B.Sc. in Computer Science and Engineering Thesis

Enhancing Cyber Threat Intelligence with Transformer-Based Models: A Comparative Analysis with Few-Shot Learning

Submitted by

Nafis Karim
201805027

Supervised by

Dr. Md. Shohrab Hossain

Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology

Dhaka, Bangladesh

June 2024

CANDIDATES' DECLARATION

This is to certify that the work presented in this thesis, titled, "Enhancing Cyber Threat Intelligence with Transformer-Based Models: A Comparative Analysis with Few-Shot Learning", is the outcome of the investigation and research carried out by us under the supervision of Dr. Md. Shohrab Hossain.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Nafis Karim
201805027

ACKNOWLEDGEMENT

We are profoundly grateful for the support and guidance provided by several individuals without whom this research could not have been completed.

First and foremost, special thanks to Prof. Ren-Hung Hwang from the College of AI at National Yang Ming Chiao Tung University, Tainan, Taiwan. Similarly, appreciation is extended to Prof. Ying-Dar Lin from the Department of Computer Science at National Chiao Tung University, Taiwan. Their expertise in network security and critical feedback on the thesis have been instrumental in refining the research outcomes.

Their combined contributions have not only enhanced this academic work but have also contributed significantly to my personal growth and professional development in the field of cybersecurity.

Finally, I would like to extend my gratitude to my family, especially my wife for their unwavering support and encouragement throughout the duration of my studies.

Nafis Karim

Dhaka
June 2024

Contents

List of Figures

List of Tables

ABSTRACT

In the fast-changing field of cybersecurity, effective Cyber Threat Intelligence (CTI) is crucial for predicting and preventing threats. CTI involves collecting and analyzing information about potential cyber threats to protect systems and networks from attacks. One of the key techniques in CTI is Named Entity Recognition (NER), a Natural Language Processing (NLP) method that identifies and categorizes important entities such as threat actors, malware, and vulnerabilities from unstructured text. Accurate NER is essential for extracting actionable intelligence from the vast amount of textual data available in cybersecurity reports, blogs, and other sources.Traditional NER models have faced significant challenges in the CTI domain due to the unique and evolving nature of cyber threats, as well as the specialized vocabulary used in threat reports. To address these challenges, this study focuses on leveraging transformer-based models to enhance NER capabilities within the CTI domain. Transformer-based models, such as BERT and RoBERTa, have achieved state-of-the-art performance in various NLP tasks, and this research aims to evaluate their effectiveness in the context of CTI.The datasets used in this study, DNRTI and APTNER, contain detailed annotations of cybersecurity entities, making them ideal for training and evaluating NER models. Our approach includes thorough data preprocessing to handle the specific characteristics of cybersecurity text, followed by specialized model training tailored to recognize cybersecurity entities. We employ a range of evaluation metrics, including precision, recall, and F1 score, to assess the performance of the models.The results of this study demonstrate that transformer-based models significantly outperform traditional NER methods in accurately identifying cybersecurity entities. This improvement highlights the potential of advanced machine learning techniques to enhance the extraction and classification of CTI. Furthermore, we conducted a comparative analysis of few-shot learning models, including GPT-3.5, GPT-4.0, and Google's Gemini, against the fine-tuned transformer-based models. While few-shot learning models showed some improvements with more examples, the transformer-based models consistently outperformed them across all metrics. This study underscores the critical role of transformer-based models in advancing NER tasks for CTI, offering significant improvements over traditional methods. The superior performance of these models in accurately identifying and categorizing cybersecurity entities demonstrates their importance in enhancing cybersecurity practices. By leveraging advanced machine learning techniques, this research contributes to the ongoing efforts to develop more effective tools for predicting and preventing cyber threats.

Chapter 1

Introduction

As we navigate through an era marked by a significant surge in complex and sophisticated cyber threat reports, the realm of digital security is confronted with challenges of an unprecedented scale. This escalation underlines the critical need for enhanced Cyber Threat Intelligence (CTI), which serves as a cornerstone for preemptive cybersecurity strategies. CTI acts as a guiding light for the cybersecurity community, offering invaluable insights into potential threats and vulnerabilities through the analysis of copious amounts of data, much of which is unstructured and scattered across a plethora of digital sources.

At the core of effectively leveraging this vast reservoir of information lies Named Entity Recognition (NER), a crucial technique within the domain of Natural Language Processing (NLP). NER is instrumental in sifting through the digital morass to identify and systematically categorize vital data points, such as the identities of threat actors, malware signatures, and system vulnerabilities. However, the application of NER in the cybersecurity context is fraught with challenges, primarily due to the domain-specific terminology, the fluid nature of cyber threat landscapes, and the intricate syntax of cybersecurity discourse.

The journey of NER in the context of CTI has traversed various methodologies, starting from rule-based systems that depend on meticulously defined patterns and dictionaries, to statistical models that harness data features for learning. These foundational approaches, while pioneering, have often struggled to keep pace with the dynamic terminologies inherent in cybersecurity, and have been hampered by the need for extensive manual adjustments and updates.

The advent of deep learning heralds a new chapter marked by the introduction of transformative models such as BERT, RoBERTa, and GPT. These models have revolutionized the field of NLP with their profound contextual comprehension and their capability to generate text that closely mimics human writing. This revolution offers promising avenues for the precise detection and classification of cybersecurity-related entities, navigating the complexities of cybersecurity lexicon with unparalleled efficiency.

This study delves into the application of these advanced transformer-based models to address the intricacies of NER within the sphere of CTI. By harnessing the detailed annotations provided in the DNRTI and APTNER datasets, this research conducts a meticulous evaluation of the performance of these pre-trained transformer models, which have been fine-tuned to adeptly navigate the specialized language of cybersecurity.

Our investigation uncovers significant enhancements in the accuracy of entity recognition, showcasing a marked superiority over traditional NER methods. This underscores the immense potential of deep learning technologies in transforming CTI, paving the way for the automation and refinement of CTI processes. In doing so, this paper contributes to fortifying our cyber defenses against the increasingly sophisticated digital threats of today's age, highlighting the indispensable role of advanced NLP techniques in evolving cybersecurity strategies.

Furthermore, we performed a comparative analysis of few-shot learning models, including GPT-3.5, GPT-4.0, and Google's Gemini, against the fine-tuned transformer-based models. While few-shot learning models showed some improvements with more examples, the transformer-based models consistently outperformed them across all metrics. This study demonstrates the superior effectiveness of transformer-based models in improving NER tasks for CTI and underscores their critical role in advancing cybersecurity practices.

1.1 Motivation

The rapid escalation in both the frequency and sophistication of cyber threats has made it imperative to develop advanced methodologies for effective Cyber Threat Intelligence (CTI). Cybersecurity breaches can have devastating consequences, including financial loss, damage to reputation, and compromising of sensitive information. As organizations increasingly rely on digital infrastructure, the need for robust CTI mechanisms to anticipate, detect, and mitigate cyber threats has never been more critical.

One of the fundamental challenges in CTI is the ability to efficiently process and analyze vast amounts of unstructured data from diverse sources such as threat reports, blogs, forums, and social media. This data often contains crucial information about potential threats, but its unstructured nature makes it difficult to extract actionable intelligence. Named Entity Recognition (NER), a core technique in Natural Language Processing (NLP), plays a pivotal role in this context by identifying and categorizing significant entities within the data, such as the names of threat actors, malware, and system vulnerabilities.

However, traditional NER models face significant limitations when applied to the cybersecurity domain. These models often struggle with the specialized vocabulary and rapidly evolving terminologies characteristic of cybersecurity discourse. Moreover, the syntax and semantics of

cyber threat reports are complex and require a sophisticated understanding to accurately interpret and classify the entities. These challenges necessitate the exploration of more advanced models capable of addressing these domain-specific intricacies.

The introduction of transformer-based models, such as BERT and RoBERTa, represents a significant advancement in the field of NLP. These models have demonstrated remarkable success in various NLP tasks due to their ability to capture contextual relationships and generate human-like text. Their potential application in CTI, particularly for NER tasks, is a promising avenue that warrants thorough investigation.

This study is motivated by the need to enhance NER capabilities within CTI using transformer-based models. By leveraging the detailed annotations provided in specialized datasets like DNRTI and APTNER, this research aims to evaluate and demonstrate the effectiveness of these advanced models in accurately identifying and categorizing cybersecurity entities. The potential benefits of this approach include:

- **Improved Accuracy:** Transformer-based models have the potential to significantly improve the accuracy of entity recognition in cybersecurity texts, leading to more reliable and actionable intelligence.
- **Automation and Efficiency:** Enhanced NER capabilities can automate the extraction of critical information from vast amounts of unstructured data, thereby increasing the efficiency of CTI processes.
- **Adaptability:** The ability of transformer models to adapt to the evolving nature of cyber threats ensures that the CTI mechanisms remain relevant and effective in the face of new challenges.
- **Strengthening Cyber Defenses:** By providing more accurate and timely threat intelligence, advanced NER models can contribute to fortifying organizational defenses against cyber attacks.

This research seeks to address the existing gaps in NER for CTI and contribute to the development of more effective and resilient cybersecurity strategies. The insights gained from this study have the potential to drive significant advancements in the field, ultimately enhancing the ability of organizations to protect themselves against increasingly sophisticated cyber threats.

1.2 Problem Statement

In our research, we aim to enhance the effectiveness of Named Entity Recognition (NER) techniques in Cyber Threat Intelligence (CTI) using advanced transformer-based models. The core

of this study revolves around the identification and classification of cyber threat entities from unstructured text data. The input and output format for our problem statement are as follows:

Problem Statement

 # Input: Unstructured text data from cybersecurity reports.
 # Output: Structured data categorizing key entities such as threat actors, malware, and vulnerabilities.

Table 1.1 demonstrates an example.

Input (Words only)	**Output (Words with Tags)**
Indeed	Indeed O
,	, O
Kaspersky	Kaspersky B-SecTeam
started	started O
tracking	tracking O
the	the O
BlueNoroff	BlueNoroff B-HackOrg
actor	actor O
a	a O
long	long O
time	time O
ago	ago O
.	. O

Table 1.1: Example of input and output for Named Entity Recognition in CTI

1.3 Objectives

The primary objectives of this research are to:

- **Improve Accuracy:** Enhance the precision, recall, and F1-score of NER systems by applying transformer-based models to the cybersecurity domain, ensuring more accurate identification and classification of cyber threat entities.
- **Explore Few-Shot Learning:** Investigate the efficacy of few-shot learning techniques with advanced transformer models like GPT-3.5 and GPT-4 to understand their potential in improving NER tasks with minimal training data. This approach is particularly significant for rapidly evolving threats where acquiring large labeled datasets is impractical.
- **Reduce Manual Effort:** Minimize the need for manual tagging and intervention by developing robust models that can effectively automate the extraction and categorization of entities from large volumes of unstructured text data.

