CEUR-WS.org/Vol-4121/Ital-IA_2025_paper_69.pdf

Towards explainability framework for cybersecurity domain: a case study using NER

Stefano Silvestri[1,*], Emanuele Damiano[1], Raffaele Guarasci[1] and Mario Ciampi[1]

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

Abstract

This work presents some activities performed within the SoBigData.it research project, with the purpose of addressing some critical challenges of Named Entity Recognition (NER) in the cybersecurity domain, by introducing a specific explainability framework. In detail, explainability techniques based on the Captum library have been experimented to deepen the analysis of model behaviour, revealing critical insights into feature importance and layer-wise contributions for entities like Malware, Vulnerability, and Indicators, showing a promising path for the adoption of similar approaches.

Keywords

Cybersecurity, Artificial Intelligence, Named Entity Recognition, Explainability

1. Introduction

In recent years, the rapid growth of cybersecurity threats and vulnerabilities across global networks has become an issue of crucial importance. Cyber Threat Intelligence (CTI), which encapsulates both structured and unstructured information necessary for proactive defense mechanisms [1], is the main tool to handle and mitigate these threats. While structured repositories like the National Vulnerability Database (NVD) and ExploitDB served as invaluable sources of clearly defined security insights, a significant portion of actionable intelligence initially surfaced in unstructured, natural language formats. Sources such as technical blogs, security forums, mailing lists, and online news platforms frequently disseminated emerging threat details before formal databases were updated [2], providing rich, dynamic information encompassing vulnerabilities, exploits, risks, and countermeasures. However, integrating such data into structured CTI repositories exposes potential risks due to prolonged time exposure [3].

Although the ability to effectively extract meaningful entities from unstructured texts, Named Entity Recognition (NER) is an established task in Natural Language Processing (NLP), applying it to the cybersecurity domain poses several critical issues because of domain-specific terminology and the lack of extensive structured annotated datasets. General-purpose NER models trained on generic corpora frequently failed to identify specialized entities within threat accurately reports [4]. Consequently, we identified a clear need for domain-focused datasets capable of training and evaluating NLP models to extract cyber threat indicators with greater accuracy and contextual relevance [5, 6]. To address this lack, several approaches have been proposed. For instance, the APTNER dataset [6] has proposed a set of entity types following the STIX 2.1 standard and compiling a substantial corpus for cyber-specific NER tasks. However, these datasets often lacked the diversity and breadth required to capture the wide-ranging threat scenarios encountered in real-world contexts. Similar limitations affected CySecNER and HackIntelNER [7, 8], where the scale or adaptability across various attack contexts remained insufficient. The emergence of increasingly refined and powerful Neural Language Models (NLMs) significantly transformed the NLP landscape, spreading across every domain, including applications in cyber security. Notable improvements in entity extraction capabilities with models such as BERT and its

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

[*] Corresponding author.

stefano.silvestri@icar.cnr.it (S. Silvestri); emanuele.damiano@icar.cnr.it (E. Damiano); raffaele.guarasci@icar.cnr.it (R. Guarasci); mario.ciampi@icar.cnr.it (M. Ciampi)

0000-0002-9890-8409 (S. Silvestri); 0000-0002-0106-8635 (R. Guarasci); 0000-0002-7286-6212 (M. Ciampi)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-07

cyber-specialized variants [9] have been proven. These models, trained on large-scale corpora, offered new opportunities to reconcile unstructured CTI data with structured outputs, enhancing the timeliness and accuracy of threat detection and facilitating the development of intelligent cyber defense tools [10, 11]. Although resources are available for general-purpose approaches, the situation is different for specific domains or low-resource languages. However, given the growing need for annotated and unannotated datasets, several strategies have been put in place to address for these shortcomings [12].

The research activity presented in the following of this paper has been developed under the tasks of the *SoBigData.it* research project [13], a research infrastructure that enhances interdisciplinary and innovative research on the multiple aspects of social complexity by combining data and model-driven approaches, supporting open science, and also pursuing some crucial objectives, such as the development of new datasets and methods available to the scientific community and the definition of trustworthy AI approaches. In detail, we first created a novel cybersecurity NER dataset, by merging and remapping the labels of CyNER and APTNER, used to train and test some LLMs for the NER task, comparing their performances and highlighting their pros and cons [14]. Finally, to better understand the behavior of NER models in the domain under investigation, as well as to improve the trustability of the models for their larger adoption and use in real-world contexts, we propose here a preliminary explainability analysis, investigating how and where some specific features impact on model predictions.

