Computers & Security 136 (2024) 103579

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

CyberEntRel: Joint extraction of cyber entities and relations using deep learning

Kashan Ahmed [a], Syed Khaldoon Khurshid [a], Sadaf Hina [b,*]

[a] University of Engineering and Technology, Lahore 54890, Punjab, Pakistan
[b] University of Salford, Salford M5 4WT, Manchester, United Kingdom

ARTICLE INFO

Keywords:
Cyber threat intelligence
Deep learning
Named entity recognition
Relation extraction
Knowledge graph

ABSTRACT

The cyber threat intelligence (CTI) knowledge graph is beneficial for making robust defense strategies for security professionals. These are built from cyber threat intelligence data based on relation triples where each relation triple contains two entities associated with one relation. The main problem is that the CTI data is increasing more rapidly than expected and existing techniques are becoming ineffective for extracting the CTI information. This work mainly focuses on the extraction of cyber relation triples in an effective way using the joint extraction technique, which resolves the issues in the classical pipeline technique. Firstly, the 'BIEOS' tagging scheme was applied to CTI data using the joint tagging technique and then the relation triples were jointly extracted. This study utilized the attention-based RoBERTa-BiGRU-CRF model for sequential tagging. Finally, the relation triples were extracted using the relation-matching technique after matching the best suitable relation for the two predicted entities. The experimental results showed that this technique outperformed the state-of-the-art models in knowledge triple extraction on CTI data. Furthermore, a 7% increase in the F1 score also proved the effectiveness of this technique for the information extraction task on CTI data.

1. Introduction

Cyber security-related information is growing rapidly on the internet as cyber-attacks are increasing day by day. Attackers mostly target military, government, and corporate sectors as these contain sensitive and classified information that requires proper defensive strategies Watters (2023). For this purpose, mainly two types of defensive techniques are used; i) detection of cyber attacks and harmful events and mitigating them in real-time when they occur, and ii) earlier prediction of potential cyber-attacks using cyber threat intelligence and making a strong defensive mechanism.

The increasing number of cyber-attacks and data breaches have made cyber security a top priority for governments, corporations, and individuals. Big data analytics has become a valuable tool in cyber security, providing a way to examine large amounts of data to specify potential threats and vulnerabilities Tang et al. (2017). In a recent research study related to cyber attacks, researchers deeply studied 60 government websites and found 18% of them highly and 33% medium vulnerable using Arachni tool, and 13% highly and 41% medium vulnerable using ZAP tool Ghazanfar et al. (2021).

Security professionals develop and deploy security solutions for the detection and mitigation of cyber-attacks. However, these do not always prove to be the best as attackers keep trying new tactics, techniques, and procedures. Notably, it's a very difficult and time-consuming task for the attackers to change their entire attacking process Zongxun et al. (2021). Studies show that most attacker groups can be identified by their attacking techniques as they incorporate advanced persistent threats (APTs) and ransomware attacks as attacking methods. For the earlier prediction of possible cyber attacks, cyber threats and attacking groups' related information can be extracted from cyber threat intelligence (CTI) data Sun et al. (2023).

Cyber threat intelligence (CTI) is a critical element of modern-day security operations Conti et al. (2018). CTI data comprises a vast amount of information about potential cyber threats, such as attacker profiles, tactics, techniques, and procedures. Effective extraction of valuable information from CTI data can help security professionals make informed decisions and develop robust defense strategies Ainslie et al. (2023). However, the exponential growth of CTI data has made it challenging to extract relevant information in a timely and accurate manner. This problem has spurred the development of various tech-

* Corresponding author.
E-mail addresses: kashanahmed867@gmail.com (K. Ahmed), khaldoon@uet.edu.pk (S.K. Khurshid), s.hina@salford.ac.uk (S. Hina).

https://doi.org/10.1016/j.cose.2023.103579
Received 24 August 2023; Received in revised form 20 October 2023; Accepted 30 October 2023
Available online 7 November 2023

niques for the extraction of entities along with relations from CTI data using deep learning.

In the past few years, natural language processing (NLP) and deep learning techniques have shown tremendous potential in solving problems related to the extraction of entity relations from unstructured text data Jo et al. (2022). These techniques can learn patterns and relationships from data without relying on pre-defined rules, making them suitable for complex and dynamic datasets such as CTI data. However, existing techniques for entity and relation extraction from CTI data suffer from several limitations, including low accuracy, scalability, and generalizability.

The above-mentioned challenges need be focused attention and there are obvious research questions that lead to a motivation to work on this research. The research questions addressed in this study include: How can cyber threat intelligence (CTI) data be effectively extracted to support the development of robust defense strategies? How can deep learning methods be utilized to improve the accuracy, scalability, and generalizability of CTI extraction? How does the proposed technique perform compared to the current state-of-the-art models? Can future research further enhance the proposed technique by incorporating multimodal data, graph-based models, and knowledge graphs? By addressing these research questions, this research enhances the existing literature on cyber threat intelligence and entity-relation extraction, providing valuable insights for researchers, practitioners, policymakers, and security professionals in the field.

The primary objective of this research study was to propose a novel technique for extracting cyber entities and relations from textual data using deep learning methods. Specifically, this study focused on the joint extraction technique, which resolves the limitations of the classical pipeline technique used in the existing techniques. Researchers in this study utilized an attention-based RoBERTa-BiGRU-CRF model for sequential tagging and relation-matching techniques to extract relation triples from CTI data. To evaluate the effectiveness of this technique, experiments were conducted using a manually annotated CTI dataset, and the interpretation of the model was compared to the current state-of-the-art models.

The outcomes of the experiments revealed that the proposed technique outperformed existing models in terms of accuracy, scalability, and generalizability. The proposed technique achieved the final F1 score of 0.86, which is 7% higher than the best-performing baseline model. Furthermore, the joint extraction technique showed promising results in addressing the challenges associated with the entity and relation extraction from CTI data, providing security professionals with more accurate and timely insights to make informed decisions. Overall, the proposed technique for extracting cyber entities and relations from textual data using deep learning has the prospect of making a considerable impact in the domain of cyber threat intelligence. The results of this research can be used by researchers, practitioners, and policymakers to generate more practical tools and strategies for CTI data analysis and decision-making Choo et al. (2019a,b).

The introduction section discussed the importance of CTI data and the limitations of existing techniques for the extraction of entities and relations. In the related study/literature review section the state-of-the-art models for CTI data extraction are reviewed and the gaps in the literature are highlighted. The methodology/technique section presents the joint extraction technique using an attention-based RoBERTa-BiGRU-CRF model for sequential tagging and relation-matching techniques for relation triple extraction. The implementation/experimentation section describes the dataset, experimental setup, and evaluation metrics used to validate the proposed technique. This paper also presented the experimental results, including comparisons with state-of-the-art models, and discussed the significance of the results. Finally, in the conclusion section, the findings are summarized, the contributions of this research are highlighted, and future directions are discussed to improve the proposed technique.