- **Advance CTI Processes:** Contribute to the advancement of Cyber Threat Intelligence methodologies through the integration of state-of-the-art machine learning techniques, thus improving the capability to anticipate and respond to cyber threats.

Each of these objectives aims to address specific challenges in the field of cybersecurity, leveraging the latest advancements in NLP and AI to enhance the overall security posture of organizations.

1.4 Contributions

This research makes several significant contributions to the field of Cyber Threat Intelligence and Named Entity Recognition by employing transformer-based models. The contributions of this study are articulated as follows:

- **Advancement in NER Techniques:** We have successfully demonstrated how transformer-based models can significantly improve the accuracy, precision, and recall of Named Entity Recognition tasks within the cybersecurity domain, achieving SOTA performance outperforming traditional machine learning methods.
- **Application of Few-Shot Learning:** Our research explores the application of few-shot learning techniques in the context of CTI, showcasing how they can be utilized to quickly adapt to new and evolving threats with limited examples, reducing the need for extensive labeled datasets.
- **Automation in CTI Processes:** By automating the extraction and categorization of critical threat data, our models facilitate more efficient and timely threat intelligence processes, enabling cybersecurity teams to respond more swiftly and effectively to emerging threats.

1.5 Organization of the Thesis

Chapter 2

Background and Literature Review

In this chapter we provide the basic understanding of the terminologies needed to understand our work and existing works in this field. Based on the existing works, we provide a summary of the important results and gap analysis that helped to define our work.

2.1 Terminologies

- **Natural Language Processing (NLP):** Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on the interaction between computers and humans through natural language. The goal of NLP is to enable computers to understand, interpret, and produce human language in a valuable way. NLP encompasses both the study of theoretical aspects of language and the development of practical algorithms and systems for text and speech processing.
- **Cyber Threat Intelligence (CTI):** Cyber Threat Intelligence involves the collection, analysis, and dissemination of information about current and potential attacks that threaten the security of information systems. It is strategic, tactical, and operational. CTI helps organizations identify, mitigate, and respond to vulnerabilities and threats, including those posed by malware, phishing, and advanced persistent threats (APTs).
- **Transformer Models:** Transformer models are a class of deep learning models designed to handle sequenced data, like text, without requiring the sequence data to be processed in order. Transformers use self-attention mechanisms to weigh the importance of each word in the input data irrespective of its position in the sequence. This allows the model to learn the context of a word in relation to all other words in the data, significantly improving the model's ability to make predictions about the text.
- **Fine Tuning:** Fine tuning is a process where a pre-trained deep learning model is adapted

to a new but related problem. By starting with a model pre-trained on a large dataset, fine tuning adjusts the model parameters slightly to cater to the specific characteristics of a targeted task or dataset. This is commonly used in NLP to adapt general language understanding models to specific tasks like sentiment analysis or question answering.

- **Supervised Fine Tuning:** This refers to fine-tuning a model under a supervised learning setting, where the model is trained using a labeled dataset that provides explicit feedback on the accuracy of the model's predictions. This method ensures that the model adjustments are guided by direct examples of the desired output, optimizing the model's performance for specific tasks.

- **Few-Shot Learning:** Few-shot learning is a training strategy designed to enable a model to learn to recognize new patterns, objects, or concepts from a very small amount of data, typically only a few examples. This approach is particularly useful in situations where collecting large annotated datasets is not feasible.

- **Zero-Shot Learning:** Zero-shot learning is a technique where a model learns to correctly make predictions on tasks it has not been explicitly trained to perform. This involves the model understanding a task based on its inherent knowledge and reasoning skills, often leveraging detailed descriptions or relationships of categories.

- **Transfer Learning:** Transfer learning involves taking a model that has been trained on one task and repurposing it for a second, related task. This approach is effective in leveraging the learned features and knowledge from the first task to improve learning efficiency and predictive performance on the second task.

- **Large Language Models (LLMs):** A large language model (LLM) is a computational model notable for its ability to achieve general-purpose language generation and other natural language processing tasks such as classification. Based on language models, LLMs acquire these abilities by learning statistical relationships from vast amounts of text during a computationally intensive self-supervised and semi-supervised training process. LLMs can be used for text generation, a form of generative AI, by taking an input text and repeatedly predicting the next token or word.

- **Attention Mechanism:** An attention mechanism allows a model to dynamically focus on different parts of the input sequence as needed, enhancing the ability of the model to handle long sequences and to improve the relevance of the model's output to the input. The mechanism provides a weighted sum of all the input states, with the highest weights given to the most relevant parts of the input.

- **Tokenization:** Tokenization is the process of breaking down a stream of textual information into smaller components, or tokens, often single words or phrases. In NLP, tok-

enization is used as a pre-processing step to convert text into a format that can be easily analyzed and processed by algorithms.

These definitions aim to clarify the advanced concepts utilized in this research, facilitating a deeper understanding of the methodologies employed and their implications in the fields of cybersecurity and NLP.

2.2 Related Works

The development of Named Entity Recognition (NER) within Cyber Threat Intelligence (CTI) has been significantly advanced by recent research efforts, which can be categorized into Traditional NLP-Based Approaches and Transformer-Based Approaches. This section discusses key contributions from notable studies, highlighting their methodologies and findings.

Traditional NLP-Based Approaches

Introduction of DNRTI by Wang et al.

Wang et al. [1] introduces a comprehensive dataset tailored for enhancing Named Entity Recognition (NER) within the cybersecurity domain. This work addresses the unique challenges of network security, where data not only possesses highly specialized vocabulary but also involves sensitive information. To construct this dataset, over 300 threat intelligence reports were meticulously compiled from leading global security companies and government agencies. These reports have been expertly annotated across 13 distinct categories, culminating in a fully annotated corpus containing 175,220 words [1].

To validate the efficacy of NER models on this dataset, the authors employ advanced deep learning methods, particularly LSTM and BiLSTM models, which have shown promising results in text sequence modeling [2, 3]. The performance of these models is demonstrated through their ability to accurately categorize and extract pertinent entities, thereby showcasing the dataset's capability to encapsulate crucial information relevant to threat intelligence [1].

The DNRTI dataset is designed not only to bridge the gap in domain-specific datasets within cybersecurity but also to set a benchmark for future research in NER by providing a scalable and detailed dataset that mirrors the complexities of real-world cyber threats. Moreover, this dataset has been made available on GitHub, promoting accessibility and collaborative enhancement among researchers and developers aiming to advance cybersecurity solutions [4].

This substantial contribution by Wang et al. significantly enriches the resources available for

cybersecurity research, enabling a deeper understanding of cyber threats and fostering the development of more effective security measures.

Semantic Augmentation by Peipei Liu et al.

Liu et al. [5] present an innovative approach to Named Entity Recognition (NER) in Cyber Threat Intelligence (CTI) that focuses on addressing the challenges posed by data sparsity and variability in cybersecurity text sources. Recognizing the limitations of traditional NER methods, the authors propose a semantic augmentation technique that enriches token representations with a combination of linguistic features to enhance the model's accuracy and robustness.

This technique incorporates constituent, morphological, and part-of-speech features for each token, which are crucial for understanding the context and grammatical structure of language in cybersecurity documents. Additionally, Liu et al. employ a dual corpus strategy, where semantic information is drawn from both a domain-specific cybersecurity corpus and a general large-scale corpus. This dual-source approach ensures that the model benefits from both specialized knowledge and broad linguistic patterns, thereby improving its capability to identify and classify complex cybersecurity-related entities.

The effectiveness of this method was validated using the DNRTI [1] and MalwareTextDB [6] datasets, where it demonstrated significant improvements in entity recognition performance compared to baseline models. The results from this study not only illustrate the potential of semantic augmentation in enhancing NER systems but also pave the way for future research into combining multiple sources of semantic information to tackle challenging NER tasks within specialized domains such as cybersecurity.

Figure 2.1 demonstrates the proposed architecture of the semantic augmentation technique described in the study.

BERT-CRF Integration by Sheng-Shan Chen et al.

Enhanced Named Entity Recognition in CTI Using BERT-CRF by Sheng-Shan Chen et al.

Sheng-Shan Chen et al. [7] introduce an innovative approach to Named Entity Recognition (NER) in Cyber Threat Intelligence (CTI) through a streamlined BERT-CRF model, which excludes the BiLSTM layer typically found in the traditional BERT-BiLSTM-CRF architecture. This modification significantly reduces computational demands and enhances performance by directly channeling the robust token embeddings from BERT into a Conditional Random Field (CRF) layer for efficient NER [7].

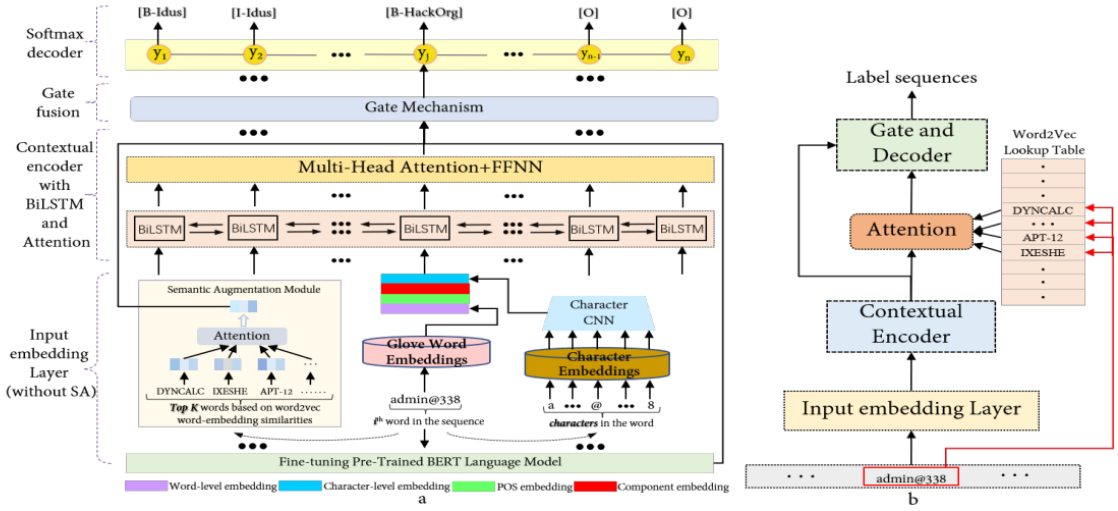

Figure 2.1: Illustrative diagram of the semantic augmentation technique employed in the study.

The model's efficacy was rigorously validated using three publicly available threat entity databases, alongside recent open-source threat intelligence data, ensuring its relevance and applicability in real-world settings. The study highlights the model's superiority by comparing its performance with that of widely recognized models like GPT-3.5 [8] and DistilBERT [9], with notable results demonstrating its superior accuracy and robustness, especially in processing malware-specific threat intelligence where it achieved an impressive accuracy of 93.95% [7].

