WAVE-27K: Bringing together CTI sources to enhance threat intelligence models

Felipe Castaño
Vicomtech
University of León
Bilbao, Spain

Amaia Gil-Lerchundi
Vicomtech
San Sebastian, Spain

Raul Orduna-Urrutia
Vicomtech
San Sebastian, Spain

Eduardo Fidalgo Fernandez
University of León
León, Spain

Rocío Alaiz-Rodríguez
University of León
León, Spain

Abstract

Considering the growing flow of information on the internet, and the increased incident-related data from diverse sources, unstructured text processing gains importance. We have presented an automated approach to link several CTI sources through the mapping of external references. Our method facilitates the automatic construction of datasets, allowing for updates and the inclusion of new samples and labels. Following this method we built a new dataset of unstructured CTI descriptions called Weakness, Attack, Vulnerabilities, and Events 27k (WAVE-27k). Our dataset includes information about 27 different MITRE techniques, containing 22539 samples related one technique and 5262 related to two or more techniques simultaneously. We evaluated five BERT-based models into the WAVE-27K dataset concluding that SecRoBERTa reaches the highest performance with a 77.52% F1 score. Additionally, we compare the performance of the SecRoBERTa on the WAVE-27K dataset and other public datasets. The results show that the model using the WAVE-27K dataset outperforms the others. These results demonstrate that the data within WAVE-27K contains relevant information and that the proposed method effectively built a dataset with a level of quality sufficient to train a machine-learning model.

1 Introduction

The growing flow of information has led to increased incident-related data from diverse sources, such as open-source intelligence (OSINT) platforms, cybersecurity analyst forums, and several other sources on the internet. Therefore, it is crucial to automatically process unstructured texts (Fujii et al., 2022) to extract information such as tactics, techniques, and procedures (TTPs) from different free-text sources to help understand and detect relevant incidents inside the local network.

In addition, algorithms designed to process unstructured texts for TTPs offer an advantage in their capacity to extract valuable insights from unconventional sources, such as Dark web forums and other suspicious platforms where malicious activities are documented and discussed. This capacity not only facilitates the detection and characterization of cyber attacks but also enables the identification of underground networks where new attacks are disseminated.

However, regardless of the advantages offered by algorithms designed to process unstructured CTI texts and their significant impact on the security of local networks, there is a need for a more extensive dataset of unstructured Cyber Threat Intelligence (CTI) texts. We hypothesize that enhancing the quality and quantity of accessible data will substantially improve the efficacy of state-of-the-art models.

To supply this lack, we acknowledge it is key to propose a methodology that takes advantage of the increasing flow of information mentioned above, providing automatic methods to create datasets and train algorithms focused on extracting and detecting TTPs from unstructured texts. Our method uses the information from Cyber Threat Intelligence (CTI) sources to automate the data collection process of information related to TTPs in unstructured text, reducing costs and ad hoc studies with limited data.

Our goal is to develop CTI tools that contribute to the community and help standardize the datasets in the state of the art allowing state of the art models comparison. Consequently, we address an approach that involves three steps: Constructing a method for automatically creating CTI datasets, collecting a dataset following the proposed method,

Proceedings of the 1st International Conference on NLP & AI for Cyber Security, pages 119–126
July 29–30, 2024.

and evaluating machine learning models to validate the built dataset. As a result, we present a dataset named WAVE-27K, including unstructured texts with Tactical Techniques and Procedures (TTP) information. WAVE-27K contains approximately 28,000 CTI descriptions associated with seven tactics and 27 different MITRE techniques. To the best of our knowledge, WAVE-27K is the largest dataset available in the CTI state of the art.

This paper is organized as follows: Section 2 provides a related work review, offering context for the research. Section 3 details our methodology, including the dataset-building process. Section 4 describes the experiments, defining details regarding the models and the metrics used for model evaluation. Finally, Section 5 presents the results, and Section 6 contains our findings and future research.

2 Background

There is two groups from the TTP pattern extraction literature differentiated by their goals. The first group extracts information from unstructured sources and transforms it into structured data. This process implies detecting different entities in a free text sample, and then identifying their relationships to generate knowledge graphs. The second group focuses on classification techniques, addressing CTI unstructured data as a classification problem. The primary goal of this group is to detect patterns within the unstructured text and categorize them according to known cyberattack techniques, enabling the identification and classification of relevant information. In this section, we detail significant results presented in the state of the art related to both groups.

