An Analysis of Pre-trained Models to Identify Cybersecurity Incidents Entities in the Healthcare Industry

Rafael Paim[1], **Luciano Ignaczak**[1]

[1] Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Brasil

rafpaim@edu.unisinos.br

lignaczak@unisinos.br

Abstract. *Healthcare institutions have always been a critical sector in any community. Cybersecurity issues, such as attacks or incidents, may impact their operations and cause damage that could eventually lead to patient death. Named Entity Recognition and Classification (NERC) can support these institutions in analyzing incidents, highlighting the incident's type, attack type, and location, just to name a few examples. This work evaluated pre-trained machine learning models to comprehend how they help in this identification. For this purpose, we analyzed two fine-tuned BERT models used in a corpus with incidents related to Healthcare institutions in the U.S. We evaluated the entity recognition using both the Strict and Partial approaches. Experiment results indicated a higher precision (above 0.776) but with low Recall, with less than 0.267. This may indicate a good performance for entity recognition. However, the models missed many entities.*

1. Introduction

In recent years, we have seen an incredible growth in technology applied to the Healthcare sector, allowing innovations such as health sensors and monitoring devices to transmit live data [Meskó et al. 2017]. Medical Devices, Electronic Medical Records (EMR), Picture Archiving Communication Systems (PACS), and distinct protocols have leveraged patient care and advanced digital health information across the industry. As mentioned in Luh and Yen 2020, technology offers a window of opportunity to take patient care to another level, and as a consequence, there is a rise in threats to Digital Healthcare. According to Bhuyan et al. 2020, cyber actors have consistently targeted the health industry to steal data. It is worth pointing out that health data is extremely sensitive and expensive. Franco et al. 2025 reported the consistent increase in data breach costs in the Healthcare sector. For twelve consecutive years, Healthcare incidents have had a higher median cost than all other sectors – roughly US$ 10,10 million as mentioned by IBM 2022.

Cyber Threat Intelligence (CTI) can play a big role in helping Healthcare providers to mitigate such adverse events. According to Future 2019, CTI is the knowledge that enables companies to avoid and mitigate attacks by providing useful information and insights about the threats, considering the context of where such an enterprise is placed. CTI can collect information from many useful sources, produce consistent and meaningful data analysis, and present results at a high level to make decision-making easier. Research in this area evaluates the use of text mining, Named Entity Recognition (NER),

and Masked-Language Model (MLM) to support the cyber-threat/incident classification and/or identification. Ranade et al. 2021 used the BERT model to identify threats and vulnerabilities, which is useful for Security Operations Centers (SOC). Aghaei et al. 2023 also used the BERT model for vulnerability classification, with a novel tokenization process derived from the RoBERTa framework and its capability to identify cybersecurity terms. Li et al. 2021 took advantage of posts from threat actors' forums to propose a BERT model to detect cybersecurity neologisms to be proactive in threat prevention. Finally, Iqbal et al. 2019 analyzed chat log sessions from criminals and used text mining and NER to capture information about illegal acts, supported by the Wordnet dataset.

In this study, we proposed to answer the question: "What is the performance of pretrained models to detect entities related to the Healthcare Industry incidents?". To achieve this, we designed an experiment to evaluate two fine-tuned BERT models against a corpus with information regarding cybersecurity incidents in the health domain, correlating four common entities between these models, gathering news from specialized websites that covered Healthcare and cybersecurity news, and extracting the information provided in these data sources to view in which states of the United States, the incidents happened, the kind of them, which institutions have suffered from cyber issues, and if there are any vulnerabilities associated with them.

This paper is divided as follows: Section 2 presents the discussion of the selected studies, Section 3 depicts the methodology applied which allowed our analysis, Section 4 holds the outcomes by the analyzed models and present some insights about Healthcare incidents, and Section 5 holds the conclusion, limitations of this work and future improvements.

2. Related Work

In this section, we discuss each selected study for this work. Four articles used BERT as the primary model to accomplish such tasks. Two of them were designed to present new, fine-tuned, trained models in the cybersecurity domain (CyBERT and SecureBERT), while the other two BERT-based papers focus on the use of the model in real-world applications, specifically vulnerability prioritization and the identification of neologisms.