2. Related study

Recent research has shown great focus on proposing various techniques and methods for extracting actionable information from CTI data, specifically in the area of relation triples extraction. Table 1 shows relevant studies for entities and relation extraction.

One of the main areas of research in this field is the development of systems and platforms for extracting and analyzing CTI data, such as the CTI View where Yinghai et al., proposed a methodology designed to process and analyze massive amounts of unstructured cyberspace threat intelligence (CTI) data for enhanced defense against advanced persistent threats (APTs). To address the challenges posed by heterogeneous CTI, their methodology incorporated a text extraction framework that utilized automated test frameworks, text recognition technology, and text denoising techniques. The framework also tackled the adaptability issues encountered when crawling diverse CTI sources. Furthermore, their study leveraged regular expressions along with blacklist and whitelist mechanisms to extract relevant indicators of compromise (IoC) and tactics, techniques, and procedures (TTPs) information from the CTI data Zhou et al. (2022). In another research, TriCTI was presented as an actionable cyber threat intelligence discovery system that functioned via a trigger-enhanced neural network. In this research, Jian et al., utilized natural language processing (NLP) technology to establish relationships between indicators of compromise (IoC) and campaign stages in cybersecurity reports. They generated actionable CTI by utilizing the concept of “campaign triggers” to enhance the explanation of campaign stages and slightly improve the performance of the classification model. Campaign trigger phrases are identified as keywords within sentences that indicate specific campaign stages. The system employed NLP techniques to train final trigger vectors that possessed similar semantic representations to the keywords found in unseen sentences Liu et al. (2022).

There is a growing interest in the use of neural network-based models for extracting information from CTI data. Zongxun et al. proposed a methodology that could capture the semantics of sentences within the context of cyber threat intelligence. The presented model was capable of identifying threat actions within sentences, providing means to mark and highlight specific actions related to threats. By leveraging existing knowledge, they constructed cyber threat ontology, to obtain comprehensive attack information. Their method enabled the generation of high-level indicators of compromise (IoC) along with tactics, techniques, and procedures (TTPs) by mapping the threat actions to the ontology and constructing TTPs based on this mapping Zongxun et al. (2021). Xuren et al. generated a new dataset and applied a neural network for the named entity recognition (NER) task. The rationality of their dataset was demonstrated through rigorous evaluation and analysis, highlighting its relevance and suitability for NER joint and multi-task learning tasks in CTI. Researchers introduced a dataset for document-level threat intelligence analysis, encompassing entities, relations, and coreference annotations. They proposed a joint learning framework that combined relation extraction and coreference resolution, treating coreference as a predefined relation using GCN Wang et al. (2020b). The above-mentioned studies utilized neural networks for extracting information from CTI data. A recent research named CDTier introduced a Chinese dataset of threat intelligence entity relationships and is a significant read, as it incorporated neural network-based models. The findings demonstrated that the developed system significantly minimized threat intelligence analysts’ work while assessing threat intelligence Zhou et al. (2023).

Another notable area of research is the construction of knowledge graphs for CTI data as proposed by Yitong et al. This research focused on constructing and applying an advanced persistent threat (APT) knowledge graph from open-source cyber threat intelligence (OSCTI) data. Researchers developed a cybersecurity platform called CSKG4APT based on the knowledge graph and employed deep learning and expert knowledge to extract and update APT threat knowledge. The proposed

Table 1
Studies concerning entities and relations extraction.

Ref.	Technique	Dataset	Entities	Relations	Joint extraction	Knowledge graph	Explainable AI
Zhou et al. (2022)	BERT-BiLSTM-GRU-CRF	Custom	✓	✗	✗	✗	✗
Liu et al. (2022)	TriCTI	Custom	✓	✓	✗	✗	✗
Wang et al. (2022a)	BERT-BiLSTM-CRF	Custom	✓	✗	✗	✗	✗
Wang et al. (2020b)	C-GCN	Public	✓	✓	✓	✓	✗
Ren et al. (2022)	BERT-BiLSTM-CRF	Custom	✓	✓	✗	✓	✗
Sarhan and Spruit (2021)	Open-CyKG	Microsoft, MalwareDB	✓	✓	✗	✓	✗
Zuo et al. (2022)	BERT-BiLSTM-CRF	Custom	✓	✓	✓	✗	✗
Zhang et al. (2020)	PCNN	NYT-10	✓	✓	✓	✓	✗
Guo et al. (2023)	BERT with BiGRU	OSINT Data	✓	✓	✓	✓	✗
Wang and Liu (2023)	Pert	Microsoft, OpenCyber	✓	✗	✗	✗	✗
Srivastava et al. (2023)	BERT with BiLSTM	Microsoft, Metasploit	✓	✗	✗	✗	✗
Proposed Technique	RoBERTa-BiGRU-CRF	Custom	✓	✓	✓	✓	✓

APT attack attribution method enhanced defense strategies by integrating fragmented intelligence and actively adjusting defense strategies, paving the way for dominance in network attack and defense Ren et al. (2022). In a similar research, proposed by Sharan et al., an open cyber threat intelligence knowledge graph framework was introduced for automated information extraction from unstructured advanced persistent threat (APT) reports. They employed an attention-based neural open information extraction (OIE) model and neural cybersecurity named entity recognizer (NER) to extract entities and generate relation triples. The structured data is then used to construct the knowledge graph. The results indicated that the proposed method enabled security professionals to retrieve valuable information through query execution Sarhan and Spruit (2021). These research studies proposed techniques for representing CTI data in a structured and organized manner to make it more useful for corporations.

Joint extraction techniques, which involve jointly extracting entities and relations from text, have also been proposed as a solution to the limitations of existing techniques that focus on extracting entities and relations separately. Junjia et al., proposed an end-to-end joint extraction model for CTI to improve the accuracy and efficiency of entity and relation extraction. The model combines a joint tagging strategy with an end-to-end sequence tagging model Zuo et al. (2022). Zhenyu et al., also proposed a joint learning framework for relation extraction and entity linking, leveraging the semantic relevance between entities and relations. The result depicted good improvement over other traditional methods Zhang et al. (2020).

Named entity recognition (NER) is also an important task in CTI data extraction where entities can be extracted via rule-based, machine-learning, deep-learning methods Zhao et al. (2020a) Poostchi and Piccardi (2019). Xuren et al., introduced a dataset that used NER for the extraction of entities from the CTI data Wang et al. (2020a). They further introduced another dataset that also utilized NER for the extraction of entities from the CTI data but by using deep learning and domain knowledge engineering techniques Wang et al. (2022b). Hank et al., proposed a CTI corpus that was created from open sources to train and test cybersecurity entity models using the spaCy framework which is commonly used for NER tasks. Self-learning methods were explored to automatically recognize cybersecurity entities, and cybersecurity domain entity linking was applied with existing world knowledge from Wikidata Hanks et al. (2022). In a similar study, Feng et al., proposed compelling techniques for NER in CTI data. Their model combined regular expressions, a known-entity dictionary, and conditional random fields (CRF) with four feature templates to capture the characteristics and correlations of security-named entities. The rule-based expressions, known-entity dictionary, and CRF-based extractor were integrated to enhance recognition performance Yi et al. (2020).