2. Related Works

Several datasets have been released in the last decade to support NER tasks in the cyber security domain. These resources vary significantly in scale, annotation strategies, and types of entities included. The early approaches were based on the combination of traditional machine learning with rule-based strategies. iGen [15] focused on eight entity types from technical blogs and reports; iACE [16] identified four classes of entities from open-source articles, and [17] used structured rules to tag entities from log files. In the cybersecurity domain, [2] presented a domain-specific NER corpus composed of 14,000 unstructured documents with 22 predefined entity types and over 7,000 labeled entities. [18] curated a malware-focused knowledge graph derived from open-source GitHub repositories, defining 14 unique entity types relevant to malware behavior. [19] developed a BiLSTM-CRF-based NER system trained on 160 threat reports, capturing 11 cybersecurity-related entity types. [20] introduced TIMiner, a hybrid framework blending rule-based and neural techniques for extracting six types of threat-related entities from a manually annotated dataset of 15,000 threat intelligence records. [21] tried to incorporate the BIO annotation scheme over 175,220 tokens; although public, the relatively limited size of this dataset restricts its generalization capability for broader threat modeling. Several works have adopted neural architectures with domain adaptations. [22] used BiLSTM with attention and feature-rich representations to extract 11 types of attack indicators from GitHub incident reports spanning 2008–2018. [23] developed the Chainsmith system using regular expressions and neural models for classifying six types of CTI indicators from over 14,000 texts. More recently, [24] explored fine-tuning a large language model for NER in the healthcare cybersecurity domain, using web-extracted news articles annotated via a semi-supervised procedure detailed in [25]. This approach demonstrates the growing efficacy of adapting LLMs to domain-specific NER tasks. A recent study proposed a hybrid approach that combines controlled resources with neural models, using the cybersecurity domain as a case study [26]. Finally, an example of a trustworthy AI framework for the cybersecurity domain has been presented by [27, 28], where a hybrid AI-enabled model that combines both linear regression and deep learning is used to prioritize vulnerabilities, also integrating the explainability and interpretability characteristics for explaining AI model's decision making and its inner working parameters.

3. Materials and Methods

To address the scarcity of open cyber security NER datasets, we combined APTNER and CyNER into a single, harmonized corpus. Each dataset originally differently defined entities, we standardized and

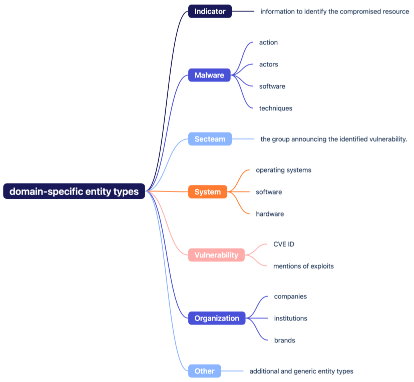

Figure 1: Cybersecurity entity types mapped in APTNER

aligned their annotations to support more robust and accurate NER model training. The resulting dataset adopts the CyNER annotation scheme, extended with an additional entity type from APTNER: *Secteam*, distinguishing cyber security teams from broader organizations. This brings the final label set to seven: *Malware, Indicator, System, Secteam, Organization, Vulnerability,* and *O*, using the BIO tagging format. To maintain generalization and reduce overfitting, we opted for a more compact schema. APTNER's richer label taxonomy was systematically mapped into this simpler structure, grouping related subtypes under broader categories like *Malware* and *Indicator*, ensuring both semantic clarity and model efficiency. The mapped entities are shown in Figure 1. The final combined cyber security NER dataset comprises 347,779 tokens and 13,601 labeled phrases, resulting in 37,684 annotated entities across six distinct entity types. More details, including label mappings, are available in [14] and the dataset is publicly available on the SoBigData repository[1]. Different NLMs, reported below, were used to evaluate the dataset, to benchmark their performances in the cybersecurity NER task when fine-tuned on our merged dataset [14]. The NER NLMs trained on the aforementioned dataset and considered for the purpose of this study are also made publicly available through the SoBigData repository[2] [3] [4].