Ranade et al. 2021 stands out for being a fine-tuned spin-off from BERT for cybersecurity terms, named CyBERT, which can be used for information extraction, prediction, and classification tasks. The authors' goal was to train a new model for cybersecurity-related tasks. Their contributions are delivering a new model that can understand and process cybersecurity terminology with high precision and creating a domain-specific corpus. Aghaei et al. 2023 developed SecureBERT as a new trained cybersecurity terminology model. This language model can analyze semantics and relationships between texts in the context of cybersecurity, allowing the use of automated CTI tasks. The training corpus had about a billion words from a large variety of cybersecurity terms and resources. This approach enables it to be used for phishing detection, sentiment analysis, and NER tasks.

Silvestri et al. 2023 used BERT with XGBOOST to classify, prioritize vulnerabilities, and predict attacks within the healthcare ecosystem. This model correlates threats and vulnerabilities based on its own way of listing the institution's assets. The last paper that uses BERT is Li et al. 2021, which proposed a novel way to detect Cybersecurity words. Their framework can automatically identify hacker neologisms from hackers' forums,

aiming to discover if a new threat is in the wild. This model used NER, Convolutional Neural Networks (CNN), and Bidirectional Long-Short Term Memory (BiLSTM), to process the texts.

Four papers used Machine Learning (ML) or Deep Learning (DL) algorithms to detect vulnerabilities [Samtani et al. 2022, Islam et al. 2022], incidents [Mumtaz et al. 2023], and cyber attacks / cyber threats [Alqudhaibi et al. 2023]. The paper presented by Samtani et al. 2022 targeted the Healthcare sector and Supervisory Control and Data Acquisition (SCADA), and aimed to link exploits mentioned in dark web forums with vulnerabilities in real entities to identify and prioritize vulnerability patching. The next study was Mumtaz et al. 2023, which evaluated the role of COVID-19 on cybersecurity incidents. The authors collected incident data from reports and divided the incidents captured into five categories: Brute Force, Denial of Service (DoS), Malware, Espionage, and APT. Using Machine Learning algorithms with data from pre- and post-pandemic times, the authors used Bag-of-Words (BoW) and N-Gram tasks to extract texts and correlate information such as year and continent.

Islam et al. 2022 developed a model based on machine learning algorithms: Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). The study identified the exploitability of vulnerabilities in the Healthcare sector using an ontology approach. The last paper selected, Iqbal et al. 2019, used the Wordnet lexical database together with cyber criminal chat-log sessions (dataset) seized by Canadian Police to apply their model combined with NER to identify illegal activities among entities in chat sessions. The system identified words that frequently appeared in the texts.

In the same way as the aforementioned studies, we used the BERT model to analyze cybersecurity incidents in the Healthcare sector. Ranade et al. 2021 and Aghaei et al. 2023 built CyBERT and SecureBERT to be pre-trained models to be used with downstream tasks. Although Li et al. 2021 also used BERT in their paper, the authors did not use a fine-tuned model as the previous ones did. Our work used the SecureBERT pre-trained model to process text data. Different from Samtani et al. 2022, Silvestri et al. 2023 as well as Islam et al. 2022 that captured data only from cyber-specialized pages, this work captured information from Healthcare / Cybersecurity News websites. Regarding related studies, only two dealt with incidents: Iqbal et al. 2019 and Mumtaz et al. 2023, but they did not consider the Healthcare sector in their evaluation. Alqudhaibi et al. 2023 had its efforts towards Industry 4.0, not focused on Healthcare. We analyzed incidents in the Healthcare sector using NER techniques and the SecureBERT model.