The objective of entity and relation extraction is to extract relevant triples from textual data. This process can be classified into two distinct techniques, the pipeline technique and the joint learning technique, based on their individual processing methods.

2.1. Pipeline extraction technique

The pipelining technique for CTI extraction is a widely used technique in the field of natural language processing. It involves breaking down the task of CTI extraction into several distinct steps or modules, each designed to address a specific subtask. One common pipeline technique utilizes an NLP-based technique to extract entity and relation information from the text Chen and Guo (2022).

In recent years, several studies have explored the use of various NLP techniques for CTI extraction using the pipeline technique. For example, Liu et al. (2023) proposed a methodology that introduced a pipeline relation extraction method for improved performance in relation extraction tasks. Unlike the traditional joint models that share span representation, their method utilized separate encoders for entity recognition and relation classification, reducing feature conflict. By incorporating advanced pre-trained models and attention mechanisms, their method fused contextual semantic representation, addressing a limitation of other pipeline models but still results were not effective. Additionally, the framework incorporated explicit entity mentions, allowing for the capture of entities' locale and type information, which is challenging to utilize in joint models. Another study by Bayer et al., proposed a pipeline technique for CTI extraction which used both supervised and unsupervised machine learning techniques. The proposed system consisted of several modules, including pre-processing, entity recognition, relation extraction, and post-processing. Researchers evaluated their system on the SemEval 2018 Task 10 dataset and achieved competitive results. They also analyzed the contributions of each module and found that the entity recognition and relation extraction modules were the most important for overall performance Bayer et al. (2022). In addition, Li et al. (2022b) proposed a pipeline technique for CTI extraction that incorporated a pre-processing step, an entity extraction step, and a relation extraction step. The system used a combination of rule-based and machine-learning techniques and achieved promising results in CTI extraction.

2.2. Joint extraction technique

Joint extraction techniques have become increasingly popular in CTI due to their ability to extract multiple pieces of information from text simultaneously. These techniques combine named entity recognition and relation extraction into a single model, resulting in more accurate and efficient CTI extraction. One such method is the multi-task learning technique, which jointly learns to predict both named entities and relations in a single model. Wang et al. Dimitriadis et al. (2021) proposed a multi-task learning technique that outperformed state-of-the-art CTI extraction models on several benchmark datasets.

Another joint extraction technique is the end-to-end neural network model, which directly extracts entities and relations from text without relying on any pre-processing steps. Li et al. (2022a) proposed a novel end-to-end neural network model for CTI extraction that leveraged

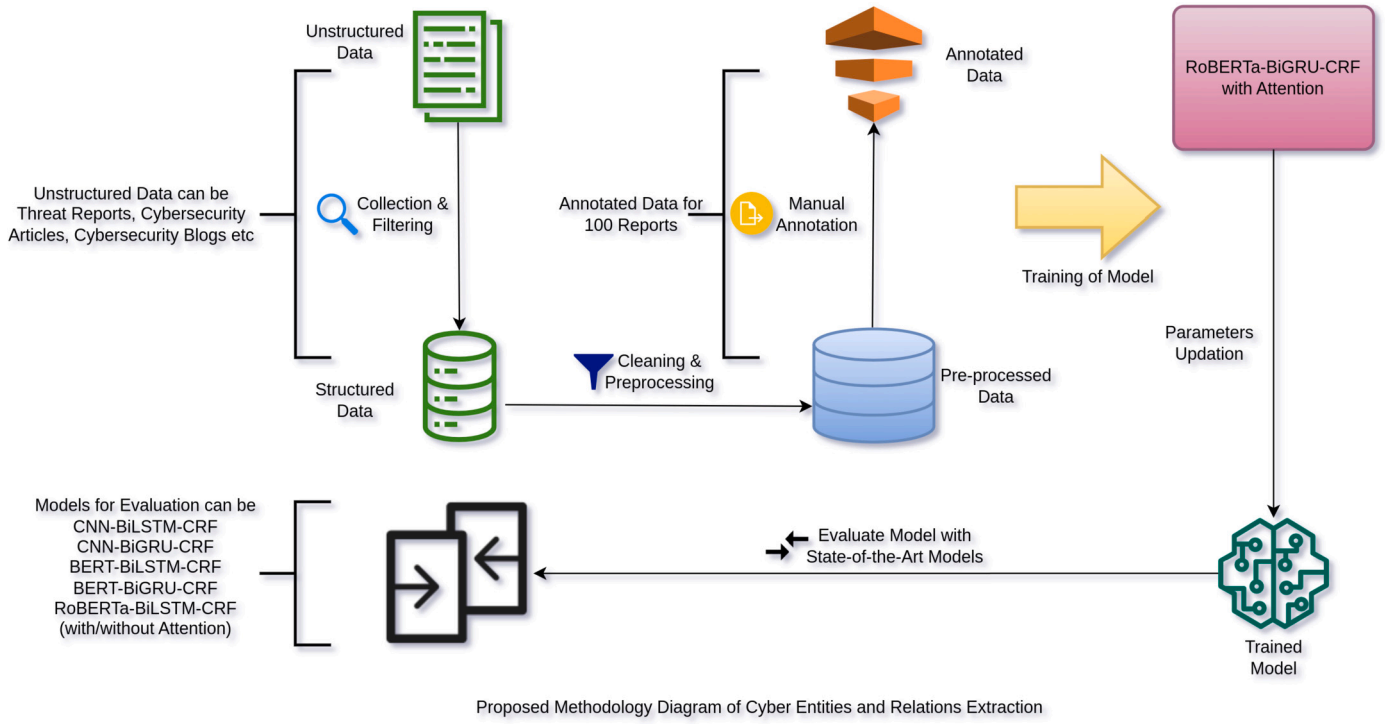

Fig. 1. Proposed methodology diagram of cyber entities and relations extraction.

contextual embeddings and self-attention mechanisms. Their model achieved state-of-the-art performance on several benchmark datasets, demonstrating the effectiveness of end-to-end techniques for CTI extraction.

In addition to multi-task learning and end-to-end neural network models, there have been other joint extraction techniques proposed for CTI extraction. Ge and Wang (2022) proposed a multi-level attention model that incorporated both local and global contextual information for CTI extraction. Similarly, Zhao et al. (2020b) proposed a graph convolutional network (GCN) for joint extraction of entities and relations. Their model utilized the graph structure of CTI data to capture global dependencies between entities and relations, resulting in improved extraction performance. Marchiori et al., used STIX format while using joint entities and relations extraction from CTI reports Marchiori (2021/2022).

Overall, joint extraction techniques have shown great potential in CTI extraction, particularly in their ability to extract multiple pieces of information from text simultaneously. Two of the famous techniques in this area are the CyberRel technique Guo et al. (2021) and the RelExt technique Pingle et al. (2019). As the field of CTI continues to grow, it is likely that joint extraction techniques will become increasingly important for accurately and efficiently extracting CTI from large volumes of text data.