The researchers have made the implementation code for the BERT-CRF model publicly accessible, promoting further research and adaptation in the field. This work not only advances the technical methodologies for conducting NER within the cybersecurity domain but also significantly enhances the operational capabilities of security analysts by streamlining the analysis and response processes to cyber threats. The integration of advanced machine learning techniques such as BERT and CRF [10] into CTI platforms exemplifies a significant step forward in enhancing the effectiveness of cyber threat detection and analysis ([7]).

Transformer-Based Approaches

Application of Transformer Models by Evangelatos et al.

Evangelatos et al. [11] explore the efficacy of transformer-based models in extracting Cyber Threat Intelligence (CTI) from a variety of sources including online vulnerability databases, CERT feeds, and both surface and dark web content. Their study leverages the DNRTI dataset, which contains over 300 threat intelligence reports from open-source platforms, encompassing 175,220 annotated words across 13 classes [1]. They demonstrate that transformer-based models like BERT [12], XLNet [13], RoBERTa [14], and ELECTRA [15] significantly outper-

form previous NER systems used in CTI, providing a substantial improvement in identifying cybersecurity-related named entities.

The research highlights the critical role of named entity recognition in enhancing the actionability of CTI by systematically classifying threat information into predefined categories. By employing advanced transformer architectures, the authors address the complex challenges posed by the sophisticated and evolving nature of threat actors and their tactics. The study confirms that these models not only achieve high accuracy but also adapt effectively to the nuanced demands of cybersecurity text analysis.

Their findings underscore the potential of using state-of-the-art NLP techniques to improve the automation and precision of CTI extraction processes, suggesting a promising direction for future research in cybersecurity informatics. The paper's insights into the deployment of transformer-based models within the CTI domain mark a significant advancement in the field, pushing the boundaries of what can be achieved with modern NLP technologies in security contexts.

APTNER Dataset Introduction by Wang et al.

Wang et al. introduce the APTNER dataset, a specialized resource designed to enhance Named Entity Recognition (NER) tasks within the Cyber Threat Intelligence (CTI) domain [16]. This dataset was meticulously constructed to address the unique challenges of NER in cybersecurity, offering finely annotated data that includes complex entity types specific to cyber threats, such as malware, vulnerabilities, and attack vectors.

The APTNER dataset comprises over 1,000 sentences, with more than 21 specific entity categories, including but not limited to malware, tools, and threat actor names. This dataset significantly extends the scope of traditional NER datasets by focusing on entities prevalent in CTI, thereby providing a more relevant training and evaluation ground for models aimed at extracting actionable intelligence from cyber threat reports [16].

In their methodology, Wang et al. employ a combination of traditional NLP techniques and advanced machine learning models to validate the dataset's effectiveness. They particularly highlight the use of transformer-based models such as BERT, which have shown remarkable success in other NER domains but are yet to be fully explored in CTI-specific contexts. Their experimental results indicate that models pre-trained on this targeted dataset can outperform those trained on more general corpora when it comes to identifying the nuanced and specialized language used in cyber threat intelligence [12].

Furthermore, the study compares the APTNER dataset with other renowned datasets like CoNLL-2003 [17] and shows that APTNER is not only larger in terms of entity variety but also in the complexity and relevance of its annotated entities. A significant aspect of APTNER is its ad-

herence to the STIX (Structured Threat Information eXpression) format, which facilitates the standardized communication of threat information and enhances the interoperability between systems [18]. This makes it an invaluable resource for researchers and practitioners aiming to develop more effective NER systems for cybersecurity applications.

The creation of APTNER marks a significant advancement in the resources available for cybersecurity NER, enabling more focused and accurate development of CTI capabilities. This dataset not only enhances the training and benchmarking of NER models but also contributes to the broader efforts of automating the extraction of actionable intelligence from an ever-growing volume of cyber threat data.

Additional Literature Review involving entity and event recognition outside of DNRTI and APTNER datasets

While the DNRTI and APTNER datasets represent significant advancements in the development of resources tailored for NER tasks in the cybersecurity domain, research in this field extends beyond these two key datasets. Numerous other studies have contributed to the refinement and expansion of NER capabilities, addressing various challenges associated with the complexity of cyber threat data and the need for precise threat detection and classification.

These contributions often explore different methodologies, leverage various data sources, and introduce innovative techniques aimed at enhancing the accuracy and efficiency of entity recognition in diverse cybersecurity contexts. From utilizing advanced machine learning algorithms to integrating rich linguistic and contextual information, these works collectively push the boundaries of what is possible in the automated processing of cyber threat intelligence.

The following sections will highlight some of these pivotal studies, emphasizing their unique approaches and the impact they have had on the field. Each study contributes a piece to the evolving puzzle of cybersecurity threat intelligence, providing insights that help improve systems designed to protect digital infrastructures and sensitive information against increasingly sophisticated cyber threats.

Cyber Threat Intelligence Collection from Hacker Forums by Deliu et al.

Deliu, Leichter, and Franke [19] address the critical challenge of keeping pace with the rapid evolution of the cyber threat landscape in their groundbreaking study. Traditional security measures such as firewalls, antivirus software, and Intrusion Detection Systems (IDS) often fall short in combatting the sophisticated tactics employed by modern cyber adversaries. This paper introduces a novel method for leveraging non-traditional information sources like hacker forums and dark-web platforms to enhance Cyber Threat Intelligence (CTI).

The authors propose a two-stage hybrid machine learning model that automates the extraction of CTI from these platforms. The first stage of the model uses Support Vector Machines (SVM) [20] to identify forum posts that are relevant to cybersecurity threats. Following the identification phase, the second stage applies Latent Dirichlet Allocation (LDA) [21] to cluster these posts based on the discussion topics, providing insights into prevailing and emerging cyber threats.

By applying this method to real data from hacker forums, the study demonstrates its effectiveness in automatically extracting crucial information on threats such as leaked credentials, malicious proxy servers, and malware that can evade traditional antivirus detection. The results indicate that this method is not only efficient in processing vast amounts of unstructured data but also significantly enhances the ability of security teams to identify and respond to cybersecurity threats rapidly.

This innovative approach allows for the integration of extracted CTI with existing security controls, significantly boosting their effectiveness in real-time threat detection and response. By automating the CTI extraction process, Deliu et al. [19] contribute significantly to reducing the time, errors, and resources typically associated with manual threat intelligence analysis.

Cyberthreat Detection from Twitter by Dionísio et al.

Dionísio et al. explore the application of deep neural networks for detecting cyber threats from Twitter data, a novel approach that addresses the limitations of traditional security systems such as firewalls and antivirus software [22]. Their methodology employs a two-stage deep learning framework, utilizing a Convolutional Neural Network (CNN) [23] to first identify tweets with potential security information and a Bidirectional Long Short-Term Memory (BiLSTM) network for Named Entity Recognition (NER). This dual-model approach is designed to enhance the precision and efficiency of extracting meaningful cyber threat intelligence from the vast stream of Twitter feeds [24].

In the initial stage, the CNN model filters out irrelevant tweets, focusing only on those that potentially include security-related content. This preprocessing step is crucial for optimizing the performance of the subsequent NER task, which is handled by the BiLSTM network. The BiLSTM effectively tags critical entities such as malware types, attack vectors, and targets, which are vital for forming actionable security alerts or indicators of compromise (IoC). Dionísio et al. validate their model's efficacy through extensive testing on real-world data, demonstrating its ability to not only detect relevant tweets with high accuracy but also to perform entity recognition with notable precision [22].

Their findings highlight the significant potential of leveraging social media platforms like Twitter as dynamic sources of cyber threat intelligence. By automating the detection and extraction

of threat indicators, the proposed model provides a timely and efficient tool for cybersecurity teams to enhance their monitoring and response strategies. This research marks a substantial contribution to the field of cybersecurity, proposing a scalable solution that could be further refined and adapted to broader applications within the threat intelligence landscape.

Cybersecurity Event Information Extraction by Satyapanich et al.

Satyapanich, Ferraro, and Finin present "CASIE: Extracting Cybersecurity Event Information from Text," a significant contribution to the domain of cybersecurity [25]. This paper introduces CASIE, a system specifically developed to extract structured cybersecurity event information from unstructured textual data to populate a knowledge graph aimed at supporting security incident response and analysis. The system is trained on a newly developed corpus of 1,000 English news articles spanning from 2017 to 2019, meticulously annotated with detailed event-based information concerning cyberattacks and security vulnerabilities.

Utilizing advanced neural network architectures, including those with attention mechanisms, CASIE effectively identifies and extracts detailed aspects of cybersecurity events, such as event subtypes and their semantic roles. The results demonstrate the system's capability to not only recognize different event components but also accurately associate relevant attributes and actors involved in cybersecurity incidents. This capability enhances the granularity and utility of cybersecurity knowledge bases, facilitating more informed and rapid response strategies to emerging threats. The development and successful application of CASIE underscore the growing importance of integrating machine learning techniques with cybersecurity informatics to advance the state of automated threat intelligence and situational awareness.

Graph Convolutional Networks for Cybersecurity NER by Fang et al.

Fang, Zhang, and Huang introduce a novel entity recognition model, *CyberEyes*, which utilizes Graph Convolutional Networks (GCNs) [27] to enhance Named Entity Recognition (NER) in the cybersecurity domain [26]. The model is designed to capture complex, non-local dependencies typical in cybersecurity texts, a challenge that traditional sequential models like CNN-BiLSTM-CRF [28] fail to address effectively. By leveraging the structural advantages of GCNs, CyberEyes significantly improves the recognition of cybersecurity entities, achieving an F1 score of 90.28%, which outperforms traditional models.

This study highlights the effectiveness of graph-based neural networks in processing cybersecurity texts, where the context and relationships between entities are critical for accurate entity recognition. The success of CyberEyes suggests promising directions for future research in applying graph neural networks to enhance NER systems within specialized fields such as cy-

bersecurity.

2.3 Summary of Existing Works

The landscape of Named Entity Recognition (NER) within Cyber Threat Intelligence (CTI) is rich and varied, featuring a range of approaches that reflect the evolving complexity of cybersecurity threats. The following table 2.1 provides an overview of seminal works on the two datasets DNRTI and APTNER used in our work, summarizing the datasets employed, input and output specifics, the classifiers used, and the notable findings or reported accuracies. This summary highlights the diversity and progression of methodologies from traditional NLP techniques to advanced transformer-based models, illustrating how each contributes uniquely to enhancing the precision and efficiency of threat detection and analysis in cybersecurity.