3. Experiment Design

In this section, we detail the experiment applied to this work, as depicted in Figure 1. In the first stage, we created a corpus with news extracted from websites, using two scripts that traversed web search results pages, capturing 24,108 news. Next, we performed a domain selection to rank the websites with a certain quantity of news in favor of websites that hold only a few news items, intending to identify the most relevant domains. The third phase comprised data preprocessing, excluding duplicated or irrelevant records. Then we applied the NERC models proposed to evaluate the results, using the F1 score metric, considering the performance of the models against the analysis carried out manually by the author. The details of this experiment are in the subsequent sections.

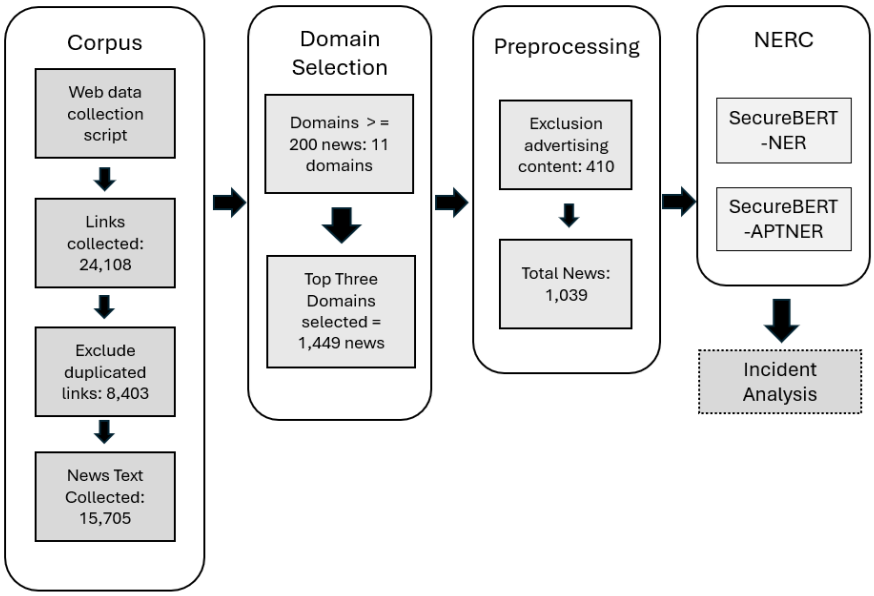

Figure 1. Methodology Overview

3.1. Data Collection

This study corpus comprises cybersecurity news regarding incidents related to the healthcare sector. For this, we chose a well-known web search mechanism to collect and store the data in a SQLite database. To accomplish this task, we needed to create two Python scripts: a crawler to query the search engine and iterate through all web search results pages, saving them in a folder repository to be parsed by our second script. The second script, a web scraper, captured the previous data and saved it on our database, recording at least the link, title, description (a metadata text inside the HTML head tag, which contains a very short summary of the webpage text), and news date.

The author used four search queries: 'hospital AND cyberattack', 'hospital AND cybersecurity AND incident', 'healthcare AND cyberattack', and 'healthcare AND cybersecurity AND incident'. This approach yielded 24,108 registers, from 2,059 news portals. We created a flag in the database to avoid considering the already scraped/crawled registers. The next step was to eliminate duplicate results. Using an SQL query, we identified 8,403 duplicated links, leaving 15,705 usable news articles for our experiment. This task was done on October 30th, 2023, and the period chosen to grab the data was from January 1st to October 30th, 2023.

3.2. News Portal Selection and Preprocessing

After eliminating duplicated entries, the remaining news summed 15,705. At this point, it was important to know the representativeness of each domain (site) as regards news quantity per site, since it would give a better overview of the information distribution in the corpus and could be used to remove news portals with less relevance. To perform this task, the author defined that portals with equal to or more than 200 news articles should be considered. This criterion was chosen after previous database evaluations that demonstrated that a greater number of sites had one or a few records, and some websites

had many more news compared to their peers. The process resulted in eleven news portals comprising 200 news or more.

Since we had eleven sites with 3,586 news, we had to choose which news portals would be used in our evaluation. Thus, we selected the top three news portals, as follows: Security Magazine (618 news), Becker's Hospital Review (489 news), and JD Supra (342 news). In the end, this approach produced 1,449 news to be used. We conducted a review of this news and excluded 410 records because their contents were not aligned with the purpose of this work, ending up with 1,039 records.