2.1 Information extraction

Noor et al. (2019) implemented a three-step approach to extract information from unstructured data. The first step focused on collecting data from CTI sources. Then, they analyzed the data using a semantic search method to identify techniques, procedures, and observables. Finally, they developed a model to predict the cyber threat actor group based on the extracted information. Their study involved collecting 327 unstructured reports collected from 2012 to 2018, related to 36 threat groups. Finally, they evaluated Naive Bayes, k-nearest neighbors, Decision tree, Random Forest, and Deep Learning Neural Network (DLNN) using this dataset, with the DLNN model demonstrating the highest effectiveness at 94% accuracy.

Jo et al. (2022) proposed a BERT-based model to extract entities from unstructured CTI data. Their approach integrated BERT (Devlin et al., 2018) and BiLSTM layers, explicitly focusing on recognizing ransomware information. Additionally, the authors built a manually annotated dataset that includes 6791 entities and 4323 relations. The authors reported that the BERT model achieved an F1-score of 97.2% for the entity recognition task.

Later, Siracusano et al. (2023) presented a method employing the GPT-3.5-Turbo prompt[1] to detect entities and relationships within CTI data. They transform this information into a Structured Threat Information Expression (STIX)[2] bundle, enabling easy comparison with existing research. This study focused on identifying malware and built a dataset including 204 publicly available reports over 2022.

Recently, Wang et al. (2024) presented the construction of a method called knowledge based Cyber Threat Intelligence Entity and Relation Extraction (KnowCTI). The authors addressed the entity extraction as a tagging task and relation extraction task as a classification task. They collected a total of 53713 samples as base knowledge. Then, they collected a second dataset for the entity extraction experiments. The second dataset contains 8872 instances and 28347 entities. Finally, the authors reported F1-scores of 90.16% for the entity recognition task and 81.83% for the relation extraction task

2.2 Classification techniques

Introducing a new perspective, Legoy et al. (2020) approached CTI information as a classification task aiming to identify MITRE ATT&CK tactics and techniques[3]. They compared TF-IDF weighting factors proposed by Christopher et al. (2008) against the Word2Vec model in the pre-processing phase. In the classification process, the authors evaluated both binary relevance presented by Luaces et al. (2012) and multi-label approaches. Their dataset comprised 1490 reports related to MITRE attacks and tactics. Finally, they found that models using TF-IDF weighting factors outperformed those using Word2Vec. Specifically, the AdaBoost Decision Tree model achieved a 61.30% F0.5 score for the multi-label approach, while Gradient T

[1] GPT-3-5-turbo Homepage
[2] Stix Homepage
[3] MITRE ATT&CK Homepage

Boosting attained a 65.04% F0.5 score for the binary relevance approach. The authors released a tool called Reports Classification by Adversarial Tactics and Techniques (rcATT) using the method proposed, and the data used to train and test the method as well[4].

Expanding on earlier work, Mendsaikhan et al. (2021) evaluated the efficacy of identifying MITRE attacks through a multi-label approach using various models, such as the fine-tuned BERT model, Multi-label k-Nearest Neighbors (Zhang and Zhou, 2005) (MlkNN), and LabelPowerset (Tsoumakas and Vlahavas, 2007). The authors performed their analysis using three publicly available datasets for training: the Threat Report ATT&CK Mapper[5] (TRAM) dataset, it includes 1482 samples describing an event linked to 80 different MITRE techniques; Katos et al. (2019) presented the second dataset, which is built using the data release in an ENISA report with data from 2018 to 2019. After preprocessing the reports, the dataset incorporates 7642 samples associated with 50 techniques and nine tactics; Finally, the authors used the dataset presented by Legoy et al. (2020) previously described in this Section. The results showed that BERT achieved the highest performance, achieving a 78.01% F1 score and following the LabelPowerset method with Multilayer Perceptron (MLP) with a 74.70% F1 score.