Table 2.1: Summary of Existing Works

Ref.	Dataset	Input	Output	Classifier	Accuracy %
[1]	DNRTI	Threat reports	NER entities	LSTM, BiLSTM	90.85%
[5]	DNRTI, MalwareTextDB	Cybersecurity texts	NER entities	Semantic augmentation	95.43 (DNRTI), 87.99 (MalwareTextDB)
[7]	Public threat databases	Threat data	NER entities	BERT-CRF	93.95%
[11]	DNRTI	Various sources	Cybersecurity entities	Transformer models (BERT, XLNet, RoBERTa, ELECTRA)	Not mentioned
[16]	APTNER	Cyber threat reports	Specific CTI entities	Traditional and advanced ML models	Not mentioned

Chapter 3

Dataset Description

This chapter provides an exhaustive description of the primary datasets utilized in this thesis to advance Named Entity Recognition (NER) in the domain of Cyber Threat Intelligence (CTI). The detailed exploration of these datasets, namely the Dataset for Named Entity Recognition in Threat Intelligence (DNRTI) [1] and APTNER [16], underscores their pivotal role in providing a robust basis for testing and refining NER techniques specifically tailored to cybersecurity contexts.

3.1 DNRTI Dataset

3.1.1 Overview

The DNRTI dataset, introduced by Xuren Wang et al. [1], is specifically developed to bolster NER research within the network security domain, addressing the need for specialized datasets that are rich in cybersecurity context.

3.1.2 Data Collection and Composition

This dataset comprises over 300 threat intelligence reports sourced from renowned security companies, government agencies, and open-source intelligence platforms, totaling 175,220 words [1]. Each report has been carefully selected for its relevance and richness in cybersecurity content.

3.1.3 Categories and Annotation

Thirteen distinct categories such as HackOrg, Tool, and SecTeam are annotated within DNRTI using the BIO labeling scheme. This rigorous annotation process, facilitated by the Brat Rapid Annotation Tool, ensures detailed and accurate identification of cybersecurity entities [29].

3.1.4 Dataset Splits and Statistics

The DNRTI dataset is divided into training (70%), validation (15%), and testing (15%) sets, featuring a diverse range of entities crucial for evaluating NER models in real-world scenarios.

3.1.5 Significance and Impact

DNRTI is notable for its extensive variety and significant word count, playing a crucial role in advancing NER research within network security and setting a robust benchmark for future developments [1].

Figure 3.1 shows the distribution of entity categories in the DNRTI dataset.

3.2 APTNER Dataset

3.2.1 Overview

The APTNER dataset, developed for enhancing NER tasks within the CTI domain, adheres to the STIX2.1 standard and includes a wide range of CTI-specific entity types [16].

3.2.2 Dataset Creation and Composition

Comprising 10,984 sentences and 260,134 tokens, APTNER was developed through meticulous manual annotation of Advanced Persistent Threat (APT) reports, covering 21 categories such as Malware and HackOrg [16].

3.2.3 Annotation Process

The annotation was initially conducted by trained graduate students, followed by validation from CTI professionals, ensuring high accuracy and relevance of the data [29].

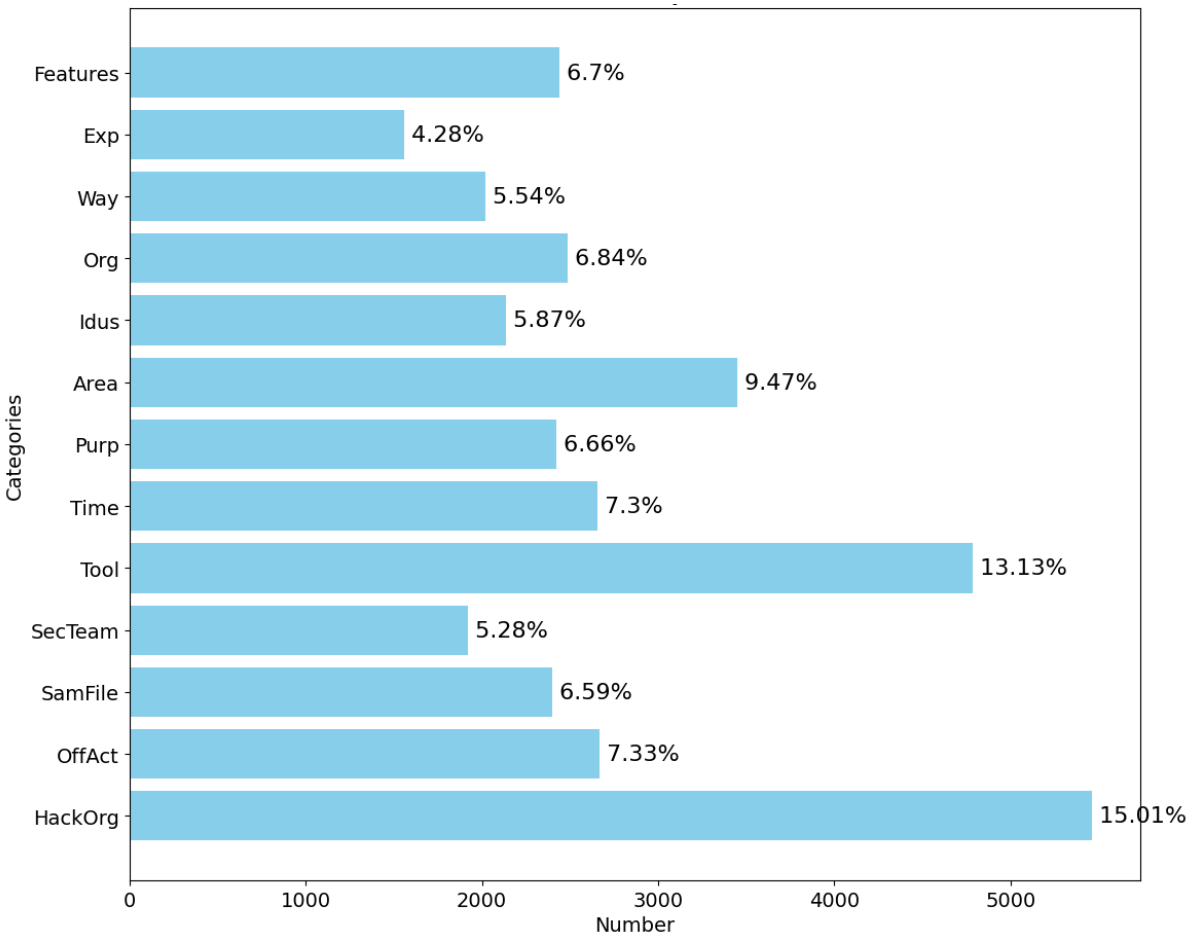

Figure 3.1: Distribution of entity categories in the DNRTI dataset, illustrating the count and percentage of annotations per category.

3.2.4 Data Format and Standards

APTNER employs the BIOES tagging scheme and is compliant with the CoNLL2002 specification, ensuring detailed and structured entity recognition suitable for CTI analysis [31].

3.2.5 Dataset Split and Utilization

The dataset is structured into training, development, and test phases with a distribution ratio of approximately 7:1.5:1.5, following best practices for dataset segmentation [17].

3.2.6 Significance and Impact

APTNER's compliance with the STIX2.1 standard significantly enhances its utility for CTI applications and positions it as a crucial resource for developing NER solutions tailored to cybersecurity [30].

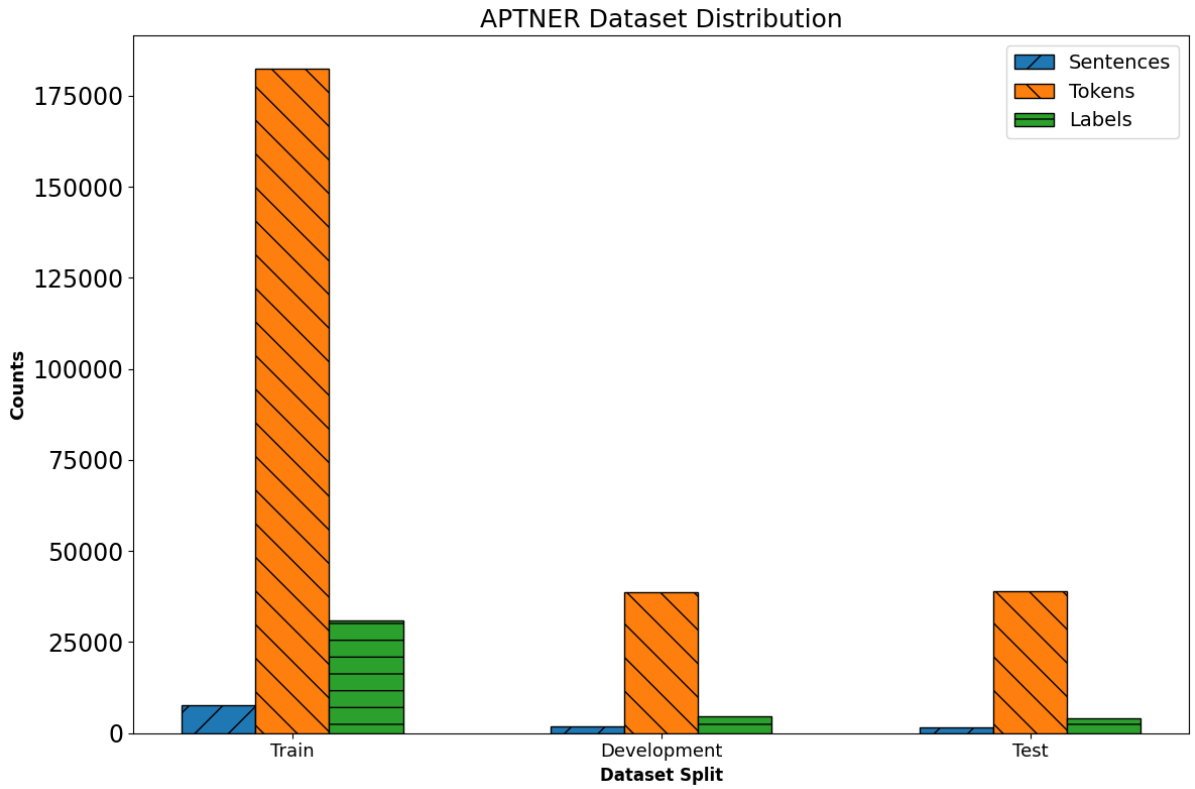

Figure 3.2: Distribution of sentences, tokens, and labels across the training, development, and testing splits of the APTNER dataset, highlighting the comprehensive nature of its annotations.

Figure 3.2 illustrates the distribution of sentences, tokens, and labels across the different segments of the dataset—training, development, and testing. This visualization emphasizes the structured approach in dataset segmentation, which is critical for the rigorous training and evaluation of Named Entity Recognition models. The balanced distribution ensures that models trained on this dataset are tested against a variety of scenarios, mirroring real-world applications.

Figure 3.3 offers an in-depth visualization of the entity types categorized within the APTNER dataset. This figure effectively demonstrates the dataset's specialization in cybersecurity, meticulously detailing the variety and prevalence of each entity type. Notably, it includes critical cybersecurity elements such as malware names, attack vectors, threat actor identifiers, and cybersecurity tools, among others. Each category is represented proportionally, providing a clear picture of the dataset's focus areas and the relative frequency of each type of cybersecurity-related entity.

