1633–1637.

[23] M. Li, R. Zheng, L. Liu, and P. Yang, "Extraction of threat actions from threat-related articles using multi-label machine learning classification method," in *Proc. 2nd Int. Conf. Safety Produce Informatization (IICSPI)*, 2019, pp. 428–431.

[24] A. Niakanlahiji, J. Wei, and B.-T. Chu, "A natural language processing-based trend analysis of advanced persistent threat techniques," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2018, pp. 2995–3000.

[25] *Threat Report ATT & CK Mapping*, Mitre Corp., McLean, VA, USA, 2022.

[26] Y. You et al., "Tim: threat context-enhanced TTP intelligence mining on unstructured threat data," *Cybersecurity*, vol. 5, no. 1, pp. 1–17, 2022.

[27] A. Tuor, S. Kaplan, B. Hutchinson, N. Nichols, and S. Robinson, "Deep learning for unsupervised insider threat detection in structured cybersecurity data streams," in *Proc. Workshops 31st AAAI Conf. Artif. Intell.*, 2017, pp. 224–231.

[28] A. Kesari, "Predicting cybersecurity incidents through mandatory disclosure regulation," SSRN.com. 2020. [Online]. Available: https://ssrn.com/abstract=3700243

[29] Z.-J. Liao, "COMAT: A cybersecurity ontology based on MITRE ATT&CK," M.S. thesis, Dept. Comput. Sci. Inf. Eng., Nat. Cent. Univ., Taiwan, 2022.

[30] W. U. Hassan, A. Bates, and D. Marino, "Tactical provenance analysis for endpoint detection and response systems," in *Proc. IEEE Symp. Security Privacy (SP)*, 2020, pp. 1172–1189.

[31] K. Oosthoek and C. Doerr, "Cyber threat intelligence: A product without a process?" *Int. J. Intell. CounterIntell.*, vol. 34, no. 2, pp. 300–315, 2021.

[32] C.-M. Chen et al., "Threat action extraction using information retrieval," in *Proc. CS & IT Conf. Proc.*, 2021, pp. 13–19.

[33] Z. Syed, A. Padia, T. Finin, L. Mathews, and A. Joshi, "UCO: A unified cybersecurity ontology," in *Proc. Workshops 30th AAAI Conf. Artif. Intell.*, 2016, pp. 195–202.

[34] N. Rastogi, S. Dutta, M. J. Zaki, A. Gittens, and C. Aggarwal, "MALOnt: An ontology for malware threat intelligence," in *Proc. Int. Workshop Deployable Mach. Learn. Secur. Defense* 2020, pp. 28–44.

[35] G. Amati, "BM25," in *Encyclopedia of Database Systems*, vol. 1. Boston, MA, USA: Springer, 2009, pp. 257–260.

[36] Z. Li, J. Zeng, Y. Chen, and Z. Liang, "AttacKG: Constructing technique knowledge graph from cyber threat intelligence reports," in *Proc. Eur. Symp. Res. Comput. Secur.*, 2022, pp. 589–609.

[37] N. Rani, B. Saha, V. Maurya, and S. K. Shukla, "TTPHunter: Automated extraction of actionable intelligence as TTPs from narrative threat reports," in *Proc. Aust. Comput. Sci. Week*, 2023, pp. 126–134.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[39] "MITREtrieval." Accessed: Feb. 5, 2024. [Online]. Available: https://github.com/wmlab-MITREtreival/MITREtrieval

[40] K. Satvat, R. Gjomemo, and V. N. Venkatakrishnan, "Extractor: Extracting attack behavior from threat reports," in *Proc. IEEE Eur. Symp. Security Privacy (EuroS P)*, 2021, pp. 598–615.

[41] X. Hu, L. Wen, Y. Xu, C. Zhang, and P. S. Yu, "SelfORE: Self-supervised relational feature learning for open relation extraction," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2020, pp. 3673–3682.

[42] A. Bastos et al., "RECON: Relation extraction using knowledge graph context in a graph neural network," 2021, *arXiv:2009.08694*.

[43] K. Zhang et al., "Open hierarchical relation extraction," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist., Human Lang. Technol.*, 2021, pp. 5682–5693.