Later, Orbinato et al. (2022) used several machine learning techniques for the classification task on a dataset created from information extracted from MITRE ATT&CK and Attack Pattern Enumerations and Classifications (CAPEC) sources. Their dataset [6] contains 12945 samples with descriptions of threat actors and their malware campaigns, the samples are related to 14 tactics and 188 distinct techniques. Additionally, they included the TRAM dataset in their evaluation. The authors used models such as Linear Regression (LR), Support Vector Machine (SVM), and SecureBERT (Aghaei et al., 2022) on both datasets. Finally, SecureBERT achieved the highest F1-score value of 72.50% in their dataset, while SVM achieved the highest F1-Score of 60.90% in the TRAM dataset.

Alves et al. (2022) analyzed 11 different combinations of hyperparameters on Transformer models, including RoBERTa (Liu et al., 2019), BERT (Devlin et al., 2018), SecRoBERTa (Liu et al., 2019), and SecBERT (Aghaei et al., 2022). Their dataset included 9909 sentences corresponding to 253 techniques, gathered from procedure examples within the MITRE ATT&CK source. The authors used accuracy to assess the performance of the models, showing RoBERTa as the model that achieved the highest performance with an accuracy of 82.64% on the testing dataset.

Recently, Branescu et al. (2024) presented a new dataset called CVE2ATT, the authors used MITRE ATT&CK tactic information as labels[7]. Following an automated process, the dataset extracts data from the ENISA register 2018 to 2019, including 9985 samples related to 14 tactics. The authors evaluated the data using several models, including CyBERT (Ranade et al., 2021), SecBERT, TARS (Halder et al., 2020), and GPT-4, in a multilabel tactic classification task. Their results revealed that SecRoBERTa achieved the highest performance with a 78.88% F1 score, closely followed by SecBERT at 78.77%.

Regarding the two groups reviewed in this Section, we have observed on the one hand, that the Information Extraction group focused on generating structured information from unstructured sources, usually representing it as a knowledge graph containing entities and relations of incident-related data. However, the building process of this kind of dataset compromises significant challenges. Despite the utility of this information extraction process in daily CTI tasks, its construction requires expertise and implies a complex process. On the other hand, the classification technique group intends to standardize the labels using the MITRE matrix, allowing the comparison between different implementations and enabling the integration of the public datasets into the training process. This standardization also allows us to work on automating the construction process of the dataset using the flow of data supplied by the CTI sources. Therefore, in this work, we have decided to focus on developing an automated construction method capable of collecting data from multiple sources to create a dataset and keep updating the dataset in sample size and class diversity.

3 Methodology

In our data collection process, we employed four primary sources that have been widely used in pre-

[4] rcATT GitHub Repository
[5] TRAM GitHub Repository
[6] cti-to-mitre-with-nlp GitHub Repository
[7] CVE2ATT GitHub Repository

vious research to construct CTI datasets; we selected these sources due to the facility to cross-reference their samples as has been proposed previously in the state of the art (Hemberg et al., 2022, 2020; Rantos et al., 2020; Branescu et al., 2024).

The first source is the MITRE ATT&CK framework, used as the foundation for standardizing datasets within the classification group (Legoy et al., 2020; Mendsaikhan et al., 2021; Orbinato et al., 2022; Alves et al., 2022; Branescu et al., 2024). This framework provides information on the tactics and techniques employed by attackers and information about campaigns, the associated threat actor groups, the tools and software used in the attacks, and potential mitigation strategies as well.

Another source employed in our data collection process is CAPEC, which assists in understanding how adversaries exploit software vulnerabilities. This list of attack patterns includes several columns providing information such as the attack pattern name, description, likelihood, related weaknesses, execution flow, severity, and additional relevant data.

Taking advantage of the information provided by CAPEC regarding software weaknesses, our third data source is the Common Weakness Enumeration (CWE), containing a list developed by the community of software and hardware vulnerabilities. This list traces each weakness with background details, affected technologies, consequences, impacted architectures, and observable examples.

Finally, the fourth data source is the Common Vulnerabilities and Exposures (CVE) repository, which lists information on known vulnerabilities. Each CVE entry includes a description of the vulnerability, its complexity, and its impact on confidentiality, integrity, and software availability.

We performed a complete review of the fields to identify potential references to external sources for each source. Some sources, such as MITRE and CAPEC, contain fields that directly present external references. In these cases, the external reference field within an entry was analyzed to verify if it included data from the selected sources. In the case of the CWE source, we analyzed the "observed example" field, which contains information about reported vulnerabilities. Using the vulnerability ID, we linked the information to the CVE source. Subsequently, all references were evaluated to determine if the target ID in the origin source was included in the data of the target source. The next phase involved matching the extracted IDs to establish new relationships and creating those relationships in an STIX format.