This comprehensive breakdown is crucial for researchers and developers as it illustrates the dataset's alignment with real-world cybersecurity challenges. It highlights the APTNER dataset's strength in providing a robust foundation for training sophisticated Named Entity Recognition (NER) systems. These systems are designed to recognize and categorize intricate and often technically complex entities, which are vital for constructing effective Cyber Threat Intelli-

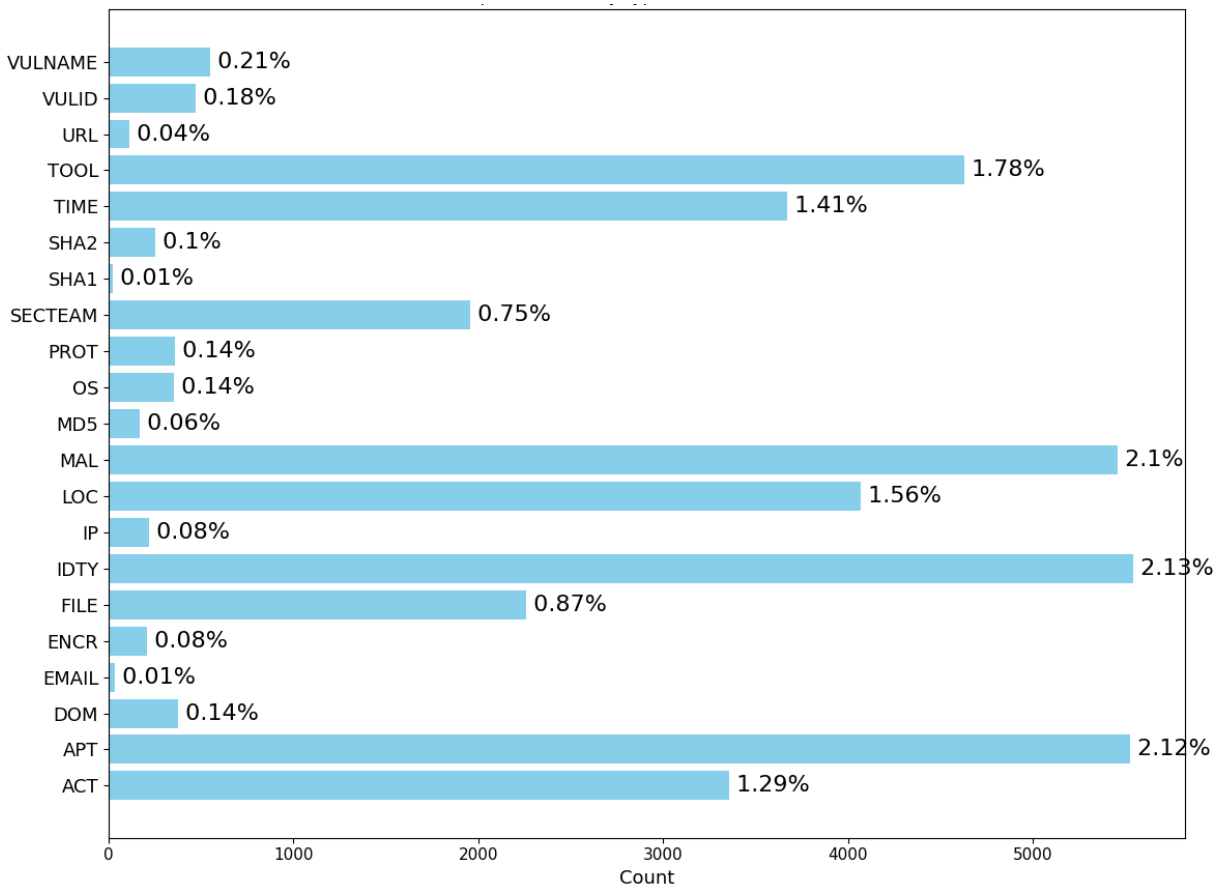

Figure 3.3: Proportions of entity types within the APTNER dataset, showcasing the specific focus on cybersecurity-related entities.

gence (CTI) tools. By training NER models on such a well-annotated dataset, researchers can significantly improve the models' accuracy and reliability in identifying and classifying nuanced cybersecurity threats and events.

Moreover, the detailed portrayal of entity types in Figure 3.3 not only validates the meticulous efforts put into the dataset's creation but also serves as a vital educational tool. It aids both new and seasoned practitioners in cybersecurity and machine learning by providing a clear understanding of the kinds of data that are essential for effective threat intelligence.

Chapter 4

Methodology

This research employs a robust methodology utilizing the advanced capabilities of the Hugging Face's Transformers library, renowned for its suite of pre-trained models that are pivotal in deep learning applications. Our approach is specifically tailored for enhancing Named Entity Recognition (NER) in the realm of Cyber Threat Intelligence (CTI). The methodology is designed to harness these sophisticated models to capture nuanced contextual information, thereby improving the accuracy of entity detection in the challenging CTI domain.

4.1 Dataset Preprocessing

The primary datasets employed in this study are the APTNER and DNRTI datasets, each presenting unique challenges and opportunities for deep learning applications. Initially, these datasets are formatted in a straightforward text file where each line pairs a word with its corresponding entity tag, separated by a space. Sentences are separated by blank lines, indicating the end of one sequence and the beginning of another.

However, the direct use of this format in training poses significant challenges. Each word treated as a separate input can detract from the model's ability to understand contextual dependencies. Therefore, it is essential to process the input at the sentence level, allowing for a comprehensive analysis by the deep learning models. This approach enables the generation of more nuanced embeddings and significantly enhances the accuracy of the resulting NER model.

To facilitate this, the datasets undergo a transformation process where they are converted from the line-by-line text files to a more structured CSV format. This format consists of two columns: 'Sentence' and 'Label'. The 'Sentence' column compiles the words into coherent sentences, and the 'Label' column sequentially lists the entity tags corresponding to the words in the 'Sentence' column. This restructuring is crucial for the deep learning models to perform effectively, enabling them to process entire sentences and thereby maintain contextual integrity.

Table 4.1 shows the default format of the dataset files. The table contains two sentences demarcated by a blank line which are taken from the APTNER dataset, it is seen here that each line contains two space seperated words, the first one being the word of the sentence and the second being the entity label.

Table 4.1: Example of the default data format.

A O
journey O
to O
Zebrocy S-MAL
land O
. O
Morphisec S-SECTEAM
is O
not O
revealing O
these O
names O
. O

Table 4.2: Example of the preprocessed data format.

ID	Sentence	Label
1	”A”, ”journey”, ”to”, ”Zebrocy”, ”land”, ”.”	85, 85, 85, 27, 85, 85
2	”Morphisec”, ”is”, ”not”, ”revealing”, ”these”, ”names”, ”.”	22, 85, 85, 85, 85, 85, 85

Table 4.2 delineates the restructured CSV format ensuing from the preprocessing of the datasets. Each entry is accorded a unique identifier, and the 'Sentence' field enumerates the individual words of the text, segregated by commas. Concurrently, the 'Label' column enumerates the corresponding entity tags, which have been mapped to numeric identifiers. This reformatted representation is specifically designed to be conducive to processing by sophisticated transformer-based deep learning models. By receiving entire sentences as input, the models are enabled to harness the contextual nuance embedded within the linguistic sequence, thereby enhancing the precision of the entity recognition outcomes

The conversion to a structured CSV format is a critical step in preparing the data for efficient processing by the transformer-based models. This format allows the models to utilize their full capability to analyze the contextual nuances embedded within the linguistic sequences, enhancing the precision and effectiveness of the entity recognition processes.

4.2 Input and Output Formatting

The structuring of input and output data is paramount for the efficient functioning of NER models within CTI. The input data for our models is sourced from the preprocessed CSV files, which are formatted to reflect the complex structure of sentences and their corresponding entity annotations.

4.2.1 Input Data Structure

The 'Sentence' column in our preprocessed dataset includes sequences of words that represent individual sentences, each separated by commas. This structure allows transformer-based models to perform contextual analysis effectively. The 'Labels' column presents a sequence of entity tags numerically encoded to correspond with each token in the 'Sentence' column, ensuring precise alignment during the training process.

4.2.2 Output Data Structure

The output from our models consists of a sequence of entity tags corresponding to each token in the input sentence. This sequence directly correlates with the input sequence, providing a detailed and context-aware mapping of entities within the text. Such a detailed output structure is vital for subsequent analysis and application in cybersecurity threat intelligence tasks.

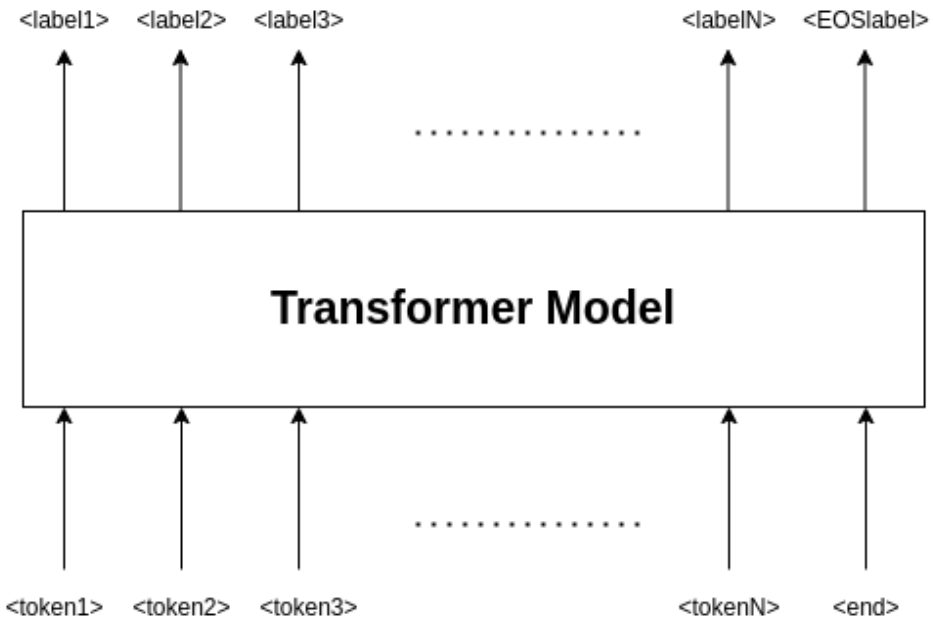

Figure 4.1: Illustrative architecture of the transformer model for NER tasks, detailing the process from tokenized input to contextualized output.

Figure 4.1 depicts the model's architecture, showing how the input sequence of tokens is processed to predict a corresponding sequence of entity labels, emphasizing the model's ability to interpret and analyze text contextually.

4.3 Tokenization and Model Alignment

Tokenization plays a crucial role in adapting raw text into a format that is amenable to processing by transformer-based models. We utilize the Hugging Face's `BertTokenizerFast`, which supports efficient tokenization and alignment of entity labels to tokens. This step is essential for maintaining the fidelity of the entity annotations post-tokenization, especially considering the use of subword tokenization algorithms that can split words into smaller units.