This section highlighted various techniques proposed for extracting actionable information from CTI data, specifically in the area of relation triples extraction. There is a growing interest in the use of neural network-based models, the construction of knowledge graphs, joint extraction techniques, and named entity recognition for CTI data extraction. The incorporated studies proposed techniques for representing CTI data in a structured and organized manner to make it more useful for corporations. This research aims to build upon the existing research in this field by proposing a new technique for extracting relation triples from CTI data and evaluating its effectiveness through experimental results.

3. Methodology

This section will discuss the joint tagging scheme, end-to-end model, and joint entities & relations extraction. Fig. 1 shows the proposed technique used in this research.

In this study, researchers first collected unstructured CTI data from relevant sources. This unstructured data was then cleaned and converted into the format that was best for deep learning models. Next, the structured data was annotated with entity and relation labels using the joint tagging technique. After annotation, the joint extraction technique was applied to extract relation triples from the annotated data using a RoBERTa-BiGRU-CRF model. A relation-matching technique was used to match the best suitable relation for two predicted entities. The model was fine-tuned to improve its performance on the specific task, followed by evaluation using appropriate metrics i.e. precision (P), recall (R), and F1.

3.1. Joint tagging technique

The CTI articles used in this research were manually annotated using an online tool, UBIAI, to create a labeled dataset. The dataset consisted of 100 CTI articles that were preprocessed to make them suitable for deep learning models. The first step in the preprocessing was data cleaning, which involved removing any irrelevant or duplicate information. The articles were then tokenized i-e. segmented into individual words, phrases, and sentences. Next, the articles were converted into a numerical format that could be processed by the deep learning models. Finally, the BIEOS tagging technique was applied to the CTI data to identify the entities and their types.

The BIEOS tagging technique is a modification of the BIO tagging technique commonly used in named entity recognition (NER). BIEOS stands for beginning, inside, end, outside, and single, and it is used to annotate entities in text. The BIEOS technique is used to mark the start, center, and stop of an entity, as well as entities that consist of a single word. This technique is used in this research to label the entities in the CTI data. The BIEOS tagging technique is important because it allows

Table 2
Relation matching rule for related entities.

Entity 1	Entity 2	Relation
corporation	locale	originate_from
application	vulnerability	is_susceptible
corporation	assault_technique	uses
corporation	malware	contains_product
malware	infected_file	contains_file
infected_file	assault_technique	belongs_to
APT_group	locale	operates_in
APT_group	campaign	launches
APT_group	application	uses
APT_group	assault_technique	uses
ransomware	malware	is_a_type_of
ransomware	application	uses
ransomware	assault_technique	uses
indicator	malware	indicates
indicator	infected_file	indicates
algorithm	application	implements

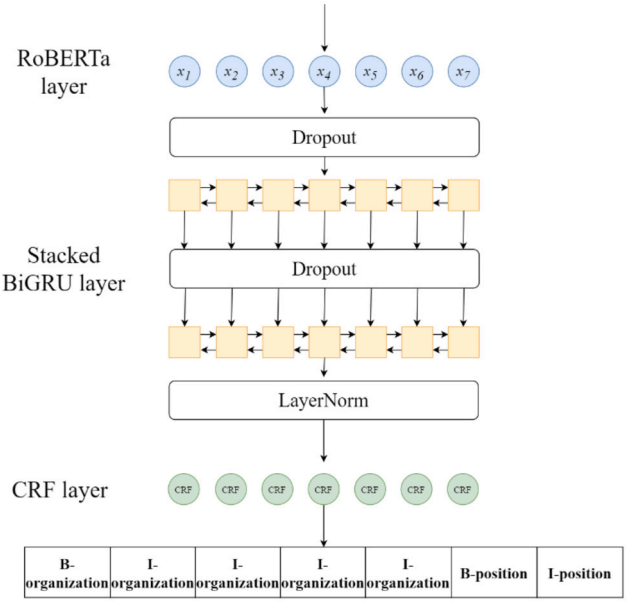

Fig. 2. The model's design for sequence tagging.

the deep learning models to recognize the boundaries of different entities in the textual data, which is essential for accurate entity recognition and the best possible relation extraction.

This study separates all entities into twelve categories based on the CTI corpus: corporation (COR), locale (LOC), application (APP), malware (MLW), vulnerability (VLN), assault_technique (TEC), infected_file (INF), APT_group (APT), ransomware (RAN), indicator (IND), campaign (CAM), and algorithm (ALG). Furthermore, the relationships are classified into eleven types: originate_from, utilizes, is_susceptible, contains_product, contains_file, belongs_to, operates_in, launches, is_a_type_of, indicates, and implements. The numerals "1" and "2" represent the entity roles where "1" signifies that the entity is the triple's start, while "2" implies, the entity is the triple's end. Table 2 shows the main criteria for entity relationships with various relations. Furthermore, "OVE" denotes that the entity to whom the term refers is involved in numerous relationships Zuo et al. (2022).

3.2. Model architecture

The RoBERTa layer, BiGRU layer, attention layer, and CRF layer construct the full model, as depicted in Fig. 2. The RoBERTa layer generates vector representations of the input words. The BiGRU layer processes the word embeddings and computes the probability distribution for every word. The attention layer catches word dependencies by computing attention scores between each pair of words. Finally, the CRF layer selects the best sequence tagging by considering the dependencies between the predicted tags.

The proposed model architecture for this research consists of the RoBERTa-BiGRU-CRF model, which is a deep learning-based technique for extracting cyber entities and relations from textual data. The architecture of the model consists of four main layers:

- RoBERTa layer: The RoBERTa layer is a pre-trained transformer-based language model that generates vector representations of the input words.
- BiGRU layer: The BiGRU layer processes the word embeddings generated by the RoBERTa layer and computes the probability distribution for every word.
- Attention layer: The attention layer catches word dependencies by computing attention scores between each pair of words. The attention scores are computed using a dot product between the BiGRU outputs of each word pair.
- CRF layer: The CRF layer selects the global best tagging sequence by considering the dependencies between the predicted tags. It applies a constraint to ensure that the output sequence forms a valid sequence of entities and relations.

3.2.1. RoBERTa layer

The RoBERTa layer is a transformer-based language instance that operates on bidirectional attention to contextualizing words in a given text sequence Liu et al. (2019). It encodes each token in the input sequence by aggregating information from all the other tokens, considering them in both the left and right directions. The outcome of this layer is a series of contextualized embeddings for every token, which is passed on to the next layer. The mathematical equation for the RoBERTa layer can be written as follows:

$$\mathbf{H} = RoBERTa(\mathbf{X})$$

where $\mathbf{X}$ shows the input series and $\mathbf{H}$ shows the output series of contextualized embeddings.