This approach aims to enhance data completeness by adding information from diverse views offered by different sources. As previously mentioned, data collection involves establishing new connections by mapping external references. Specifically, we used the following fields: MITRE ATT&CK external references to associate with CAPEC IDs, CAPEC external references to correlate with CWE IDs, and CVE weaknesses to align with CWE IDs. CWE plays a central role in this process, as depicted in Figure 1.

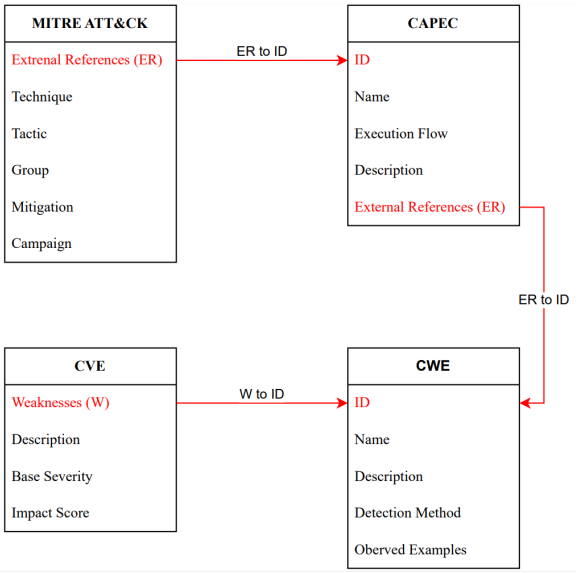

Figure 1: Sources integration process. Red highlighting represents the fields that have provided external links to relate information with other sources

The strength of this methodology lies in its automatic construction which enables updates and the addition of new samples for entry as well as the creation of new labels if new MITRE techniques are reported, besides the possibility of standardizing the dataset construction and normalizing the labels in the state of the art datasets. However, a potential limitation is the coverage of the dataset since there is limited control over the class balance within the dataset. Moreover, this approach suggests that the samples collected represent the prevailing trends and patterns observed in cyber attacks, providing valuable insights into real-world threat scenarios.

As a result of this methodology, we have created a comprehensive superset called Weakness, Attack, Vulnerabilities, and Events (WAVE). This superset

contains all the information downloaded from the sources, as well as all the relationships established through the external reference matching process. With this superset, we can link information from vulnerabilities (CVE) to MITRE ATT&CK techniques and even MITRE ATT&CK mitigations.

3.1 WAVE-27K dataset building method

We created a subset using the descriptions of vulnerabilities from the CVE source to validate the use and quality of the information contained in the superset WAVE. We selected the CVE description since it contains information about vulnerabilities written as unstructured text and it has been used in other research (Katos et al., 2019). This subset contains the CVE description with their related tactics and techniques; this subset is called WAVE-27K. The data was retrieved in the last quarter of 2023, collecting 27801 samples, where 22539 of them are associated with a single technique and the remaining 5262 samples are linked to two or more techniques. WAVE-27K contains 27 distinct labels of MITRE techniques.

As a result, we present the largest dataset compared to those in the state of the art, which also contains the largest number of samples per class, as shown in Table 1. Besides providing a larger number of samples per technique, WAVE-27K contains a more detailed description of the CTI event.

Dataset	Samples	Tactics	Techniques	AVG samples / techniques	AVG words in description
CTI_NLP	12945	14	188	68	15
TRAM	1482	14	80	18	28
ENISA	7642	9	50	1465	45
WAVE-27K	**27801**	7	27	1830	45

Table 1: Datasets description and distribution, comparison between public datasets and WAVE-27K

4 Experiments

We use different models to validate the data and establish a baseline for our dataset. Taking into account the results of Mendsaikhan et al. (2021); Orbinato et al. (2022); Alves et al. (2022); Branescu et al. (2024), which highlighted the efficacy of BERT models, we decided to use BERT-based models for our experiments. Specifically, we implemented BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), SecBERT (Aghaei et al., 2022), secRoberta (Liu et al., 2019), CyBERT (Ranade et al., 2021). To assess the performance of the models, we used the total of the data available into WAVE-27K and split the data on 80-20, assigning 80% of the data to the training set and 20% to the test set.