- **SecureBERT** : built on the RoBERTa framework, is trained on large-scale cyber security text and has demonstrated improved performance on security-related NLP tasks [29].
- **BERT-base-cased**: a standard BERT model not specifically adapted for cyber security.
- **RoBERTa-base**: a robustly optimized variant of BERT.

Using deep learning models for NER in cybersecurity requires a rigorous explainability analysis to ensure transparency, trust, and validation by domain experts [30]. The Explainability focuses on the importance of input tokens for a specific model prediction. The aim is to clarify which words or text fragments influence the prediction in classifying a particular span as a particular entity. For this purpose, we adopted **Captum AI** [31], a PyTorch-based library, to interpret transformer-based models in cybersecurity NER tasks, which includes several attribution algorithms available in Captum suitable for text-based NER analysis, such as *Integrated Gradients* (IG), which approximates gradients along a path from baseline to input [30], *Layer Integrated Gradients* (LIG), for the attribution of output to specific layers, *Feature Ablation*, to evaluate the importance of the features by removing/masking them, *GradientShap*, which Combines IG with Shapley values , and *LIME* (Local surrogate model approximation [32]). Methods like IG and LIG are particularly relevant to our analysis. IG computes the importance of the features by integrating gradients along a straight line path from the input of the baseline to the actual input [30]. For *Baseline selection* of text in NER task, common choices include

[1] https://data.d4science.org/ctlg/ResourceCatalogue/cybersecurity_ner_dataset
[2] https://data.d4science.org/ctlg/ResourceCatalogue/cybersecurity_ner_securebert_model,
[3] https://data.d4science.org/ctlg/ResourceCatalogue/cybersecurity_ner_bert-base-cased_model,
[4] https://data.d4science.org/ctlg/ResourceCatalogue/cybersecurity_ner_roberta-base_model

padding tokens, unknown tokens, or zero-embedding, vectors [33]. These are crucial for meaningful attributions; on the other hand, *Token attribution* aggregates embedding dimension scores (sum/mean) for per-token importance. LIG extends IG to attribute output to specific network layers (for example, RoBERTa's embedding layer). For SecureBERT, the attribution to a layer l can be defined as:

$$\text{Attribution} = \int_{\alpha=0}^{1} \frac{\partial f(\mathbf{x}_0 + \alpha(\mathbf{x} - \mathbf{x}_0))}{\partial \mathbf{h}^{(l)}} d\alpha \tag{1}$$

where $\mathbf{h}^{(l)}$ is the output of layer l, $\mathbf{x}$ is the input, and $\mathbf{x}_0$ is the baseline input.

Feature Ablation is another perturbation-based method that assesses the importance of tokens by observing the changes in output when tokens are replaced with baseline values [34]. For NER, the change in score can be represented as: $\Delta\text{Score} = f(\mathbf{x}) - f(\mathbf{x}_{\text{ablated}})$, where $f(\mathbf{x}_{\text{ablated}})$ is the model's output when specific tokens in $\mathbf{x}$ are masked or replaced. Strategies for text attribution to contextualize attribution scores derived from such methods, like *Padding Token* (use model's padding token as baseline [33]), *Unknown Token* (use `[UNK]` token for out-of-vocabulary representation), *Zero Embedding* (creating baseline with zero vectors in embedding space), can serve as simple benchmarks to evaluate the effectiveness of more complex text attribution methods.

4. Results and Discussion

Applying the proposed explainability framework to SecureBERT, we identified the most influential tokens for recognizing key cybersecurity entities. Table 1 presents these tokens alongside their attribution scores and example contexts. For Malware, "*APT*" holds the highest attribution score (0.82), as seen in "*APT29 spearphishing campaign*." "*CVE-*" (0.79) is key for Vulnerability identification, exemplified by "*CVE-2023-1234 in Apache Log4j*." The token "*http://*" (0.75) is most influential for Indicator entities, such as "*Malicious URL http://phish.site*." For system entities, "*Windows*" (0.68) is significant, as in "*Windows registry keys modified*." Lastly, "*Microsoft*" (0.61) is the top token for Organization entities, illustrated by "*Microsoft security advisory*." These findings underscore the importance of specific technical terms and formats in the model's entity recognition process.