3.3. NERC Approaches

According to Aghaei et al. 2023, large language models, such as BERT, are trained in more general corpora in the English language, becoming excellent in NLP tasks such as sentiment analysis or text prediction. However, they do not deal well with some specific domains such as healthcare or cybersecurity. Specific domains hold many words (tokens) as a regular language. The challenge is that those words can possess different meanings within a particular context than in regular language. For instance, the word virus has a meaning in the health domain but is not the same in the cybersecurity domain.

As mentioned by Ajagbe and Zhao 2022, the transfer learning capability of the BERT model can improve the model's understanding of new features for a given domain, enabling the model to learn new information more efficiently. With this in mind, we considered using SecureBERT from Aghaei et al. 2023, due to the number of documents used to train this model and the higher F1 score achieved. Unfortunately, the use of SecureBERT was not satisfactory. We ran all the texts against the model, as initially proposed. However, this did not produce useful results to be analyzed, since we received tags as Label_0 and Label_1. So, we had to change our approach, looking for SecureBERT fine-tuned models. We found 41 fine-tuned models derived from SecureBERT in the Hugging Face Portal. As our scope is to use models trained in the Token Classification task, we ended up with six models. We considered models trained with cybersecurity content and bibliographic references to filter the results, so we chose the following models: SecureBERT-NER and SecureBERT-APTNER.

Both models were trained in the APTNER dataset [Wang et al. 2022], created with information collected from APT reports of many cybersecurity companies, and properly annotated for NER tasks. The models have at least 20 entities: SecureBERT-NER recognizes 22 entities, and SecureBERT-APTNER recognizes 20. For this work, we chose the following entities present in both models: B-IDTY, B-ACT, B-VUL, and B-LOC. The entity B-IDTY refers to identity, in other words, an organization or institution. Our dataset would include hospitals, clinics, and/or some organizations related to the healthcare sector. The B-ACT entity pointed to the attack action involved in the incident. B-VUL refers to the vulnerability discovered in the texts. B-LOC stands for location, which would be a city, state, region, and/or Country.

3.4. Evaluation Approach

In order to evaluate the model's performance, we decided to manually extract the entities of the texts and compare the results. To do this, we selected 60 news articles (20 of each portal) and reduced these texts to the first 400 tokens. Since many texts have more than

800 words, this reduction allowed us to be more assertive in this step. The author manually extracted the entities and classified them according to the tags mentioned. Ultimately, we got more than 1,000 entities distributed across these four tags. With all information organized, we calculated the True Positive (TP), False Negative (FN), and False Positive (FP) for each result returned by the models compared to the author's manual analysis.

For calculation, we considered the Strict and Partial approaches as mentioned by Segura-Bedmar et al. 2013. The Strict approach is based on the concept that when the model makes a prediction, there is a match between the type (e.g.: IDTY, ACT, VUL, or LOC) and the boundaries (the beginning and the end of the word) of the entity predicted, thus considering the prediction as a true positive. The Partial approach is not as rigorous as the Strict; it only cares about a partial matching of the entity boundaries, without considering the assigned tag type, so it could be helpful in some scenarios when part of the entity is essential for the prediction results. Regarding the metrics used to evaluate the models, we used the F1-score metric, as mentioned by Sundheim 1995, to assess the performance of the NERC systems. The F1-score consists of the harmonic mean of precision and recall, which, according to Powers 2011, produces a balanced result between both metrics.

4. Results

This section contains the results obtained and proposes some discussion. It was divided into the following sections: section 4.1, which presented an overall analysis of each model's performance, based on the sixty news articles manually collected by the author; section 4.2 demonstrated analysis of the entity identified by each model; finally, the section 4.3 displayed results by running the models against the entire corpus.