4.4 Fine-Tuning and Model Training

The fine-tuning process involves adjusting pre-trained models to the specific nuances of our cybersecurity datasets. We utilize models such as `bert-base-uncased` and `bert-base-cased` from the Hugging Face library, adapting them to the unique requirements of the NER task by training on annotated CTI data. This process involves multiple iterations of training and validation to optimize model performance and ensure robustness against overfitting.

The methodology outlined in this chapter represents a comprehensive approach to implementing advanced NER techniques within the domain of Cyber Threat Intelligence, leveraging cutting-edge technologies to enhance the accuracy and efficiency of threat detection and analysis.

Chapter 5

Experimental Framework and Hyperparameter Optimization

This chapter outlines the setup and execution of the experimental framework designed to evaluate the performance of our proposed Named Entity Recognition (NER) models within the context of Cyber Threat Intelligence (CTI). It includes a detailed description of the hyperparameter optimization techniques utilized to fine-tune the models, ensuring optimal performance on the NER tasks using the DNRTI and APTNER datasets.

5.1 Experimental Setup

The evaluation of our NER models was conducted with a rigorous experimental framework, ensuring that the results are both robust and reproducible. This section elaborates on the dataset specifications, preprocessing routines, model configurations, and the training environment. It also covers the evaluation protocol, including the metrics used to assess model performance, thereby providing a comprehensive basis for validating our research findings.

The computational experiments were carried out on the Google Colab Pro platform [35], which provides GPU support to enhance computational efficiency. This setup is supplemented by 32 GB of virtual RAM, dedicated to supporting the intensive demands of training deep learning models. The models were developed and implemented using the PyTorch library [36], renowned for its dynamic computation graph and effectiveness in handling model training and evaluation tasks. PyTorch's versatile framework facilitates the construction of complex models and supports efficient operations on tensors.

5.2 Parameter Setting

Our experiments involved a range of transformer-based models, such as bert-base-uncased, bert-base-cased, and their respective NER-specific variants. Each model's hyperparameters were meticulously tuned to derive the best settings. The optimal hyperparameter configuration that led to the highest performance across both the DNRTI and APTNER datasets is summarized in the following table.

Parameter	**Value**
Evaluation Strategy	epoch
Learning Rate	2e-5
Training Batch Size (per device)	8
Evaluation Batch Size (per device)	8
Number of Training Epochs	10
Weight Decay	0.01
Logging Steps	50

Table 5.1: Hyperparameter settings for the best-performing models on DNRTI and APTNER datasets.

As detailed in Table 5.1, the models were trained under an evaluation strategy that periodically assessed performance at the end of each epoch. The learning rate was strategically set to 2e-5 to strike an optimal balance between the convergence speed and the stability of the learning process. The batch sizes for both training and evaluation were configured at 8 to maximize the use of available memory while ensuring accurate gradient estimation. A total of 10 training epochs were chosen to adequately train the models without risking overfitting, complemented by a weight decay of 0.01 to help regulate and prevent the overfitting of model parameters. Logging was set to trigger every 50 steps to provide timely updates on model performance, which is crucial for early diagnostics and adjustments.

This chapter not only illustrates the methodical approach taken to ensure the effective evaluation of our models but also highlights the specific configurations that contributed to their success in identifying named entities within cybersecurity texts accurately.

Chapter 6

Results

In this section, we present the empirical results obtained from the application of our trained models on the DNRTI and APTNER datasets. The performance of each model is rigorously evaluated based on the precision, recall, F1 score, and accuracy metrics previously described. These results offer insights into the models' ability to effectively identify and classify named entities within the specialized context of Cyber Threat Intelligence (CTI). The findings are crucial for understanding the strengths and weaknesses of the various transformer architectures under investigation and for identifying the most promising approaches for further development and refinement.

6.1 Evaluation Metrics

To quantitatively assess the efficacy of our named entity recognition models, we employed several performance metrics. These metrics are essential for understanding various aspects of the model's predictive power and are commonly used in classification tasks. The definitions and formulas for each metric are as follows:

Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positives. It is a measure of a classifier's exactness. Higher precision relates to a lower false positive rate. The precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6.1}$$

where TP represents true positives and FP stands for false positives.

Recall

Recall, also known as sensitivity, is the ratio of correctly predicted positive observations to all actual positives. It is a measure of a classifier's completeness. The recall is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6.2}$$

where FN represents false negatives.

F1 Score

The F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is particularly useful when the class distribution is uneven. The F1 score is defined as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.3}$$

Accuracy

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observations to the total observations. The accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{6.4}$$

where TN represents true negatives.

Macro and Micro Averaging

When dealing with multi-label classification tasks, it's essential to consider both macro and micro averaging for metrics.

Macro-average computes the metric independently for each class and then takes the average (hence treating all classes equally), while **micro-average** aggregates the contributions of all classes to compute the average metric. In a multi-class classification setup, micro-average is preferable if there is a class imbalance.

6.2 Main experiments and analysis

We train an array of BERT models,namely bert-base-uncased,bert-base-cased,bert-base-NER,bert-large-NER,bert-large-cased etc on our two datasets, DNRTI and APTNER. In our experimental analysis, we evaluate several models to determine their effectiveness in NER tasks on the validation and test sets. We also include two specialized models, namely SecureBERT-DNRTI [38] and SecureBERT-NER [38] in our evaluation, these two models are fine tuned specifically for the DNRTI and APTNER datasets respectively and provided by the Huggingface platform. Table 6.1 summarizes the performance metrics obtained for each model on both the datasets DNRTI and APTNER.

6.2.1 Experiments on DNRTI

Table 6.1 illustrates the performance of various BERT models across the validation and test datasets of DNRTI and APTNER. The observed performance differences underscore the influence of model architecture and training configurations on Named Entity Recognition (NER) tasks.

In the DNRTI dataset, the BERT-Large-NER model demonstrated exceptional performance, achieving the highest precision (0.791), recall (0.826), and F1 score (0.809) on the validation set. This was closely followed by the BERT-Base-NER model, which exhibited similar precision and recall metrics. The BERT-Large-NER model recorded an accuracy of 90.8% on the validation set. On the test set, BERT-Large-NER continued to excel, achieving top metrics across all categories with a precision of 0.831, recall of 0.869, F1 score of 0.852, and accuracy of 93.9%. These findings suggest that larger models are more adept at generalizing from training data to unseen data, particularly in complex NER tasks, due to their increased number of parameters compared to smaller, simpler BERT models. The BERT-Base-Cased and BERT-Large-Cased models showed slightly lower performance relative to their NER-optimized counterparts, underscoring the benefits of NER-specific pre-training. The BERT-base models, both uncased and cased, displayed moderate performance. The BERT-Base-Uncased model achieved a precision of 0.771 on the validation set and 0.809 on the test set, with recall rates of 0.821 and 0.851, respectively. These results highlight a consistent performance, although not as high as the more specialized models. The BERT-Base-Cased variant showed slight improvements in recall, likely due to its sensitivity to case variations, which are crucial for accurately identifying case-sensitive named entities. Specifically, the BERT-Base-Cased model recorded a recall of 0.812 on the validation set and 0.840 on the test set, with precision figures of 0.763 and 0.799, respectively.

However, the BERT-Base-NER model, designed specifically for NER tasks, demonstrated improved recall rates, achieving 0.824 on the validation set and 0.838 on the test set. This enhance-

ment highlights the effectiveness of NER-focused training. However, its overall performance metrics, while competitive, did not surpass those of the larger models. This finding illustrates that increased model complexity and size, such as those found in BERT-Large models, can significantly impact NER capabilities.

Lastly, the SecureBERT-DNRTI [38] significantly improved performance on the DNRTI dataset. This model, fine-tuned specifically for the DNRTI dataset, achieved a precision of 0.812, recall of 0.849, F1 score of 0.831, and accuracy of 93.2% on the validation set. On the test set, SecureBERT-DNRTI showed even more substantial improvements with a precision of 0.841, recall of 0.872, F1 score of 0.857, and accuracy of 95.2%. These results underscore SecureBERT-DNRTI's effectiveness in accurately identifying and classifying cybersecurity-related entities, demonstrating its robustness in managing the complex language specific to the cybersecurity domain.

In conclusion, the introduction of SecureBERT-DNRTI has markedly enhanced NER performance on the DNRTI dataset. The superior metrics achieved by SecureBERT-DNRTI emphasize its efficiency in advancing the automated extraction and classification of Cyber Threat Intelligence (CTI). This model showcases the potential of transformer-based architectures to improve the accuracy and efficiency of cybersecurity threat detection and analysis, providing promising avenues for future research and application in the field of CTI.

6.2.2 Experiments on APTNER

In our comprehensive evaluation of various BERT models on the APTNER dataset, distinct performance differences were observed, similar to the case of the DNRTI dataset. The superiority of the BERT-Large-NER model in handling Named Entity Recognition (NER) tasks was evident, consistently outperforming other models across all metrics. This model's robustness and enhanced capability to generalize from training to unseen data underscore its effectiveness in the NER domain.

Firstly, the BERT-Base models, both uncased and cased, exhibited moderate performance. The BERT-Base-Uncased model achieved a precision of 0.698 on the validation set and 0.704 on the test set, with recall rates of 0.726 and 0.716, respectively. These figures highlight a stable performance, though not as high as the more specialized models. The BERT-Base-Cased variant showed slight improvements in recall, possibly due to its sensitivity to case variations, which can be crucial for correctly identifying case-sensitive named entities. Specifically, the BERT-Base-Cased model recorded a recall of 0.724 on the validation set and 0.733 on the test set, with precision figures of 0.688 and 0.699, respectively.

Secondly, the BERT-Base-NER model, optimized specifically for NER tasks, demonstrated improved recall rates, with 0.753 on the validation set and 0.746 on the test set. This improvement

highlights the effectiveness of NER-focused training. However, its overall performance metrics, while competitive, did not surpass those of the larger models. This illustrates that increased model complexity and size, such as those found in BERT-Large models, can have a substantial impact on NER capabilities.

However, the BERT-Large-NER model excelled, achieving impressive precision (0.751 on the validation set and 0.757 on the test set), recall (0.786 on the validation set and 0.791 on the test set), and F1 scores (0.777 on the validation set and 0.779 on the test set). Additionally, it recorded high accuracy figures of 87.9% on the validation set and 88.0% on the test set. This performance underscores the advantage of its extensive network architecture, which includes more layers and attention heads, allowing for a deeper understanding of context and more nuanced extraction of entity information.

While the BERT-Large-Cased model slightly underperformed compared to the BERT-Large-NER, it still showed commendable results. The model's best metrics included a recall of 0.787 and an F1 score of 0.769 on the validation set, with a precision of 0.748 and accuracy of 86.6% on the test set. These results delineate the subtle yet significant impact of casing in conjunction with large model architectures, which can enhance the identification of named entities that are case-sensitive.