3.2.2. BiGRU layer

The bidirectional gated recurrent unit (BiGRU) layer is used to capture sequential dependencies within the input sequence Lu et al. (2021). It consists of two layers of gated recurrent unit (GRU) cells, one processing the input sequence from left to right, and the other from right to left. The outcomes of both GRU layers are combined to yield the final outcome. The mathematical equation for the BiGRU layer can be represented as follows:

$$\vec{\mathbf{h}}t = GRU_{\text{forward}}(\mathbf{h}_{t-1}, \mathbf{x}_t)$$

$$\overleftarrow{\mathbf{h}}t = GRU_{\text{backward}}(\overleftarrow{\mathbf{h}}_{t+1}, \mathbf{x}_t)$$

$$\mathbf{h}_t = [\vec{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t]$$

where $\mathbf{x}_t$ represents the t^{th} token in the input series, $\vec{\mathbf{h}}_t$ represents the latent state of the forward GRU cell, $\overleftarrow{\mathbf{h}}_t$ shows the latent state of the backward GRU cell and $\mathbf{h}_t$ shows the final outcome of the BiGRU for the t^{th} token.

3.2.3. Attention layer

The attention layer is used to compute the attention scores between the outcome of the BiGRU layer and a learnable query vector to capture the relative importance of each token in the input sequence for a given relation. The attention scores are utilized to calculate a weighted aggregate of the output of the BiGRU layer, which represents the context vector. The context vector is then passed on to the CRF layer for decoding. The mathematical equation for the attention layer can be represented as follows:

$$\mathbf{e}_t = \tanh(\mathbf{W}_1\mathbf{h}_t + \mathbf{W}_2\mathbf{q})$$

$$\alpha_t = \frac{\exp(\mathbf{v}^T\mathbf{e}_t)}{\sum_{j=1}^{n}\exp(\mathbf{v}^T\mathbf{e}_j)}$$

$$\mathbf{c} = \sum_{t=1}^{n}\alpha_t\mathbf{h}_t$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{v}$ are learnable weight matrices, $\mathbf{q}$ is a learnable query vector, $\mathbf{h}_t$ is the hidden state of the BiGRU layer at time-period stage t, $\mathbf{e}_t$ is the attention vector at time-period stage t, and α_t is the attention score due to the token at time-period stage t. The attention score α_t is calculated by utilizing the softmax method over the dot product of the attention vector $\mathbf{e}_t$ and the learnable weight vector $\mathbf{v}$. The context vector $\mathbf{c}$ is computed as the weighted aggregate of the output of the BiGRU layer $\mathbf{h}_t$ using the attention scores α_t. The context vector captures the most relevant information in the input sequence for a given relation and is used as input to the CRF layer for decoding the final tagging sequence Yang and Xiao (2022).

3.2.4. CRF layer

The CRF layer is the final layer in our model and brings the context vector from the attention layer as input and outputs the most probable tagging sequence for the input sentence. The CRF layer models the dependencies between neighboring tags and takes into account the global structure of the tagging sequence Alves-Pinto et al. (2022). The mathematical equation for the CRF layer must be represented as written below:

$$\mathbf{y}^* = \text{argmax}\mathbf{y}\left(\sum_{i=1}^{n}A_{y_{i-1},y_i} + \sum_{i=1}^{n}E_{y_i}(\mathbf{h}_i)\right)$$

where $\mathbf{y}$ is a tagging sequence of length n, A is a transition matrix that models the possibility of transitioning from label y_{i-1} to label y_i, $E_{y_i}(\mathbf{h}_i)$ is the emission score for label y_i given the context vector $\mathbf{h}_i$, and $\mathbf{y}^*$ is the most possible tagging sequence. The transition matrix and the emission scores are learned during training using maximum likelihood estimation. During inference, the Viterbi algorithm Noravesh (2023) is employed to efficiently compute the most possible sequence tagging given the input text and the learned parameters of the final and trained model.

The RoBERTa-BiGRU-CRF model is introduced end-to-end using a joint tagging technique and a BIEOS tagging technique. The joint tagging technique allows the model to predict the entity and relation tags simultaneously, while the BIEOS tagging technique provides a more granular representation of the entity boundaries. The model is prepared using a combination of negative log-likelihood loss and L2 regularization loss and is optimized using the Adam algorithm as an optimizer. Furthermore, after the setting of different hyper-parameters, the final alpha is set to 0.0002.

3.3. Joint extraction technique

The joint extraction technique is a technique used to extract multiple relations simultaneously from a single sentence. This technique is applied to the proposed model to extract multiple relations. In this technique, multiple relations are modeled jointly in the CRF layer, which enables the model to capture the dependencies among different relations and their corresponding entities.

The joint extraction technique involves modifying the traditional CRF layer by incorporating multiple tag sets. Specifically, for each relation type, a separate tag set is used. For example, if there are three relation types, then there will be three tag sets, one for each relation type. Each tag set consists of tags that correspond to the entities involved in that relation type.

In the stage of model training, the methods learn to assign the appropriate tag for each entity in the sentence for each relation type. During inference, the model outputs multiple tags for each entity, one for every type of relation. The final output of the method is a group of relation triples, and every relation triple consists of a relation type and the corresponding entities.

The joint extraction technique is proved to be effective in capturing the dependencies among different relations and their corresponding entities. It enables the model to extract multiple relations from a single sentence with high accuracy and efficiency. In the proposed model, the joint extraction technique was applied to extract multiple relations from CTI textual data, which typically contain complex and overlapping relations.

3.4. Relation matching technique

The relation-matching technique desires to enhance the overall performance of the model by incorporating external knowledge sources such as knowledge graphs. In this step, a relation-matching module was utilized to calculate the semantic similarity between the extracted relation and the relation in the generated knowledge graph. This is done by first retrieving all the candidate relations from the knowledge graph that are related to the entities in the extracted relation. Then, a relation-matching score is calculated between each candidate relation and the extracted relation.

To calculate the relation-matching score, a graph convolutional network (GCN) based technique was incorporated Zhong et al. (2023). The GCN takes the nodes and edges of the knowledge graph as input and applies a series of graph convolutions to learn a representation of each node that captures its neighborhood information. The output of the GCN is a matrix containing the learned expressions of all the nodes in the constructed knowledge graph Cheng et al. (2022).

To compute the relation-matching score of the extracted relation and a candidate relation, researchers first obtained the embeddings of the entities in the extracted relation using the RoBERTa model. Later, these embeddings were concatenated with the embeddings of the corresponding entities in the candidate relation obtained from the GCN output. The concatenated embeddings are later forwarded via a feedforward neural network to obtain a scalar score that represents the similarity between the two relations.

Finally, the relation-matching score is combined with the confidence score obtained from the joint extraction technique to obtain the final relation score. Relations with high final scores are selected as the predicted relations for the input sentence. The constructed knowledge graph can also be used for link prediction Li et al. (2018).

3.5. Training and evaluation

The training and evaluation phase is critical to the model creation process. For this, researchers began by dividing the dataset into 80:10:10 training, validation, and testing collections. The training collection is employed for the training of the model, the validation collection is employed for the revision of hyperparameters and also to bypass the overfitting of the model, and the testing collection is employed for the estimation of the model's interpretation of unknown textual data.