4.1. Overall Performance Analysis

As mentioned previously, we manually extracted articles from our dataset. This process gave us 1,041 entities. This was our comparison baseline for both models. The next step was to run each BERT model proposed against the same 60 texts. We considered the Strict and the Partial approach for SecureBERT-NER and SecureBERT-APTNER, as cited by Ignaczak et al. 2023. Table 1 depicts the number of entities found by each model, in both the Strict and Partial approaches.

Table 1. Total quantity of entities found by the fine-tuned models, SecureBERT-NER and SecureBERT-APTNER, in both Strict and Partial approaches, using the dataset created with healthcare news.

Model	Type	Correct	Incorrect	Partial	Missed
SecureBERT-NER	Strict	253	73	0	788
	Partial	255	46	31.0	786
SecureBERT-APTNER	Strict	116	80	0	925
	Partial	116	23	61.0	925

The SecureBERT-NER model found 253 correct entities in strict type and 255 in partial type. The number of missed entities was 788 and 786 for strict and partial types, respectively. Only for the partial mode, we had 31 entities identified. Regarding

the incorrect entities identified, we had 73 entities for strict type and 46 hits for partial type. The partial entities identified were relevant to the context since they identified hospitals' names that were some kind of abbreviation in the texts after the full name was first presented. The strict approach disregarded these pieces of information.

The SecureBERT-APTNER model returned 116 correct entities in strict type and the same value for the partial type. Regarding the missed entities, we got 925 in the strict and partial types. The model also returned 61 entities only for the partial type. We had 80 in the strict type and 23 in the partial type for the incorrect entities. The results achieved by the SecureBERT-APTNER model were very close to those reached by the SecureBERT-NER model. We highlighted the low number of correct entities discovered by both models.

Table 2 presents the metrics calculated for both models. We calculated the precision, recall, and F1-score for the strict and partial types. For both models, we had a precision above 0.7 (or 70%); however, the recall was low. For SecureBERT-NER, we got 0.243 (strict) and 0.267 (partial). For SecureBERT-APTNER, we got 0.111 (strict) and 0.161 (partial); pushing the F1-score to 0.370 (strict) and 0.407 (partial) for the SecureBERT-NER, and 0.188 (Strict) and 0.272 (partial) for the SecureBERT-APTNER. This may indicate that both models performed well in detecting true positives, but not all true positive entities that should have been identified were truly detected.

Table 2. Precision, Recall, and F1-score results for each model considering the set of 60 healthcare news

Model	Type	Precision	Recall	F1-Score
SecureBERT-NER	Strict	0.776	0.243	0.370
	Partial	0.861	0.267	0.407
SecureBERT-APTNER	Strict	0.592	0.111	0.188
	Partial	0.885	0.161	0.272

Both models missed a significant number of entities. SecureBERT-NER detected 24% of all information in the strict approach. One reason for explaining the performance is how the models were trained. Since the corpora used for training these models was created with various reports from Advanced Persistent Threats (APT) of many security companies, as mentioned by Wang et al. 2022, probably, they did not have texts containing healthcare terms. We believe this could be a reason for the models' present difficulty in identifying words and contexts related to the healthcare industry.

4.2. Performance Analysis by Entity

Based on the results presented earlier, we verified the performance of these models by looking at each tag. We chose four tags to identify information: B-IDTY, B-ACT, B-VUL, and B-LOC; which means Identity, Attack Action, Vulnerability type, and Location. Our purpose was to check if the models could identify as many entities as possible in the healthcare sector and collect data about how the threat actors attacked their targets, which vulnerabilities were used, and in which locations the hospitals/clinics were targeted. We were aware that not all entities could be found in the texts provided.

Table 3 shows the total of entities identified by the models, considering each news portal selected. It is possible to see that the entity IDTY was the most frequently found information. Both models correctly identified hospitals, clinics, and government institutions. Only a few entities were different from the healthcare sector. However, we accepted these texts due to their contents, which were news about the healthcare sector.

We have a certain number of detected ACT and LOC. But, the number of VUL entities found was low, with only six discovered. We believed it was because of the particularities of the texts, mainly in two of the three news portals. Except for the news portal SC Magazine - a webpage focused on cybersecurity news, the other two are somehow the opposite of this one - enlightening the healthcare sector, plus adding some cybersecurity news. This situation tends to provide news with less specific cybersecurity terms, impacting the classification task to extract this kind of token.