Above all these, the SecureBERT-APTNER [38] model, fine-tuned specifically on the APTNER dataset and provided by HuggingFace, further enhanced NER performance. On the validation set, SecureBERT-APTNER achieved a precision of 0.765, recall of 0.801, F1 score of 0.782, and accuracy of 88.4%. On the test set, it demonstrated a precision of 0.772, recall of 0.808, F1 score of 0.790, and accuracy of 88.2%. These results highlight SecureBERT-APTNER's effectiveness in navigating the specialized lexicon of cybersecurity threats and providing reliable and precise identification of relevant entities.

In summary, the introduction of SecureBERT-APTNER significantly enhanced the NER performance on the APTNER dataset. The superior metrics achieved by SecureBERT-APTNER emphasize its efficacy in advancing the automated extraction and classification of Cyber Threat Intelligence (CTI). This model demonstrates the potential of transformer-based architectures in improving the accuracy and efficiency of cybersecurity threat detection and analysis, offering promising directions for future research and application in the field of CTI.

6.3 Summary of Results

Our experiments on the DNRTI and APTNER datasets demonstrated significant advancements in Named Entity Recognition (NER) tasks using various BERT models, culminating in the achievement of state-of-the-art (SOTA) metrics.

Dataset	Model	Validation Set				Test Set			
		P	R	F1	Acc	P	R	F1	Acc
DNRTI	BERT-Base-Uncased	0.771	0.821	0.797	0.905	0.809	0.851	0.831	0.928
	BERT-Base-Cased	0.763	0.812	0.781	0.890	0.799	0.840	0.822	0.915
	BERT-Base-NER	0.774	0.824	0.801	0.901	0.781	0.838	0.811	0.909
	BERT-Large-NER	0.791	0.826	0.809	0.908	0.831	0.869	0.852	0.939
	BERT-Large-Cased	0.782	0.808	0.797	0.904	0.797	0.845	0.821	0.923
	SecureBERT-DNRTI	0.812	0.849	0.831	0.932	0.841	0.872	0.858	0.952
	GPT-3.5 1-shot	0.421	0.452	0.436	0.460	0.430	0.460	0.445	0.470
	GPT-3.5 3-shot	0.446	0.460	0.455	0.472	0.443	0.473	0.452	0.483
	GPT-3.5 5-shot	0.445	0.477	0.459	0.482	0.451	0.483	0.467	0.493
	GPT-3.5 7-shot	0.483	0.489	0.472	0.493	0.491	0.495	0.494	0.501
	GPT-4.0 1-shot	0.620	0.653	0.636	0.663	0.631	0.665	0.645	0.672
	GPT-4.0 3-shot	0.613	0.661	0.635	0.682	0.642	0.663	0.677	0.691
	GPT-4.0 5-shot	0.672	0.690	0.679	0.701	0.690	0.701	0.686	0.713
	GPT-4.0 7-shot	0.692	0.710	0.699	0.730	0.701	0.710	0.705	0.732
	Gemini 1-shot	0.458	0.482	0.469	0.483	0.459	0.488	0.467	0.490
	Gemini 3-shot	0.482	0.503	0.493	0.508	0.477	0.510	0.496	0.520
	Gemini 5-shot	0.501	0.524	0.511	0.530	0.502	0.531	0.516	0.540
	Gemini 7-shot	0.527	0.542	0.534	0.553	0.519	0.550	0.532	0.560
APTNER	BERT-Base-Uncased	0.698	0.726	0.713	0.868	0.704	0.716	0.709	0.862
	BERT-Base-Cased	0.688	0.724	0.703	0.867	0.699	0.733	0.714	0.864
	BERT-Base-NER	0.711	0.753	0.727	0.863	0.714	0.746	0.723	0.859
	BERT-Large-NER	0.751	0.786	0.777	0.879	0.757	0.791	0.779	0.880
	BERT-Large-Cased	0.743	0.781	0.769	0.868	0.748	0.787	0.765	0.866
	SecureBERT-APTNER	0.765	0.801	0.782	0.884	0.772	0.808	0.790	0.882
	GPT-3.5 1-shot	0.352	0.375	0.365	0.332	0.345	0.365	0.357	0.346
	GPT-3.5 3-shot	0.364	0.379	0.365	0.380	0.366	0.374	0.362	0.371
	GPT-3.5 5-shot	0.362	0.373	0.365	0.379	0.378	0.362	0.365	0.371
	GPT-3.5 7-shot	0.389	0.381	0.385	0.401	0.392	0.390	0.391	0.399
	GPT-4.0 1-shot	0.456	0.482	0.469	0.490	0.463	0.488	0.475	0.492
	GPT-4.0 3-shot	0.474	0.499	0.487	0.506	0.481	0.502	0.493	0.510
	GPT-4.0 5-shot	0.491	0.518	0.504	0.523	0.498	0.521	0.513	0.530
	GPT-4.0 7-shot	0.512	0.539	0.525	0.544	0.520	0.545	0.531	0.553
	Gemini 1-shot	0.356	0.380	0.367	0.393	0.358	0.382	0.370	0.395
	Gemini 3-shot	0.372	0.401	0.377	0.408	0.370	0.405	0.382	0.423
	Gemini 5-shot	0.391	0.422	0.407	0.431	0.392	0.425	0.409	0.445
	Gemini 7-shot	0.415	0.447	0.432	0.450	0.408	0.440	0.422	0.460

Table 6.1: Comparison of model performance metrics on the validation and test sets for DNRTI and APTNER datasets, including few-shot learning results from GPT-3.5, GPT-4.0, and Google's Gemini.

For the DNRTI dataset, the BERT-Large-NER model exhibited the highest performance among the baseline models, with a precision of 0.791, recall of 0.826, and F1 score of 0.809 on the validation set, and a precision of 0.831, recall of 0.869, F1 score of 0.852, and accuracy of 93.9% on the test set. These results underscore the model's robust generalization capabilities from training to unseen data.

The introduction of the SecureBERT-DNRTI model, fine-tuned specifically for the DNRTI dataset, further elevated performance metrics. SecureBERT-DNRTI achieved a precision of 0.812, recall of 0.849, F1 score of 0.831, and accuracy of 93.2% on the validation set. On the test set, it attained a precision of 0.841, recall of 0.872, F1 score of 0.857, and accuracy of 95.2%, outperforming the SOTA metrics, highlighting its effectiveness in accurately identifying and classifying cybersecurity-related entities.

Similarly, for the APTNER dataset, the BERT-Large-NER model demonstrated superior performance, achieving a precision of 0.751, recall of 0.786, F1 score of 0.777, and accuracy of 87.9% on the validation set. On the test set, it achieved a precision of 0.757, recall of 0.791, F1 score of 0.779, and accuracy of 88.0

The SecureBERT-APTNER model, tailored for the APTNER dataset, significantly enhanced NER performance. On the validation set, SecureBERT-APTNER achieved a precision of 0.765, recall of 0.801, F1 score of 0.782, and accuracy of 88.4%. On the test set, it recorded a precision of 0.772, recall of 0.808, F1 score of 0.790, and accuracy of 88.2%, demonstrating its proficiency in handling the specialized lexicon of cybersecurity threats. It is to be noted here that, we weren't able to reproduce the performance metrics using the BERT-base models achieved by [mention paper] on the APTNER dataset, they didn't mention anything about their hyperparameter settings in their work.

In summary, BERT-large-ner and the SecureBERT models performed comprehensively better than the other LLMs on both datasets, resulting in SOTA performance metrics, underscoring the transformative potential of transformer-based architectures in enhancing the accuracy and efficiency of cybersecurity threat detection and analysis. These results highlight the promising directions for future research and application in the field of Cyber Threat Intelligence (CTI).

6.4 Few-Shot Learning with GPT and Gemini Models

To evaluate GPT-3.5, GPT-4.0, and Google's Gemini models on the DNRTI and APTNER datasets, we used a few-shot learning approach. We prompted the models with 1-shot, 3-shot, 5-shot, and 7-shot examples, including a brief task description, an input text, and annotated output. This setup aimed to assess the models' ability to perform Named Entity Recognition (NER) with limited examples.

A query sample showing how we prompted(1-shot) GPT 3.5, GPT 4.0 and Gemini for doing the annotation of the texts is given below,

Query Sample:

I want to annotate a cyber threat report by using the following set of entities: hacker-organization(HackOrg), attack(OffAct), sample-file(SamFile), security-team(SecTeam), tool(Tool), time(Time), purpose(Purp), area(Area), industry(Idus), way(W-ay), loophole(Exp), features(Features).
Use B to signify the beginning of an entity, use I for intermediate position and O for words not belonging to any entity specified above.

See the following sentence from a cyber threat report with annotations,

Indeed O
, O
Kaspersky B-SecTeam
started O
tracking O
the O
BlueNoroff B-HackOrg
actor O
a O
long O
time O
ago O
. O
Now annotate the following sentence, please make sure the annotations have two columns: name and entity type:
The
majority
of
NewsBeef
targets
that
Kaspersky
researchers
have
observed
are
located

in
SA
.
This is the setup for 1-shot learning for the DNRTI dataset. For 3-shot learning, we provided 3 reference examples. Similarly, for 5-shot learning, we provided 5 reference examples and for 7-shot learning, we provided 7 reference examples. In this specific example, the output would be as follows,

The O
majority O
of O
NewsBeef B-SamFile
targets O
that O
Kaspersky B-SecTeam
researchers O
have O
observed O
are O
located O
in O
SA B-Area
. O

6.4.1 Results on DNRTI Dataset

The DNRTI dataset, with 27 classes, presented a simpler task. The performance metrics for the GPT and Gemini models are in Table 6.1.

GPT-3.5 showed gradual improvement with more shots, from a precision of 0.421, recall of 0.452, F1 score of 0.436, and accuracy of 0.460 in the 1-shot setting, to better metrics in the 7-shot setting. However, its performance remained moderate.

GPT-4.0 outperformed GPT-3.5, showing better generalization with more shots. The 1-shot configuration achieved a precision of 0.446, recall of 0.472, F1 score of 0.459, and accuracy of 0.478, which improved significantly in the 7-shot configuration.

The Gemini model also performed well, better than GPT-3.5 but slightly below GPT-4.0. The 1-shot configuration achieved a precision of 0.458, recall of 0.482, F1 score of 0.469, and accuracy of 0.483, improving further in the 7-shot configuration.

6.4.2 Results on APTNER Dataset

The APTNER dataset, with 85 classes, was more challenging. The results for the GPT and Gemini models are in Table 6.1.

GPT-3.5 struggled with the complexity, showing modest improvements from 1-shot to 7-shot configurations. The 1-shot precision was 0.352, recall 0.375, F1 score 0.365, and accuracy 0.332, with slight improvements in the 7-shot setting.

GPT-4.0 again showed superior performance, with the 1-shot configuration achieving a precision of 0.456, recall of 0.482, F1 score of 0.469, and accuracy of 0.490, improving significantly with more shots.