In the second experiment, we evaluate the performance of the best BERT-based model identified in the first experiment across the publicly available dataset presented in Section 2. Specifically, we used the CTI and TRAM datasets, each containing single output samples relevant to cybersecurity threats, and the WAVE-27K single output samples that comprise 22539 samples. The model was trained separately on each dataset using the configuration 80-20 to divide the samples, generating three different models. Finally, each model was tested into the test set of their corresponding dataset.

In addition to the above experiments, we conducted a comparative analysis among the datasets. This experiment presented a challenge as we observed variations in the subsets of MITRE Techniques used as labels across datasets despite the MITRE matrix acting as a shared set of labels. Thus, we focused on assessing the common elements shared between our dataset and publicly available datasets. Using the WAVE-27K dataset as a reference point, we observed that the CTI, TRAM, and ENISA datasets have limited overlap in labels, as shown in Table 2. Specifically, the CTI, TRAM, and ENISA datasets incorporate only 12, 9, and 8 labels that overlap with WAVE-27K, respectively. This indicates a relatively small intersection of shared labels between WAVE-27K and these datasets, suggesting differences in the types of threats or techniques covered by each dataset.

These differences between the labels in the datasets may emerge from variations in the methodologies employed in the dataset construction process. While some datasets are built by extracting information from the MITRE matrix using NLP algorithms, others include manually annotated CVE descriptions. These diverse construction processes restrict the data to specific types of information and introduce complexities in direct data comparison, making an assessment more complicated. We consider the overlapped classes between WAVE-27K and the other datasets to allow comparison. This approach aims to provide a representative measure of data quality relative to existing datasets in the literature.

Label	WAVE-27K	CTI	ENISA	TRAM
T1021	✓	✓		✓
T1072	✓	✓		
T1505	✓	✓	✓	✓
T1543	✓	✓	✓	✓
T1546	✓	✓	✓	✓
T1547	✓	✓	✓	✓
T1550	✓	✓		
T1552	✓	✓	✓	
T1553	✓	✓	✓	✓
T1562	✓	✓	✓	✓
T1566	✓	✓		
T1574	✓	✓	✓	✓

Table 2: Intersection of labels in the dataset using as reference WAVE-27K

4.1 Metrics

The F1 score is a widely used metric for assessing binary and multi-class classification tasks, providing a balanced assessment of the ability of the model to classify both positive and negative instances by considering both precision and recall. In this specific context, since we are evaluating a model trained in WAVE-27K that includes multi-label and multi-output samples, we selected the micro-average F1 score. The micro-average F1 score provides consistency across all classes by considering each instance equally, regardless of its class, providing unbiased results in multi-class and multi-output settings.

We selected the micro-average F1 score as it is best suited for evaluating our dataset. However, we are aware that two datasets in the comparison include only single-output samples, which could lead to this metric penalizing them. To address this, in addition to evaluated directly using single output samples in the WAVE-27K (experiment 2), we included accuracy in our evaluation as well, as it is a commonly used performance metric in the state of the art for TTPs related tasks (Noor et al., 2019; Alves et al., 2022).

5 Results

For the first experiment, after training the five proposed BERT-based models in Section 4 to establish a baseline for comparison on WAVE-27K, the results indicate that SecRoberta achieved the highest performance with a 77.52% F1 score and a 83.51% accuracy, followed by BERT with 77.31% F1 score and 83.29% accuracy, as shown Table 3.

In the second experiment, we used secRoBERTa as it had the highest performance in the previous experiment. After the training phase, the secRoBERTa model from the WAVE-27K dataset

Model	Accuracy (%)	F1 Score (%)
BERT	83.29	77.31
CyBERT	81.13	73.88
RoBERTa	83.67	76.91
SecBERT	82.83	76.12
secRoBERTa	**83.51**	**77.52**

Table 3: Experiment 1. Performance metrics of different models using the WAVE-27K dataset

achieved an accuracy of 91.39% on the test set as shown in Table 4, demonstrating the highest performance among the single output models tested.