Table 3. Total of entities found by the SecureBERT-NER (Model 1) and SecureBERT-APTNER (Model 2) compared to the author's manual entity extraction (in parentheses).

Model	Data source	IDTY		ACT		VUL		LOC	
Model 1	SC Magazine	46	(208)	21	(67)	3	(10)	16	(26)
	Becker's	30	(106)	18	(54)	1	(2)	7	(48)
	JD Supra	103	(330)	4	(150)	0	(0)	4	(40)
	Total	**179**	**(644)**	**43**	**(271)**	**4**	**(12)**	**27**	**(114)**
Model 2	SC Magazine	9	(208)	12	(67)	1	(10)	16	(26)
	Becker's	12	(106)	9	(54)	1	(2)	22	(48)
	JD Supra	28	(330)	2	(150)	0	(0)	4	(40)
	Total	**49**	**(644)**	**23**	**(271)**	**2**	**(12)**	**42**	**(114)**

The texts related to cybersecurity had more affinity with the way the models were trained and, consequently, influenced the token classification task. In this case, we have identified more ACT and VUL entities in the SC Magazine news portal than in the other two portals. In SC Magazine, for example, we could find ACT entities as *Ransomware*, *Phishing*, and VUL tag; entities like *Shadow IT* and *Insecure Passwords* were found. About the LOC tag, we had a reasonable number of entities identified since some texts make explicit this information. For example, words like *Illinois* or *Philadelphia* were correctly found.

Table 4 shows the performance of the SecureBERT-NER and SecureBERT-APTNER models against each data source or news portal. We calculated the metrics precision, recall, and F1-score to verify how each model would perform with the data source provided, considering the 60 texts analysis. To calculate this metric, we chose the results from the partial approach since it was the best approach. As expected, the behavior is similar to the results presented in Table 2, in which we had higher precision values, but lower recall values, indicating an unbalanced model.

4.3. Healthcare News Analysis

We performed an additional analysis applying the entire corpus in each model to verify which entities and insights we received from it. Since we used three news portals as our

Table 4. Models' performance in entity recognition across different data sources

Model	Data source	Precision	Recall	F1-Score
SecureBERT-NER	SC Magazine	0.843	0.288	0.429
	Becker's Hospital	0.892	0.274	0.419
	JD Supra	0.862	0.252	0.390
SecureBERT-APTNER	SC Magazine	0.964	0.163	0.278
	Becker's Hospital	0.873	0.272	0.415
	JD Supra	0.838	0.113	0.199

data source to create the corpus and further conduct the evaluations, the results presented by these two models did not represent the entirety of cybersecurity incidents in the healthcare sector in the U.S. Country. Still, considering the low F1-score metrics of each model, the discussions presented are a glimpse of the incidents in this particular sector.

The models provided some insights regarding cybersecurity incidents in the healthcare sector. Analyzing the incidents in the U.S., the following ACT entities highlighted themselves in the results: Data Breach, Ransomware, and Phishing. We point out that in some cases, we had a combination of these incidents in the same place. The more general ACT entities, like "exploits" or "cyberattacks", were not included in our evaluation. Considering the same concept used for the ACT entity, for LOC entities we did not consider words like the "U.S.", because we wanted tokens that contained more specific regions, such as Georgia, or cities, such as "Chicago" or "Altamonte Springs" to which we could identify the U.S. state to which it belongs.

Another situation to justify the results is that the texts written in this news had their own particularities, syntax, and terminology that contributed to the performance shown by our models. The healthcare jargon and the lack of information about the cybersecurity incident probably contributed to the models not extracting more relevant ACT or VUL entities. Additionally, each news portal had its singularities: SC Magazine is a cybersecurity source of news; Becker's Hospital Review is a source of hospitals, clinics, and healthcare general information news; and JD Supra is strictly focused on managerial and C-level business information.