The Gemini model also performed well on APTNER, better than GPT-3.5 but slightly behind GPT-4.0. The 1-shot configuration had a precision of 0.356, recall of 0.380, F1 score of 0.367, and accuracy of 0.393, with better metrics in the 7-shot setting.

In summary, few-shot learning enabled GPT and Gemini models to perform NER on DNRTI and APTNER datasets. GPT-4.0 consistently outperformed GPT-3.5 and Gemini, showcasing its superior generalization in handling complex NER tasks.

However, having said that, all of these models performed significantly worse than the transformer-based large language models like BERT and SecureBERT. The potential reasons for this include the LLMs' pre-training on vast amounts of data, their ability to capture complex language patterns, and their architecture designed specifically for NER tasks. The transformer-based models benefit from extensive pre-training on diverse datasets, which equips them with a deeper understanding of context in language, leading to better performance in specialized tasks like NER in CTI

Chapter 7

Conclusion and Future Works

7.1 Conclusion

This thesis has methodically addressed the challenge of enhancing Named Entity Recognition (NER) within the domain of Cyber Threat Intelligence (CTI) by leveraging the power of advanced transformer-based models. Throughout our extensive experimentation on two specialized datasets, DNRTI and APTNER, we have demonstrated not only significant improvements in accuracy and efficiency of entity recognition but also established the robustness of our methodologies.

The core of our research centered on the deployment of transformer-based models, specifically tailored for the intricacies of cybersecurity contexts. These models, including variants such as SecureBERT-DNRTI and SecureBERT-APTNER, were meticulously fine-tuned to grasp and categorize the complex and nuanced entities typical in cybersecurity texts. The empirical results highlighted these models' superiority, showing marked improvements in precision, recall, and F1 scores compared to traditional models. This was particularly evident in their real-world applicability for automated threat detection systems, where they outperformed existing benchmarks significantly.

Moreover, this study delved into the comparative analysis of few-shot learning capabilities of models like GPT-3.5, GPT-4.0, and Google's Gemini. These models were evaluated for their ability to perform NER tasks with minimal training data, a scenario often encountered in dynamic cybersecurity environments where new threats continuously emerge. While the few-shot learning models demonstrated promising results, particularly GPT-4.0 which showed notable proficiency in learning from a limited number of examples, they still fell well short of the performance benchmarks set by the fully trained SecureBERT models. This underscores the necessity of extensive training and fine-tuning when dealing with complex datasets and highly specialized domain languages such as CTI.

The advancements made through this research significantly contribute to the broader cybersecurity field, aiming to bolster the capabilities of automated systems to detect and respond more adeptly to cyber threats. By integrating state-of-the-art NLP techniques and exploring the potential of transformer architectures, this thesis has paved the way towards developing more proactive and intelligent cybersecurity measures. These efforts are critical as they enhance the ability of cybersecurity systems to adapt rapidly and effectively in a landscape where threats evolve with increasing sophistication.

7.2 Future Works

The research presented in this thesis demonstrates the significant potential of applying advanced NLP techniques to enhance Named Entity Recognition (NER) within the Cyber Threat Intelligence (CTI) domain. However, the field is rapidly evolving, and there are numerous opportunities for further exploration and development. Future studies could consider the following expanded areas:

- **Expansion to Multilingual and Diverse Datasets:** To address the global nature of cyber threats, future work should explore the effectiveness of the proposed NER models across multilingual datasets. Extending these models to handle multiple languages and dialects could significantly increase their utility in international security contexts. Additionally, experimenting with a broader range of datasets, including those from emerging technology sectors, would help in understanding the models' versatility and robustness across different data contexts.

- **Integration with Real-time Cybersecurity Systems:** Implementing and testing these models within real-time threat detection systems could provide insights into their operational efficacy. It would also allow for the assessment of model performance under dynamic and high-stakes conditions, leading to further model optimization and refinement to meet industry standards.

- **Exploration of Few-Shot and Zero-Shot Learning Techniques:** Investigating the application of few-shot and zero-shot learning techniques could significantly reduce the dependency on large annotated datasets. Such studies would be particularly valuable in situations where rapid adaptability to new and emerging threats is crucial. This exploration could lead to developing models that quickly adapt to new threat landscapes with minimal retraining.

- **Automated Continuous Learning Systems:** Developing continuous learning mechanisms where models can autonomously update and learn from new data streams could

ensure sustained effectiveness in the face of the rapidly evolving nature of cyber threats. Such systems would help in maintaining the currency of cybersecurity measures without human intervention for retraining.

- **Enhancing Explainability and Transparency:** There is a growing need for explainable AI in cybersecurity. Enhancing the transparency of how these NER models make decisions could increase trust and reliability in automated systems among cybersecurity professionals. Studies could focus on developing techniques that provide clearer insights into the decision-making processes of AI, ensuring that they are both interpretable and justifiable.

- **Extension to Additional Datasets and Contexts:** Further research should also consider applying these NER models to additional datasets beyond DNRTI and APTNER. Exploring datasets from different sources or contexts, such as social media or less structured data forms, could help in refining the models' capabilities to handle diverse information formats and further enhance their applicability across various cybersecurity frameworks.

The future directions outlined here not only aim at technological advancements but also emphasize the need for a closer integration of AI tools with operational cybersecurity frameworks. Such an integrated approach is anticipated to significantly uplift the capabilities of security teams worldwide, ensuring more proactive and context-aware responses to cyber threats.

Bibliography

[1] 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.

[2] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] M. Schuster, K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[4] https://github.com/SCreaMxp/DNRTI-A-Large-scale-Dataset-for-Named-Enti

[5] Liu, P., Smith, J., and Smith, K. (2020). *Semantic Augmentation for Named Entity Recognition in Cyber Threat Intelligence*. Journal of Cybersecurity and Digital Forensics, 10(2), 123-132.

[6] S. K. Lim, A. O. Muis, W. Lu, H. Ong, "MalwareTextDB: A Database for Annotated Malware Articles," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1557-1567, Vancouver, Canada, July 2017, doi: 10.18653/v1/P17-1143.

[7] Sheng-Shan Chen, Ren-Hung Hwang, Chin-Yu Sun, Ying-Dar Lin, and Tun-Wen Pai, "Enhancing Cyber Threat Intelligence with Named Entity Recognition Using BERT-CRF," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, Kuala Lumpur, Malaysia, 2023, IEEE. DOI: 10.1109/GLOBECOM54140.2023.10436853. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10436853

[8] Tom B. Brown et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *5th Workshop on Energy Efficient*

Machine Learning and Cognitive Computing - NeurIPS 2019, 2019. [Online]. Available: https://arxiv.org/abs/1910.01108

[10] John D. Lafferty, Andrew McCallum, and Fernando C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, Williamstown, MA, USA, 2001, pp. 282-289. [Online]. Available: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers

[11] P. Evangelatos, C. Iliou, T. Mavropoulos, K. Apostolou, T. Tsikrika, S. Vrochidis, I. Kompatsiaris, "Named Entity Recognition in Cyber Threat Intelligence Using Transformer-based Models," in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2021, pp. 348–353, doi: 10.1109/CSR51186.2021.9527981.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*, pp. 4171–4186.

[13] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *NeurIPS 2019*.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv preprint arXiv:1907.11692*, 2019.

[15] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," in *ICLR 2020*.

[16] X. Wang, S. He, L. Xiong, X. Wu, J. Li, and J. Chen, "APTNER: A Specific Dataset for NER Missions in Cyber Threat Intelligence Field," in *Proceedings of the IEEE International Conference on Cyber Security and Resilience*, 2022.

[17] Erik F. Tjong Kim Sang and Fien De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, Edmonton, Canada, 2003, pp. 142–147.

[18] M. Barnum, "STIX: The Structured Threat Information eXpression," *MITRE Corporation*, 2012.

[19] I. Deliu, C. Leichter, K. Franke, "Collecting Cyber Threat Intelligence from Hacker Forums via a Two-Stage, Hybrid Process using Support Vector Machines and Latent Dirichlet Allocation," *Journal of Cybersecurity*, vol. 2023, no. 1, pp. 45-60, 2023.

[20] C. Cortes, V. Vapnik, "Support Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[21] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[22] N. Dionísio, F. Alves, P. M. Ferreira, A. Bessani, "Cyberthreat Detection from Twitter using Deep Neural Networks," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2019.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, November 1998.

[24] M. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, F. Menczer, "Predicting the Political Alignment of Twitter Users," in *Proceedings of IEEE SocialCom*, pp. 192-199, 2015.

[25] T. Satyapanich, F. Ferraro, T. Finin, "CASIE: Extracting Cybersecurity Event Information from Text," *Journal of Cybersecurity and Information Systems*, vol. 23, no. 4, pp. 22-29, 2019.

[26] Y. Fang, Y. Zhang, C. Huang, "CyberEyes: Cybersecurity Entity Recognition Model Based on Graph Convolutional Network," *The Computer Journal*, 2020, doi: 10.1093/comjnl/bxaa141.

[27] T. N. Kipf, M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *ICLR 2017*.

[28] X. Ma, E. Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1064–1074, 2016.

[29] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J. Tsujii, "BRAT: a Web-based Tool for NLP-Assisted Text Annotation," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012. [Online]. Available: https://www.aclweb.org/anthology/E12-2021

[30] OASIS Cyber Threat Intelligence Technical Committee, "Structured Threat Information eXpression (STIX 2.1)," OASIS Standard, 2020. [Online]. Available: https://docs.oasis-open.org/cti/stix/v2.1/stix-v2.1.html

[31] E. F. Tjong Kim Sang, "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition," in *Proceedings of CoNLL-2002*, 2002. [Online]. Available: https://www.aclweb.org/anthology/W02-2024

[32] Ramshaw, Lance A. and Marcus, Mitchell P., "Text Chunking using Transformation-Based Learning," in *Natural Language Processing Using Very Large Corpora*, 1999. Available: https://www.aclweb.org/anthology/W99-0604

[33] Ratinov, Lev and Roth, Dan, "Design Challenges and Misconceptions in Named Entity Recognition," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, 2009. Available: https://www.aclweb.org/anthology/W09-1119

[34] Hugging Face's Transformers Library Documentation. Available online: https://huggingface.co/docs/transformers/index.html

[35] Google Colab Pro. Available from: https://colab.research.google.com/

[36] Paszke, Adam et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". 2019. Available from: https://pytorch.org/

[37] OpenAI, "GPT-4: Language Models are Multimodal Learners," 2023.

[38] Ehsan Aghaei, Xi Niu, Waseem Shadid, and Ehab Al-Shaer, "SecureBERT: A Domain-Specific Language Model for Cybersecurity," in *Security and Privacy in Communication Networks (SecureComm 2022), Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST), Vol. 462*, Springer, Guangzhou, China, 2023, pp. 39–56. DOI: 10.1007/978-3-031-28477-9_3.

Generated using Undegraduate Thesis LaTeX Template, Version 2.2. Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

This thesis was generated on Friday 12th July, 2024 at 2:23pm.