Experiment	Dataset name	Classes	N. Test Samples	ACC
Complete Test Set	CTI	188	1942	90.73
	TRAM	80	221	83.26
	WAVE-27K	27	4507	**91.39**

Table 4: Experiment 2. Detailed performance of the models trained using single output datasets

Regarding the comparison between overlapped classes of WAVE-27K and the public datasets, the results demonstrate that the model trained with WAVE-27K outperforms those trained with the CTI, TRAM, and ENISA datasets, achieving Micro F1-scores of 96.46%, 95.50%, and 92.15%, respectively, as shown in Table 5. The last result presents a quantitative insight into the proficiency of one model across various datasets, highlighting its robust performance in classifying cybersecurity-related data. However, as we described in Section 4, the discrepancy of labels across datasets prevents direct comparison. Therefore, we rely on these results to validate that the data within WAVE-27K includes pertinent information for incident classification, demonstrating a sufficient level of quality for machine learning model training.

Experiment	Dataset name	Classes	N. Test Samples	ACC	F1 Micro
CTI - WAVE-27K	CTI	12	266	74.43	74.43
	WAVE-27K	12	1363	**91.25**	**96.46**
TRAM -WAVE-27K	TRAM	9	37	59.46	19.04
	WAVE-27K	9	1327	**91.61**	**96.50**
ENISA - WAVE-27K	ENISA	8	431	**80.22**	83.48
	WAVE-27K	8	1177	79.86	**92.15**

Table 5: Comparison of available datasets with WAVE-27K, results using only the common classes by each public dataset and WAVE-27K.

6 Conclusions and future work

This paper presents an automated approach to link several CTI sources through the mapping of external references, resulting in a more complete dataset. The previous is due to the inclusion of insights from

different four sources. Our method facilitates the automatic construction of datasets, allowing for updates and the inclusion of new samples and labels.

To assess the data collection method, we used a subset of features extracted from the consolidation of the four sources, namely Weakness, Attack, Vulnerabilities, and Events 27K dataset (WAVE-27K). The WAVE-27K includes the CVE description as a free-text sample and the corresponding MITRE techniques related to the description. While one potential limitation derives from the coverage of the dataset, with limited control over class balance, this approach suggests that the collected samples reflect prevalent trends and patterns in cyberattacks, providing valuable insights into real-world threat scenarios.

Wave27K contains 27801 samples, where 22539 of them are associated with a single technique and the remaining 5262 samples are linked to two or more techniques. WAVE-27K includes 27 distinct labels of MITRE techniques.

We trained five BERT-based models in the evaluation process, finding that SecRoBERTa reaches the highest performance with a 77.52% F1 score. Subsequently, the model trained with the WAVE-27K dataset achieved a 91.39% accuracy in the single output test. Finally, in the comparison of overlapping classes, our model using the WAVE-27K dataset outperforms others, achieving an F1 score of up to 96.46%. These findings demonstrate that the data within WAVE-27K contains relevant information for incident classification. The results show that the proposed method effectively built a dataset with a level of quality sufficient to train a machine-learning model.

In future research, we aim to explore additional machine learning models that were not considered in this study. Additionally, we plan to study the possibility of training specialized models for each class to assess the effectiveness of classification in such a scenario. Furthermore, we will explore a cascading classification approach, initially classifying tactics followed by a technique classification using a stacking method. This approach will allow us to determine if hierarchical classification enhances overall performance.

We aim to face the challenge of processing longer unstructured text and associating it with relevant tags as well. Incorporating state of the art models into this context will improve the capabilities of attack classification systems into a more realistic scenario. This advancement will facilitate the automatic generation of alerts from free and unstructured text, enhancing the efficiency of threat detection and response mechanisms.

Acknowledgment

This work has been partially supported by the European Union's Horizon Europe Framework under the project ATLANTIS (Grant Agreement No. 01073909).

References

Ehsan Aghaei, Xi Niu, Waseem Shadid, and Ehab Al-Shaer. 2022. Language model for text analytic in cybersecurity. *arXiv preprint arXiv:2204.02685*.

Paulo M.M.R. Alves, Geraldo P.R. Filho, and Vinicius P. Goncalves. 2022. Leveraging bert's power to classify ttp from unstructured text. *2022 Workshop on Communication Networks and Power Systems, WCNPS 2022*.

Ioana Branescu, Octavian Grigorescu, and Mihai Dascalu. 2024. Automated mapping of common vulnerabilities and exposures to mitre att&ck tactics. *Information*, 15:214.