Table 1
Top influential tokens for cybersecurity entity recognition in SecureBERT.

Entity Type	Token	Attribution Score	Example Context
Malware	`APT`	0.82	*APT29 spearphishing campaign*
Vulnerability	`CVE-`	0.79	*CVE-2023-1234 in Apache Log4j*
Indicator	`http://`	0.75	*Malicious URL http://phish.site*
System	`Windows`	0.68	*Windows registry keys modified*
Organization	`Microsoft`	0.61	*Microsoft security advisory*

Table 2 shows which layers of the SecureBERT model contribute to entity identification: the deepest layers (9-12) are essential for recognizing Malware, Vulnerability and Indicator entities, while the initial layers (1-4) classify entities as 'Other', including less frequent or more context-dependent entities that do not fall into the primary categories.

Table 2
Layer contribution percentages for entity types in SecureBERT.

Layer Group	Malware	Vulnerability	Indicator	Other
1-4	12.3%	9.8%	8.2%	69.7%
5-8	28.7%	22.1%	25.4%	23.8%
9-12	59.0%	68.1%	66.4%	6.5%

In summary, the use of Captum combined with explainability techniques allows for the complication of decision-making processes within an NLM. In some cases, it has been observed that specific tokens have been misattributed to non-organizational contexts, showing that the model may overfit or misinterpret clues coming from the context. Moreover, explainability contributes meaningfully to guiding data augmentation strategies. Visualizing attribution scores and analyzing prediction confidence across entity classes allowed to identify underrepresented categories, such as *Secteam*, which contains only 1,345 labeled instances in the combined dataset. These imbalances often go unnoticed through aggregate performance scores alone, but interpretability techniques help prioritize which classes may benefit most from targeted data enrichment. This is particularly relevant in cybersecurity, where emerging threat types or niche entities may be crucial despite their lower frequency. Another important implication concerns standardization. As threat intelligence initiatives increasingly rely on structured formats such as STIX 2.1 [35], it becomes essential that NER models not only achieve high accuracy, but also produce outputs that align semantically and contextually with standardized schemas. Explainability analyses can aid in verifying and improving this alignment.

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 authors have not employed any Generative AI tools.

References

[1] Shackleford, The SANS state of cyber threat intelligence survey: CTI important and maturing, 2016. URL: https://www.sans.org/reading-room/whitepapers/analyst/membership/37177.

[2] 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.

[3] N. McNeil, R. A. Bridges, M. D. Iannacone, B. Czejdo, N. Perez, J. R. Goodall, Pace: Pattern accurate computationally efficient bootstrapping for timely discovery of cyber-security concepts, in: 2013 12th International Conference on Machine Learning and Applications, volume 2, 2013, pp. 60–65.

[4] 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.

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

[6] 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.

[7] J. Hutson, S. Muhammad, N. C. Rowe, HackIntelNER: Named entity recognition for hacker intelligence, in: 2021 IEEE European Symposium on Security and Privacy, IEEE, 2021, pp. 182–187.

[8] N. Alzahrani, Y. Atif, S. Matwin, Cysecner: A fine-grained named entity recognition dataset for cyber security, Data in Brief 39 (2021) 107566.

[9] C. Sabottke, O. Suciu, T. Dumitras, Vulnerability disclosure in the age of social media: Exploiting twitter for predicting real-world exploits, in: USENIX Security Symposium, 2015, pp. 1041–1056.

[10] 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.

[11] N. Capodieci, C. Sanchez-Adames, J. Harris, U. Tatar, The impact of generative AI and LLMs on the cybersecurity profession, in: SIEDS, IEEE, 2024, pp. 448–453.

[12] A. Minutolo, R. Guarasci, E. Damiano, G. De Pietro, H. Fujita, M. Esposito, A multi-level methodology for the automated translation of a coreference resolution dataset: an application to the italian language, Neural Computing and Applications 34 (2022) 22493–22518.