To minimize the negative log-likelihood loss function, the Adam algorithm was employed and alpha was chosen to 0.0002 Kohli et al. (2022). The model was trained with exactly 100 epochs before being stopped earlier to prevent overfitting. Researchers assessed the model's interpretations on the validation collection during the training phase and store the best model on the basis of the best F1 score.

For this study, researchers employed precision (P), recall (R), and F1 score measures to assess the model's interpretation. Precision is the ratio of the true positive interpretations to all positive interpretations. The ratio of true positive interpretations to all actual positive occurrences is measured by the recall. And F1 score is actually the harmonic mean of both precision and recall. The macro-average version of the F1 score, actually the average of the F1 scores across all classes, is reported. Fur-

General Relation Triple

Input Sentence: It's reported that EvilFlower group is an APT organization from Pakistan

Tags: O O O S-ORG-1 O O O O O O S-LOC-2

Result: (EvilFlower, comes_from, Pakistan)

Overlap Relation Triple

Input Sentence: DarkHotel utilized a malicious file called Duuzer, to steal sensitive information from various organizations in the healthcare sector.

Tags: S-APT-1 O O O O O S-MF-2 O O O O O O O O O B-ORG-3 E-ORG-3

Result: (DarkHotel, utilizes, Duzzer) (DarkHotel, launches, healthcare sector)

Fig. 3. General and overlap relation triples extraction.

thermore, statistical significance testing was used for the evaluation of the model with other cutting-edge models.

3.6. Entities and relations extraction

The relation triples for the knowledge graph were constructed by using the pairing rules. Algorithm 1 depicts the realization.

Algorithm 1: Formation of relation triples from entities and relations.

Data: tag_sequence x, set_general_entity $E_{gen} = \{\}$, set_overlapping_relational_entity $E_{ove} = \{\}$
Result: knowledge_triple (e_i, r, e_j)
while $E_{gen} \neq \emptyset$ **do**
  $e_i \leftarrow$ Extract next complete entity through boundary tag;
  if e_i *is a general entity* **then**
    Add e_i to E_{gen};
  else if e_i *is an overlapping relational entity* **then**
    Add e_i to E_{ove};
  else
    for $e_j \in (E_{gen} \cup E_{ove})$ **do**
      if e_i *and* e_j *can form a knowledge triple* **then**
        Form the triple (e_i, r, e_j);
    Remove e_i from E_{gen};
return knowledge_triple (e_i, r, e_j)

Fig. 3 depicts the generic relational triple in the first example. The terms "EvilFlower" and "Pakistan" are derived from tagging data. Pairable things are looked for both before and after "EvilFlower." Because this statement contains just one linkable object, the matched result is [EvilFlower, comes-from, Pakistan]. They will no longer be involved in the linking, and the process of extraction for the current phrase is complete.

Fig. 3 depicts the overlapping relational triple in the second example. The files "DarkHotel", "Duuzer", and "healthcare sector" were extracted. First, "DarkHotel" was used to locate linkable entities, and the closest linkable entity "Duuzer" was found. [DarkHotel, utilizes, Duuzer] was the first paired result. As overlapping relation entities cannot actively engage in pairing, researchers avoided "Duuzer" and searched for linkable entities forward and backward using "healthcare sector". Similarly, researchers might obtain the nearest linkable object "healthcare sector" and the second triple [DarkHotel, launches, healthcare sector]. The extraction process was completed when the general entity set became empty.

4. Implementation

4.1. Dataset

This work constructs a web scraper to gather 1098 cyber security reports and articles from Microsoft, Cisco, McAfee, Kaspersky, Fortinet, CrowdStrike, and some similar references. As there are unrelated texts

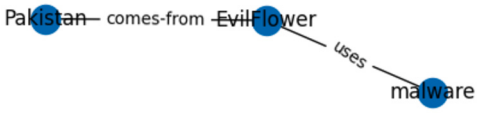

Fig. 4. Knowledge graph constructed from the dataset.

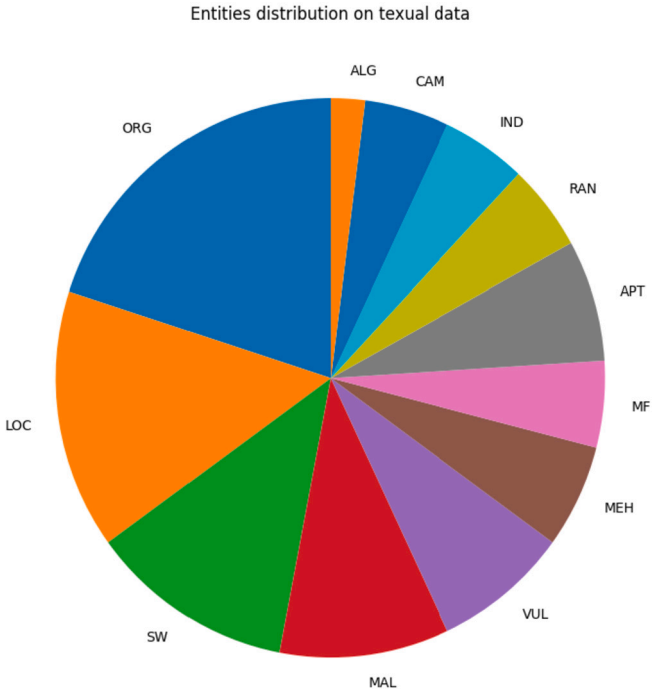

Fig. 5. Entities distribution on textual data.

in the CTI reports, researchers preferably screened them based on the following Husari et al. (2017).

- Article length: Based on researchers' expertise and observations, APT and Ransomware reports are lengthier. They must include detailed descriptions of specific attack occurrences, such as assault purpose, attack mode, and so on. Other reports normally do not have such a large quantity of words.
- Keyword density: It is believed that APT and Ransomware reports include many cyber-security terms, like CVE numbers, hashes, or MAC addresses. As a result, the maximum keywords in each article are counted and their viscosity is determined, which is the proportion of key terms in the article divided by the maximum of words in the report.

Following the screening, 852 APT and Ransomware reports were retained that used regular expressions to remove the URL, IP, and hash information. Researchers only picked the best 100 reports, which were mostly related to Pakistan using keyword filtering techniques, for manual annotation. The entities and relations in APT and Ransomware reports are then tagged using UBIAI UBIAI (2020) which is very similar to YEDDA Yang et al. (2017). Finally, 2049 entities and 1286 relation pairs were labeled. The distribution of entities is also illustrated via a pie chart in Fig. 5. The knowledge graph for the earlier discussed example is shown in Fig. 4

4.2. Experiment

Pytorch created the experimental model in this study. The Adam Kingma and Ba (2014) optimizer was utilized during the training procedure, and the alpha (LR) was selected at 0.0002. Table 3 shows some

Table 3
Details of hyperparameters.

