A Dataset for the Fine-tuning of LLM for the NER Task in the Cyber Security Domain

Stefano Silvestri[1,*,†], Giuseppe Felice Russo[1,†], Giuseppe Tricomi[1,†] and Mario Ciampi[1]

[1]Institute for High Performance Computing and Networking, National Research Council of Italy (ICAR-CNR), Via Pietro Castellino 111, Naples, 80131, Italy

Abstract

The increasing complexity of cyber threats necessitates robust cyber security measures. Effective threat detection and mitigation depend on Cyber Threat Intelligence, which includes structured and unstructured data critical for proactive defense strategies. While databases like the NVD and ExploitDB offer structured security information, a significant amount of vital intelligence initially appears in unstructured formats, such as blogs, mailing lists, and news sites. Extracting meaningful information from these sources is particularly challenging in cyber security, requiring specialized Named Entity Recognition (NER) tools to identify domain-specific entities. This paper presents a NER dataset obtained by merging two cyber security domain datasets, CyNER and APTNER, creating a unified resource that enhances NER model training. Experimental results with advanced NER models show significant performance gains, underscoring the value of the proposed dataset in advancing cyber security practices, and highlighting the needs of such kind of resources.

Keywords

Cyber Threat Intelligence, Named Entity Recognition, Cybersecurity, Large Language Model, NLP

1. Introduction

In today's interconnected digital landscape, cyber security remains a critical concern, due to the proliferation of sophisticated cyber threats and vulnerabilities across global networks. The timely identification and mitigation of these threats rely heavily on Cyber Threat Intelligence (CTI), which encompasses the structured and unstructured information essential for preemptive defense strategies. While structured databases like NVD [1] or ExploitDB [2] provide valuable and well-defined security information, a significant amount of critical intelligence emerges initially in unstructured formats, often in natural language, such as blogs, mailing lists, and news sites [3]. These sources contain valuable and constantly updated information about cyber threats, vulnerabilities, risks, mitigation strategies, but their nature, due to the intrinsic complexity of natural language, often delays the classification and integration of new information into structured databases. The challenge of extracting actionable intelligence from unstructured sources is exacerbated in the cyber security, where domain-specific entities require specialized Named Entity Recognition (NER) tools, often integrated with ontologies and Knowledge Bases [4, 5], because general-purpose NER tools trained on broad corpora often fail to capture the specialized terminology and entity types in cyber security reports [6]. As a result, there is a need for domain-specific datasets that facilitate the training and evaluation of NLP models capable of extracting cyber threat indicators with high accuracy and relevance [7, 8]. In recent years, efforts such as the development of the APTNER dataset [8] have aimed to address this need, providing a substantial corpus for training and evaluating NER models in the CTI domain. However, existing datasets often lack the scale and diversity necessary to comprehensively cover the breadth of cyber threat scenarios encountered in practice. The advent of Large Language Models (LLMs) has introduced

Discovery Science - Late Breaking Contributions 2024

*Corresponding author.

†These authors contributed equally.

stefano.silvestri@icar.cnr.it (S. Silvestri); giuseppefelice.russo@icar.cnr.it (G. F. Russo); giuseppe.tricomi@icar.cnr.it (G. Tricomi); mario.ciampi@icar.cnr.it (M. Ciampi)

0000-0002-9890-8409 (S. Silvestri); 0009-0001-2090-9647 (G. F. Russo); 0000-0003-3837-8730 (G. Tricomi); 0000-0002-7286-6212 (M. Ciampi)

Table 1
Mapping Scheme of the Dataset

APTNER entity types	**Mapped entity** types
APT (threat participants), EMAIL (malicious mailbox), MAL (malware), ACT (attack action)	Malware
SECTEAM (security team)	Secteam
IDTY (authentication identity)	Organization
OS, TOOL	System
LOC (location), TIME, PROT (protocol), ENCR (encryption algorithm)	Other
DOM (domain), IP, URL, FILE, MD5, SHA1, SHA2	Indicator
VULNAME (vulnerability name), VULID (vulnerability number)	Vulnerability

new perspectives in the NLP domain, significantly improving, among the others, the understanding and extracting domain-specific entities from complex unstructured texts. These models offer the potential to bridge the gap between structured and unstructured CTI sources, enabling more timely and accurate threat detection and response and allowing the development of effective NLP-based cyber security tools and methods [9, 10, 11]. On the other hand, LLMs needs to be fine-tuned on the specific task and domain, but, as mentioned above, in cyber security doimain there is a lack of effective annotated NER resources. Therefore, this paper proposes to merge two prominent cyber security NER datasets: CyNER [7] and APTNER [8]. CyNER aggregates a wide array of openCTI from diverse sources, and complements APTNER's focus on structured NER tasks. By merging these datasets, we aim to establish a more comprehensive resource that enriches the availability of NER datasets for the research community, and also supports enhanced threat detection, incident response, and vulnerability mitigation strategies.