It is possible to verify the differences among the media outlets in terms of the news published on them. For example, many JD Supra texts did not mention the vulnerability or the attack action involved. They used expressions like *"...Upon discovering that sensitive consumer data may have been accessible to an unauthorized party...", "...began sending out data breach notification letters to all individuals who were impacted by the recent data security incident..."*. It may be easy for humans to understand that the incidents involved could be data breaches, but this is not clear for the models. Regarding the IDTY tag, we found the institution's full name in most parts of the news and its short name later in the texts. This behavior might have impacted the models' capability since sometimes the model identified both versions or just one of them. For example, the news used Mena Regional Health System or Mena Regional as the same institution.

Finally, we may tell both models to have good precision in identifying entities (even if they missed a lot of words), which means that if the models found ransomware is

an ACT entity, we can have good confidence that this prediction is correct. However, their low performance in identifying all possible (positive) entities lowered the F1-score results, making both models unsuitable for useful insights about healthcare incidents, without manual revision of the results by humans. Regarding this, CrowdStrike 2023 denotes that, despite many technologies that could be used to aid the customers to be protected using CTI, it is needed a human to take part in the analysis, reviewing the information and considering the customer/institution context to decide which information could be valuable to an institution.

5. Conclusion

This study carried out an analysis of specific news gathered from healthcare portals and evaluated two fine-tuned models available from the Huggingface website, considering four entities: IDTY, ACT, VUL, and LOC. The analysis involved news about healthcare and cybersecurity from three news portals, each of which had its own kind of audience, from more technical to more managerial, which brought texts with diverse grammar and semantics for the fine-tuned models used.

Our evaluation was based on the results provided by the SecureBERT-NER and SecureBERT-APTNER models. For evaluation metrics, we used the F1-score as it was designed to be used when you need to consider precision and recall metrics. Both models performed an F1-score lower than 0.4; SecureBERT-NER achieved an F1-score of 0.370, and SecureBERT-APTNER reached 0.272 in the same metric. Although they presented a good precision - 0.776 and 0.592 - their recall was very low, 0.243 and 0.111 - indicating that these models missed a reasonable amount of valid entities presented in the texts. One of the possible reasons was the model's lack of understanding of healthcare jargon.

In future work, studies can use more than three sources of news/information to extract entities to be analyzed. This would increase the amount of information collected and evaluated. Also, extending the period by a year or two, rather than the eight months chosen by the author to collect the data is possible. A second suggestion is to grab the original SecureBERT model and fine-tune/train it to understand the healthcare jargon to evaluate its performance. One of the main issues already discussed in this study was the lack of understanding of this specific terminology, which resulted in many entities being despised, missed, or taken as spurious. Finally, since not all reports will have details about the vulnerability explored or the attack action, other studies can collect data from different news sources, such as Twitter/ X, Telegram, and other social networks.

References

[Aghaei et al. 2023] Aghaei, E., Niu, X., Shadid, W., and Al-Shaer, E. (2023). Securebert: A domain-specific language model for cybersecurity. In Li, F., Liang, K., Lin, Z., and Katsikas, S. K., editors, *Security and Privacy in Communication Networks*, pages 39–56, Cham. Springer Nature Switzerland.

[Ajagbe and Zhao 2022] Ajagbe, M. and Zhao, L. (2022). Retraining a bert model for transfer learning in requirements engineering: A preliminary study. In *Retraining a BERT Model for Transfer Learning in Requirements Engineering: A Preliminary Study*, pages 309–315.

[Alqudhaibi et al. 2023] Alqudhaibi, A., Albarrak, M., Aloseel, A., Jagtap, S., and Salonitis, K. (2023). Predicting cybersecurity threats in critical infrastructure for industry 4.0: A proactive approach based on attacker motivations. *Sensors*, 23(9).

[Bhuyan et al. 2020] Bhuyan, S. S., Kabir, U. Y., Escareno, J. M., Ector, K., Palakodeti, S., Wyant, D., Kumar, S., Levy, M., Kedia, S., Dasgupta, D., and Dobalian, A. (2020). Transforming healthcare cybersecurity from reactive to proactive: Current status and future recommendations. *Journal of Medical Systems*, 44(5):98.