[44] A. Belfadel, M. Gagnant, M. Dussartre, J. Picault, L. Crochepierre, and S. Tmar, "A semi-automatic framework towards building electricity grid infrastructure management ontology: A case study and retrospective," in *Proc. Int. Conf. Web Inf. Syst. Eng.*, 2023, pp. 865–874. [Online]. Available: https://doi.org/10.1007/978-981-99-7254-8_67.

[45] A. Piplai, S. Mittal, A. Joshi, T. Finin, J. Holt, and R. Zak, "Creating cybersecurity knowledge graphs from malware after action reports," *IEEE Access*, vol. 8, pp. 211691–211703, 2020.

[46] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the IOC game: Toward automatic discovery and analysis of open-source cyber threat intelligence," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 755–766.

[47] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.

[48] H. Xu, B. Liu, L. Shu, and P. S. Yu. "BERT post-training for review reading comprehension and aspect-based sentiment analysis," 2019, *arXiv:1904.02232*.

[49] B. Bhutani, N. Rastogi, P. Sehgal, and A. Purwar, "Fake news detection using sentiment analysis," in *Proc. 12th Int. Conf. Cont. Comput. (IC3)*, 2019, pp. 1–5.

[50] M. Ding, C. Zhou, H. Yang, and J. Tang, "Cogltx: Applying bert to long texts," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Assoc., Inc., 2020.

[51] J. Lu, M. Henchion, I. Bacher, and B. Mac Namee, "A sentence-level hierarchical bert model for document classification with limited labelled data," in *Proc. Int. Conf. Discovery Sci.*, 2021, pp. 231–241.

[52] "NLTK documentation." 2017. [Online]. Available: https://www.nltk.org/

[53] "Apt & cybercriminals campaign collection." CyberMonitor. 2022. [Online]. Available: https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections

[54] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Am. Ch. Assoc. Comput. Linguist., Human Lang. Technol.*, 2016, pp. 1480–1489.

Yi'-Ting Huang received the Ph.D. degree in information management from National Taiwan University in 2015. She was a Postdoctoral Fellow with Academia Sinica. She is currently an Assistant Professor with the National Taiwan University of Science and Technology. Her primary research interests include malware analysis, deep learning, and natural language processing in educational applications.

R. Vaitheeshwari received the B.E. and M.E. degrees in electronics and communication engineering from Anna University, Chennai, India, in 2017 and 2019, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering, National Central University, Taiwan. She worked as a Lecturer with VelTech University, Chennai, until 2021. Her research interests include applications of artificial intelligence, natural language processing, and cybersecurity.

Meng-Chang Chen received the Ph.D. degree in computer science from the University of California at Los Angeles in 1989. He was with AT&T Bell Labs. He is currently a Research Fellow/Professor with the Institute of Information Science, Academia Sinica, Taiwan. His current research interests include computer and network security, deep learning for complicated applications, and data and knowledge engineering.

Ying-Dar Lin (Fellow, IEEE) received the Ph.D. degree in computer science from The University of California at Los Angeles in 1993. He is a Chair Professor with National Yang Ming Chiao Tung University and the Vice President of the National Institute of Cyber Security, Taiwan. He was a Visiting Scholar with Cisco Systems, San Jose, from 2007 to 2008, the CEO with Telecom Technology Center, Taiwan, from 2010 to 2011, and the Vice President of National Applied Research Labs, Taiwan, from 2017 to 2018. He was the Founder and the Director of Network Bench Marking Lab (NBL) from 2002 to 2018, which reviewed network products with real traffic and automated tools, and has been an approved test lab of the Open Networking Foundation (ONF). He also cofounded L7 Networks Inc. in2002, later acquired by D-Link Corporation, and O'Prueba Inc. as pin-off from NBL, in 2018. He published a textbook, *Computer Networks: An Open Source Approach*, with Ren-Hung Hwang and Fred Baker (McGraw-Hill, 2011). His research Interests include network security, wireless communications, network softwarization, and machine learning for communications. His work on multihop cellular was the first along this line, and has been cited over 1000 times and standardized into IEEE 802.11s, IEEE 802.15.5, IEEE 802.16j, and 3GPP LTE-Advanced. He received the Research Excellence Award and the K. T. Li Breakthrough Award in 2017. He has served or is serving on the editorial boards of several IEEE journals and magazines, including the Editor-in-Chief of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS from 2017 to 2020. He was an IEEE Distinguished Lecturer from 2014 to 2017 and an ONF Research Associate from 2014 to 2018.