2. Dataset

APTNER [8] and CyNER [7], two datasets for NER in the cyber security domain, were combined to create a new merged dataset. Each source has different definitions and classifications for entities and it is required a mapping scheme for these different labels. To balance efficiency and simplicity, while improving coverage, the decision was made to utilize the CyNER scheme and extend it with one entity type from APTNER (i.e., *Secteam*). The APTNER annotation scheme comprises a larger label set that could be remapped into the final annotation scheme, and standardizing entity types and annotations makes it easier to create strong NER models for improved cyber security analysis and response. Moreover, the labels used in APTNER include several subtypes of CyNER labels, so a natural mapping approach and association were made using the mapping scheme shown in Table 1, aggregating some subtypes into a single type. The combined cyber security NER dataset's annotation scheme resulted in seven labels, with the following meanings:

1. *Indicator* represents information useful to identify the resource compromised or the technology affected by the attack.
2. *Malware* represents all possible threat elements extracted from the corpora, such as action, actors, software, techniques, and so on.
3. *Secteam* represents the group announcing the vulnerability identified.
4. *System* represents operating system, software, and hardware.
5. *Vulnerability* represents both CVE ID and mention of exploits.
6. *Organization* represents companies, organizations, institutions, brands, and others.
7. *Other* includes the additional and generic entity types that are not annotated in one of the considered dataset and cannot be mapped in a specific category of the other one.

As result, we obtained a merged dataset, further split in a training set and a test set (approximately 70% and 30%), whose statistics and distributions of the various entity types are summarized in Table 2.

3. Experimental Evaluation

The assessment of the proposed augmented dataset presented in this study is performed by using it in the NER fine-tuning of LLMs, comparing the obtained results with the performances obtained

Table 2
Statistics of the dataset (split in train and test) and distribution of Entities by category

Dataset	Sentences	Tokens	Total Entities	Malware	Indicator	System	Secteam	Organization	Vulnerability
Train	9,556	250,734	20,274	9,160	3,047	2,876	1,067	3,180	944
Test	4,045	97,045	6,416	2,333	1,057	2,057	278	594	97

Table 3
Performances of each LLM on the merged dataset (M), the APTNER dataset (A) and the CyNER dataset (C)

Model	Precision (M)	Recall (M)	F1-score (M)	Precision (A)	Recall(A)	F1-score A)	Precision (C)	Recall (C)	F1-score (C)
SecBERT	49.44	51.14	50.28	42.44	50.15	45.97	50.15	36.82	42.46
SecureBERT	62.25	61.89	62.07	54.97	61.56	58.08	**63.84**	**48.33**	**55.01**
BERT-base-cased	62.00	60.89	61.44	56.06	**64.62**	**60.03**	62.83	38.77	47.95
RoBERTa-base	**63.68**	**64.5**	**64.09**	**56.73**	60.79	58.69	61.05	44.28	51.33

using the original two datasets. We fine-tuned two NER models tailored to the cyber security domain, namely SecBert[1] and SecureBERT [12], both available on the Hugging Face LLM repository. These models, respectively built upon BERT [13] and RoBERTa [14] architectures, have been pre-trained on large corpora in the cyber security domain, demonstrating that they are both able to provide improved results when fine-tuned for NLP tasks in the same domain. In addition to these models, we also trained two baseline models, *BERT-base-cased* and *RoBERTa-base*, using them as benchmarks for a comparison against the specialized models. The evaluation is based on standard NER metrics (P, R, and F1) calculated at the token level. The obtained results, summarized in Table 3, provided insights into each model's effectiveness and generalization capability on the cyber security NER task. When comparing the performance across all datasets, it is evident that the merged dataset significantly improves the NER model results, showcasing its ability to enhance the performance through richer and more diverse entity coverage, also when used with general-domain LLMs. On the other hand, the datasets (including the original ones) have some very unbalanced classes, and this can limit their performances, as well as their generalization capabilities, causing in some experiments lower performances, compared with the expected ones.

The merged dataset, including the documentation on the annotation scheme, and the fine-tuned NER models are publicly available on the SoBigData research infrastructure[2].