D Christopher, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. Scoring term weighting and the vector space model. *Introduction to information retrieval*, 100:2–4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shota Fujii, Nobutaka Kawaguchi, Tomohiro Shigemoto, and Toshihiro Yamauchi. 2022. Cyner: Information extraction from unstructured text of cti sources with noncontextual iocs. In *International Workshop on Security*, pages 85–104. Springer.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. Task-aware representation of sentences for generic text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213.

Erik Hemberg, Jonathan Kelly, Michal Shlapentokh-Rothman, Bryn Reinstadler, Katherine Xu, Nick Rutar, and Una-May O'Reilly. 2020. Linking threat tactics, techniques, and patterns with defensive weaknesses, vulnerabilities and affected platform configurations for cyber hunting. *arXiv preprint arXiv:2010.00533*.

Erik Hemberg, Ashwin Srinivasan, Nick Rutar, and Una-May O'Reilly. 2022. Sourcing language models and text information for inferring cyber threat, vulnerability and mitigation relationships. In *AI4Cyber:*

AI-enabled Cybersecurity Analytics and Deployable Defense workshop.

Hyeonseong Jo, Yongjae Lee, and Seungwon Shin. 2022. Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text. *Computers & Security*, 120:102763.

Vasilis Katos, Shahin Rostami, Panagiotis Bellonias, Nigel Davies, Agnieszka Kleszcz, Shamal Faily, et al. 2019. State of vulnerabilities 2018/2019: analysis of events in the life of vulnerabilities. *Report/Study*.

Valentine Legoy, Marco Caselli, Christin Seifert, and Andreas Peter. 2020. Automated retrieval of att&ck tactics and techniques for cyber threat reports. *arXiv preprint arXiv:2004.14322*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Oscar Luaces, Jorge Díez, José Barranquero, Juan José del Coz, and Antonio Bahamonde. 2012. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, 1:303–313.

Otgonpurev Mendsaikhan, Hirokazu Hasegawa, Yukiko Yamaguchi, and Hajime Shimada. 2021. Automatic mapping of threat information to adversary techniques using different datasets. *International Journal on Advances in Security Volume 14, Number 1 & 2, 2021*.

Umara Noor, Zahid Anwar, Tehmina Amjad, and Kim Kwang Raymond Choo. 2019. A machine learning-based fintech cyber threat attribution framework using high-level indicators of compromise. *Future Generation Computer Systems*, 96:227–242.

Vittorio Orbinato, Mariarosaria Barbaraci, Roberto Natella, and Domenico Cotroneo. 2022. Automatic mapping of unstructured cyber threat intelligence: An experimental study:(practical experience report). In *2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*, pages 181–192. IEEE.

Priyanka Ranade, Aritran Piplai, Anupam Joshi, and Tim Finin. 2021. Cybert: Contextualized embeddings for the cybersecurity domain. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3334–3342. IEEE.

Konstantinos Rantos, Arnolnt Spyros, Alexandros Papanikolaou, Antonios Kritsas, Christos Ilioudis, and Vasilios Katos. 2020. Interoperability challenges in the cybersecurity information sharing ecosystem. *Computers*, 9.

Giuseppe Siracusano, Davide Sanvito, Roberto Gonzalez, Manikantan Srinivasan, Sivakaman Kamatchi, Wataru Takahashi, Masaru Kawakita, Takahiro Kakumaru, and Roberto Bifulco. 2023. Time for action: Automated analysis of cyber threat intelligence in the wild. *arXiv preprint arXiv:2307.10214*.

Grigorios Tsoumakas and Ioannis Vlahavas. 2007. Random k-labelsets: An ensemble method for multilabel classification. In *European conference on machine learning*, pages 406–417. Springer.

Gaosheng Wang, Peipei Liu, Jintao Huang, Haoyu Bin, Xi Wang, and Hongsong Zhu. 2024. Knowcti: Knowledge-based cyber threat intelligence entity and relation extraction. *Computers & Security*, 141:103824.

Min-Ling Zhang and Zhi-Hua Zhou. 2005. A k-nearest neighbor based algorithm for multi-label classification. In *2005 IEEE international conference on granular computing*, volume 2, pages 718–721. IEEE.