Ren-Hung Hwang received the Ph.D. degree in computer science from the University of Massachusetts at Amherst, Amherst, MA, USA, in 1993. He is the Dean of the College of Artificial Intelligence, National Yang Ming Chiao Tung University (NYCU), Taiwan. Before joining NYCU, he was with National Chung Cheng University, Taiwan, from 1993 to 2022. His research interests include deep learning, network security, wireless communications, Internet of Things, and cloud and edge computing. He received the Best Paper Award from the 6th International Conference on Internet of Vehicles 2019, IEEE Ubi-Media 2018, IEEE SC2 2017, and IEEE IUCC 2014, and the IEEE Outstanding Paper Award from IEEE IC/ATC/ICA3PP 2012. He received the Outstanding Technical Achievement Award of the IEEE Tainan Section in 2022. He is currently on the editorial boards of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS and *IEICE Transactions on Communications*. He served as the General Chair for the International Computer Symposium 2016, the International Symposium on Pervasive Systems, Algorithms, and Networks 2018, the International Symposium on Computer, Consumer, and Control 2018, and IEEE DataCom 2019 (The 5th IEEE International Conference on Big Data Intelligence and Computing).

Po-Ching Lin (Member, IEEE) received the Ph.D. degree in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 2008. He joined the faculty of the Department of Computer Science and Information Engineering, CCU in August 2009. He is currently a Professor with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan. His research interests include network security, network traffic analysis, and performance evaluation of network systems.

Yuan-Cheng Lai received the Ph.D. degree from the Department of Computer and Information Science, National Chiao Tung University in 1997. He joined the faculty of the Department of Information Management, National Taiwan University of Science and Technology in August 2001, where he has been a Distinguished Professor since June 2012. His research interests include performance analysis, wireless networks, network security, and machine learning.

Eric Hsiao-Kuang Wu (Member, IEEE) received the B.S. degree in computer science and information engineering from National Taiwan University in 1989, and the master's and Ph.D. degrees in computer science from the University of California at Los Angeles in 1993 and 1997, respectively. He is currently a Professor of Computer Science and Information Engineering with National Central University, Taiwan. His research interests include wireless networks, mobile computing, and broadband networks. He is also a member of the Institute of Information and Computing Machinery.

Chung-Hsuan Chen received the master's degree in computer science and information engineering from National Central University, Taiwan, in 2022, where he delved deep into the intricacies of computer systems and programming languages. His research interest includes machine learning, cybersecurity, and ontology.

Zi-Jie Liao received the master's degree in computer science and information engineering from National Central University, Taiwan, in 2022. His research interests include machine learning, cybersecurity, and ontology.

Chung-Kuan Chen received the Ph.D. degree in computer science and engineering from National Chiao-Tung University (NCTU). He is currently the Research Director of CyCraft, and responsible for organizing the research team, and an Adjunct Assistant Professor with Soochow University, Taiwan. His research focuses on cyber-attack and defense, machine learning, software vulnerability, malware, and program analysis. He tries to utilize machine learning to assist malware analysis and threat hunting, and build automatic attack and defense systems. He has published several academic journal and conference papers, and has been involved in many large research projects from digital forensic, incident response to malware analysis. He also dedicates to security education. As the Founder of NCTU hacker research clubs, he trained students to participate in world-class security contests, and has experience of participating DEFCON CTF (2016 in HITCON Team and 2018 as coach in BFS Team). He organized the BambooFox Team to join some bug bounty projects and discover some CVEs in COTS software and several vulnerabilities in campus websites. Besides, he has presented technical presentations in technique conferences, such as BlackHat, HITCON, CHITB, RootCon, CodeBlue, FIRST, and VXCON. As an Active Member in Taiwan Security Community, he is the Chairman of HITCON Review Committee as well as the Director of Association of Hacker in Taiwan.