4. Conclusions and Future Works

This work presented a cyber security NER dataset, obtained by combining APTNER and CyNER datasets, with the purposes of addressing the scarcity of open cyber security NER corpora and improving the performances of the original ones. The merged dataset standardizes and harmonizes entity types across different sources, providing a comprehensive and diverse set of annotations. Our experiments demonstrated that the proposed dataset significantly enhances model performance, highlighting its ability to improve the recognition capabilities of NER models in the cyber security domain.

Future work could focus on further expanding the dataset by integrating additional cyber security corpora to cover a wider range of entities and scenarios, as well as to reduce the unbalance of the dataset, also leveraging augmenting techniques or semi-supervised annotation approaches [15]. Additionally, exploring transfer learning techniques to apply the knowledge gained from this dataset to other related tasks in cyber security, such as threat detection and incident response, could be highly beneficial. Finally, it would be valuable to investigate the impact of different annotation schemes and entity definitions on model performance, to refine further and optimize the dataset.

[1] https://github.com/jackaduma/SecBERT
[2] https://data.d4science.org/ctlg/ResourceCatalogue/cybersecurity_ner_securebert_model
https://data.d4science.org/ctlg/ResourceCatalogue/cybersecurity_ner_roberta-base_model
https://data.d4science.org/ctlg/ResourceCatalogue/cybersecurity_ner_bert-base-cased_model
https://data.d4science.org/ctlg/ResourceCatalogue/cybersecurity_ner_dataset

Acknowledgments

This work is supported by the European Union - NextGenerationEU - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013 - Avviso n. 3264 del 28/12/2021. We thank Simona Sada and Giuseppe Trerotola for the administrative and technical support provided.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] T. B. Robert Byers, Chris Turner, National vulnerability database, national institute of standards and technology, https://nvd.nist.gov (accessed on 1/9/2024), 2024.

[2] Offsec, Exploit data base, https://www.exploit-db.com (accessed on 1/9/2024), 2024.

[3] F. Yi, B. Jiang, L. Wang, J. Wu, Cybersecurity named entity recognition using multimodal ensemble learning, IEEE Access 8 (2020) 63214–63224. doi:10.1109/ACCESS.2020.2985625.

[4] T.-M. Georgescu, Natural language processing model for automatic analysis of cybersecurity-related documents, Symmetry 12 (2020). doi:10.3390/sym12030354.

[5] T. Satyapanich, F. Ferraro, T. Finin, Casie: Extracting cybersecurity event information from text, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 8749–8757.

[6] I. Deliu, C. Leichter, K. Franke, Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks, in: 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3648–3656. doi:10.1109/BigData.2017.8258359.

[7] M. T. Alam, D. Bhusal, Y. Park, N. Rastogi, CyNER: A Python library for cybersecurity named entity recognition, 2022. arXiv:2204.05754.

[8] X. Wang, S. He, Z. Xiong, X. Wei, Z. Jiang, S. Chen, J. Jiang, APTNER: A specific dataset for NER missions in cyber threat intelligence field, in: Proceedings of the 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2022, pp. 1233–1238.

[9] S. Silvestri, S. Islam, D. Amelin, G. Weiler, S. Papastergiou, M. Ciampi, Cyber threat assessment and management for securing healthcare ecosystems using natural language processing, International Journal of Information Security 23 (2024) 31–50. doi:10.1007/s10207-023-00769-w.

[10] S. Silvestri, S. Islam, S. Papastergiou, C. Tzagkarakis, M. Ciampi, A machine learning approach for the NLP-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem, Sensors 23 (2023). URL: https://www.mdpi.com/1424-8220/23/2/651. doi:10.3390/s23020651.

[11] N. Capodieci, C. Sanchez-Adames, J. Harris, U. Tatar, The impact of generative AI and LLMs on the cybersecurity profession, in: 2024 Systems and Information Engineering Design Symposium (SIEDS), 2024, pp. 448–453. doi:10.1109/SIEDS61124.2024.10534674.

[12] E. Aghaei, X. Niu, W. Shadid, E. Al-Shaer, SecureBERT: A domain-specific language model for cybersecurity, in: International Conference Security and Privacy in Communication Networks (SecureComm), Springer, Cham, 2023, pp. 39–56.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics NAACL-HLT 2019, ACL, Minneapolis, MN, USA, 2019, pp. 4171–4186. doi:10.18653/V1/N19-1423.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). arXiv:1907.11692.

[15] G. Aracri, A. Folino, S. Silvestri, Integrated use of KOS and deep learning for data set annotation in tourism domain, Journal of Documentation 79 (2023) 1440–1458. doi:10.1108/JD-02-2023-0019.