Parameter	Value
total_layers	24
hidden_layer_neurons	1536
hidden_activation_function	ReLU
maximum_position_embedding	1024
LSTM_dimension	250
GRU_dimension	250
total_epochs	100
single_batch_size	128
alpha	0.0002
dropout_ratio	0.5

Table 4
Evaluation of experimental results.

Model	All			Overlap		
	P	R	F1	P	R	F1
CNN-BiLSTM-CRF	0.657	0.635	0.647	0.622	0.634	0.628
CNN-BiGRU-CRF	0.668	0.647	0.659	0.633	0.646	0.640
BERT-BiLSTM-CRF	0.821	0.704	0.758	0.779	0.672	0.721
BERT-BiGRU-CRF	0.843	0.727	0.781	0.801	0.695	0.744
RoBERTa-BiLSTM-CRF	0.872	0.755	0.809	0.830	0.723	0.772
RoBERTa-BiGRU-CRF	0.894	0.778	0.832	0.852	0.746	0.795

of the important parameters, with the dropout set at 0.5 to reduce overfitting.

4.3. Result

When determining if the results of the extraction of cyber entities and relations from cyber security data are accurate, this study concludes that the final outcomes are precise only if the border, category, and of cyber entities and relations are precisely labeled by deep learning. To completely assess the effectiveness of this technique, this work employed different metrics including precision (P), recall (R), and the F1 score, which are typically famous metrics in the domain of natural language processing.

For the purpose of demonstrating the efficacy of this technique, researchers made a comparison to other various standard entity and relation extraction joint extraction methods, like CNN-BiLSTM-CRF, CNN-BiGRU-CRF, BERT-BiLSTM-CRF Miwa and Bansal (2016), BERT-BiGRU-CRF Bekoulis et al. (2018), and RoBERTa-BiLSTM-CRF. Table 4 displays the precision (P), recall (R), and the F1 metrics of entity-relation extracted triples including overlaying extracted relational triples.

According to the results of this study, the attention-based RoBERTa-BiGRU-CRF technique outperformed other methods in the CTI information extraction challenge, achieving an F1 score of 0.832. This technique provides a direct solution to both the extraction of entities and relations, effectively leveraging the semantic connection between the two tasks. When compared to the CNN-BiLSTM-CRF method, the inclusion of the BERT model particularly improved the precision of the task, resulting in an increase of 18.6% in the F1 score. This can be attributed to the BERT's ability to better capture grammatical and semantic nuances across diverse contexts, thereby improving the overall abstraction of the method. While corresponding BERT-BiLSTM-CRF to the proposed method, the final F1 score is improved by 7.4%, indicating that the attention mechanism assisted the decoder module in making the input distinct at various times. Fig. 6 shows the F1 scores of different models used in the entities-relations extraction method.

The training time of different models can vary depending on various elements like dataset, architecture, computing power, etc. The use of pre-trained models can significantly reduce the training time as the model has already learned relevant patterns from a large corpus of data. Overall, the training time is an important consideration when selecting a model, and it is important to balance the complexity of the model with the available computing resources and the desired performance. Fig. 7 shows the training time (in hours) of different models used in the entities-relations extraction method.

4.4. Discussion

When compared to other models, the proposed model's F1 score rose to varied degrees. This is due to the fact that collaborative methods normally extract entities initially followed by relations that can lead to issues like duplicate entities or transmission mistakes. Overlapping relational extraction has a lower precision (P), and recall (R) than total non-overlapping extraction of relation triples. The fundamental reason for this is that certain overlapping interactions are ambiguous, and the same thing in one phrase may have many expressions, making it difficult to identify the model. As a result, the joint extraction model suggested in this study has the potential for further development.

Entities and relations extraction is a domain-specific task that involves identifying entities and the relationships between them in a given text. Various models have been proposed for this task, but not all models perform equally well. In particular, CNN-based models have been found to be less useful for entities and relations extraction, which may be due to their better performance on visual data rather than textual data.

On the other hand, transformer-based models like BERT have shown good performance, but RoBERTa is found to perform better than BERT. This is primarily because RoBERTa improves on BERT by pretraining on a larger and more diverse dataset. In this study, researchers found that RoBERTa outperformed XLNet, another transformer-based model, on the custom dataset.

In this study, both BiLSTM and BiGRU networks were used for the experiments, and researchers found that BiGRU gave slightly better results. This may be because the training data was small and custom, rather than a public dataset. In general, BiGRU models tend to achieve better results on small datasets, and BiLSTM models achieve better results on large datasets. BiGRU models are also computationally less expensive and more efficient in terms of training.

Furthermore, attention-based models were found to perform better than non-attention-based models. Finally, the experiments showed that RoBERTa-BiGRU-CRF gave the highest F1 score, which was a 7% increase over other techniques that were tested.

To attain informatics and explanations for the decision-making process of the RoBERTa-BiGRU-CRF model, various explainable artificial intelligence (XAI) techniques were utilized Zhang et al. (2022). Specifically, the attention visualization technique was employed to visualize which parts of the input text were most important in the model's decisions. Researchers also utilized LIME (local interpretable model-agnostic explanations) to generate local explanations for individual predictions, and SHAP (shapley additive explanations) to compute feature importance scores for each input token. These techniques allowed researchers to acquire a more reasonable interpretation of how the model arrived at its predictions and provided valuable insights for improving the model's performance.

Finally, one of the crucial challenges in CTI is the ability to extract and represent relevant knowledge. The OpenCTI dashboard OpenCTI (2019) provides a promising solution to this challenge by providing a centralized platform for CTI analysts to collaborate, manage, and share CTI knowledge. In this research, a novel technique for constructing a knowledge graph from extracted CTI triples using the OpenCTI dashboard is proposed. As depicted by the notable results, the proposed technique has the potential to significantly enhance the effectiveness and efficiency of CTI knowledge management, and thereby contribute to the overall security of cyberspace. Dashboard, as an example, populating cyber threat intelligence is also illustrated in Fig. 8.