[13] SoBigData.it, Strengthening the Italian RI for Social Mining and Big Data Analytics, 2025. https://pnrr.sobigdata.it.

[14] S. Silvestri, G. F. Russo, G. Tricomi, M. Ciampi, A dataset for the fine-tuning of LLM for the NER task in the cyber security domain, in: Proceedings of the Discovery Science Late Breaking Contributions 2024, volume 3928 of *CEUR Workshop Proceedings*, CEUR-WS.org, Pisa, Italy, 2025.

[15] A. Panwar, iGen: Toward automatic generation and analysis of indicators of compromise (IOCS) using convolutional neural network, Ph.D. thesis, Arizona State University, 2017.

[16] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, R. Beyah, Acing the IOC game: Toward automatic discovery and analysis of open-source cyber threat intelligence, in: 2016 ACM SIGSAC Conference on Computer and Communications Security, ACM, New York, NY, USA, 2016, p. 755–766.

[17] M. Balduccini, S. Kushner, J. Speck, Ontology-driven data semantics discovery for cyber-security, in: Practical Aspects of Declarative Languages, Springer, Cham, 2015, pp. 1–16.

[18] N. Rastogi, S. Dutta, M. J. Zaki, A. Gittens, C. Aggarwal, Malont: An ontology for malware threat intelligence, in: Deployable Machine Learning for Security Defense, Springer, 2020, pp. 28–44.

[19] G. Kim, C. Lee, J. Jo, H. Lim, Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network, International Journal of Machine Learning and Cybernetics 11 (2020).

[20] J. Zhao, Q. Yan, J. Li, M. Shao, Z. He, B. Li, TIMiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data, Computers & Security 95 (2020) 101867.

[21] X. Wang, X. Liu, S. Ao, N. Li, Z. Jiang, Z. Xu, Z. Xiong, M. Xiong, X. Zhang, DNRTI: A large-scale dataset for named entity recognition in threat intelligence, in: IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications, 2020, pp. 1842–1848.

[22] S. Zhou, Z. Long, L. Tan, H. Guo, Automatic identification of indicators of compromise using neural-based sequence labelling, in: Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, PACLIC 2018, ACL, Hong Kong, 2018.

[23] Z. Zhu, T. Dumitras, Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports, in: 2018 IEEE European Symposium on Security and Privacy (EuroS&P), 2018, pp. 458–472. doi:10.1109/EuroSP.2018.00039.

[24] 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.

[25] 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.

[26] E. Cardillo, A. Portaro, M. Taverniti, C. Lanza, R. Guarasci, Towards the automated population of thesauri using bert: A use case on the cybersecurity domain, in: International conference on emerging internet, data & web technologies, Springer, 2024, pp. 100–109.

[27] S. Islam, N. Basheer, S. Silvestri, S. Papastergiou, M. Ciampi, Intelligent dynamic cybersecurity risk management framework with explainability and interpretability of AI models for enhancing security and resilience of digital infrastructure, Research Square (2024).

[28] F. Gargiulo, A. Minutolo, R. Guarasci, E. Damiano, G. De Pietro, H. Fujita, M. Esposito, An electra-based model for neural coreference resolution, IEEE Access 10 (2022) 75144–75157.

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

[30] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 35th International Conference on Machine Learning, volume 70, 2017, pp. 3319–3328.

[31] P. Contributors, Captum ai documentation, 2020. URL: https://captum.ai/, accessed: 2025-05-16.

[32] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?" Explaining Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[33] P. Sturmfels, S. M. Lundberg, S.-I. Lee, Visualizing the impact of feature attribution baselines, 2020.

URL: https://distill.pub/2020/attribution-baselines/, accessed: 2025-05-16.

[34] C. Frye, Q. Le, I. Feige, Asymmetric shapley values: Incorporating asymmetry into explanations of neural network decisions, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 11007–11018. doi:10.48550/arXiv.2006.08602.

[35] O. Open, Stix™ version 2.1, 2021. URL: https://www.oasis-open.org/standard/6426/, accessed: 2025-05-16.