[CrowdStrike 2023] CrowdStrike (2023). Threat intelligence. Accessed: 2024-10-30.

[Franco et al. 2025] Franco, M. F., Soares, L. R., and Nobre, J. C. (2025). Saúde sob ataque: Da avaliação de riscos ao desenvolvimento de estratégias de investimentos em cibersegurança na Área da saúde. In *Anais do 25º Simpósio Brasileiro de Computação Aplicada à Saúde – SBCAS 2025*. Sociedade Brasileira de Computação. Acessado em julho de 2025.

[Future 2019] Future, R. (2019). The threat intelligence handbook - second edition.

[IBM 2022] IBM (2022). Ibm cost of data breach report 2022.

[Ignaczak et al. 2023] Ignaczak, L., Martins, M. G., da Costa, C. A., Donida, B., and da Silva, M. C. P. (2023). An evaluation of nerc learning-based approaches to discover personal data in brazilian portuguese documents. *Discover Data*, 1(1):5.

[Iqbal et al. 2019] Iqbal, F., Fung, B. C. M., Debbabi, M., Batool, R., and Marrington, A. (2019). Wordnet-based criminal networks mining for cybercrime investigation. *IEEE Access*, 7:22740–22755.

[Islam et al. 2022] Islam, S., Abba, A., Ismail, U., Mouratidis, H., and Papastergiou, S. (2022). Vulnerability prediction for secure healthcare supply chain service delivery. *Integr. Comput.-Aided Eng.*, 29(4):389–409.

[Li et al. 2021] Li, Y., Cheng, J., Huang, C., Chen, Z., and Niu, W. (2021). Nedetector: Automatically extracting cybersecurity neologisms from hacker forums. *Journal of Information Security and Applications*, 58:102784.

[Luh and Yen 2020] Luh, F. and Yen, Y. (2020). Cybersecurity in science and medicine: Threats and challenges. *Trends in Biotechnology*, 38(8):825–828.

[Meskó et al. 2017] Meskó, B., Drobni, Z., Éva Bényei, Gergely, B., and Győrffy, Z. (2017). Digital health is a cultural transformation of traditional healthcare. *Mhealth*, 3:38.

[Mumtaz et al. 2023] Mumtaz, G., Akram, S., Iqbal, W., Usman Ashraf, M., Almarhabi, K. A., Alghamdi, A. M., and Bahaddad, A. A. (2023). Classification and prediction of significant cyber incidents (sci) using data mining and machine learning (dm-ml). *IEEE Access*, pages 1–1.

[Powers 2011] Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, 2(1):37–63. Acessado em julho de 2025.

[Ranade et al. 2021] Ranade, P., Piplai, A., Joshi, A., and Finin, T. (2021). Cybert: Contextualized embeddings for the cybersecurity domain. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3334–3342.

[Samtani et al. 2022] Samtani, S., Chai, Y., and Chen, H. (2022). Linking exploits from the dark web to known vulnerabilities for proactive cyber threat intelligence: An attention-based deep structured semantic model1. *MIS quarterly*, 46(2).

[Segura-Bedmar et al. 2013] Segura-Bedmar, I., Martínez, P., and Herrero-Zazo, M. (2013). SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts. In Manandhar, S. and Yuret, D., editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 341–350, Atlanta, Georgia, USA.

[Silvestri et al. 2023] Silvestri, S., Islam, S., Papastergiou, S., Tzagkarakis, C., and Ciampi, M. (2023). A machine learning approach for the nlp-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem. *Sensors*, 23(2).

[Sundheim 1995] Sundheim, B. M. (1995). Overview of results of the muc-6 evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 13–31, Columbia, Maryland. ACL.

[Wang et al. 2022] Wang, X., He, S., Xiong, Z., Wei, X., Jiang, Z., Chen, S., and Jiang, J. (2022). Aptner: A specific dataset for ner missions in cyber threat intelligence field. In *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1233–1238.