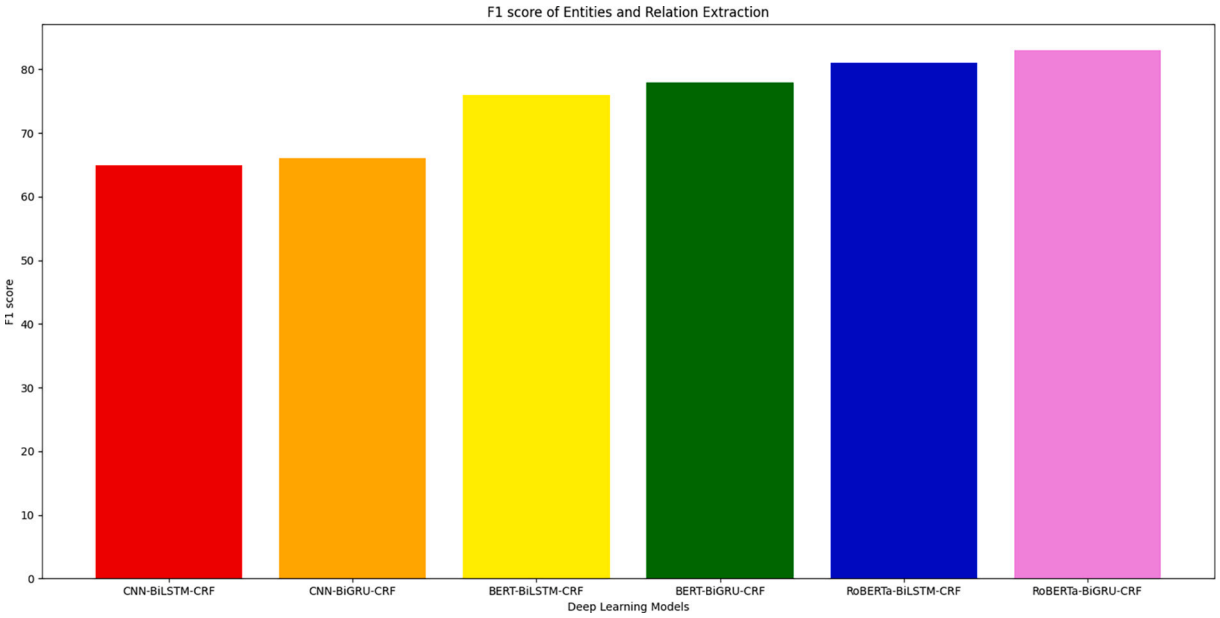

Fig. 6. F1 scores of entities and relation extraction.

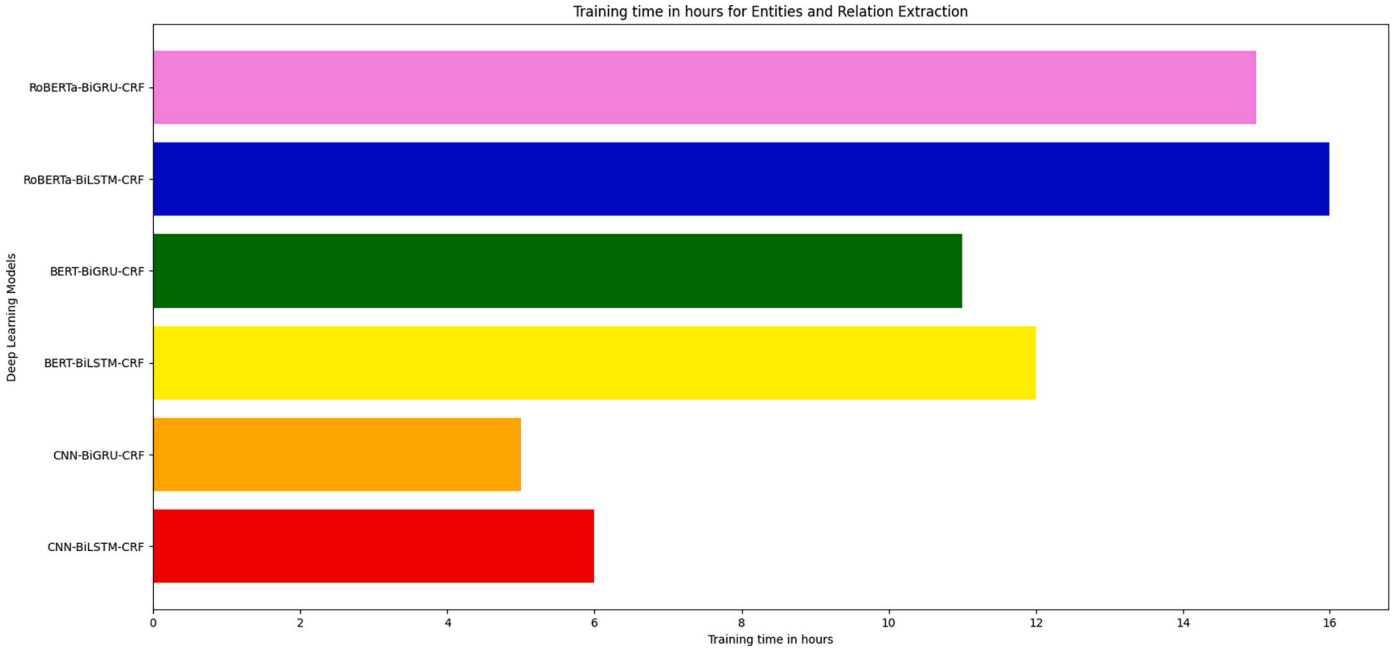

Fig. 7. Training time for entities and relation extraction.

4.5. Future work

Some possible directions for future research are listed in this section:

- Scaling up the proposed technique to handle even larger CTI data.
- Exploring the effectiveness of the proposed technique in different languages and domains.
- Developing a user-friendly tool for cybersecurity professionals to extract and integrate CTI from various sources using the proposed technique.
- Investigating the possibility of integrating machine learning and NLP techniques to improve the effectiveness and efficiency of CTI extraction.
- Evaluating the proposed technique on real-world CTI data to validate its effectiveness in practical scenarios.
- Developing techniques to automatically update and maintain the CTI knowledge graph to keep it up-to-date with emerging threats.
- Evaluating the real-world performance and impact of the proposed technique by conducting practical assessments of the model's efficiency and its ability to support decision-making in the field of cybersecurity.
- Exploring the correlation between model scores and the actions taken in response to its recommendations.
- Collaborating with cybersecurity practitioners to incorporate their feedback and insights will be crucial in refining the model and aligning it with the needs of industrial cybersecurity professionals.

Fig. 8. Cyber threat intelligence dashboard.

5. Conclusion

This research proposed a novel technique that is used for extracting CTI from unstructured text using a combination of joint extraction and relation-matching techniques. The RoBERTa-BiGRU-CRF model was developed for entity-relation extraction and attained state-of-the-art results on the CTI corpus. The evaluation of the model showed promising results, indicating its effectiveness in extracting CTI from unstructured text. The technique was tested on different scenarios related to APT and ransomware attacks, demonstrating its capability of handling complex CTI extraction tasks. Overall, this research contributes to the growing field of CTI extraction and provides a foundation for future work in this area. The proposed technique can be expanded to other related domains and can be used to automatically extract and integrate CTI from various sources, including social media, the dark web, and open-source intelligence. The size of CTI data can be increased in the future for diverse investigations. The proposed technique can also aid in the development of more effective and efficient CTI sharing and collaboration platforms, enabling cybersecurity professionals to stay ahead of emerging threats and better protect against cyberattacks.

CRediT authorship contribution statement

Kashan: Conceptualization, Data curation, Software, Formal Analysis, Validation, Writing – Original draft preparation. **Khaldoon:** Conceptualization, Supervision, Resources. **Sadaf:** Conceptualization, Visualization, Investigation, Supervision, Writing – Reviewing and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Ainslie, S., Thompson, D., Maynard, S., Ahmad, A., 2023. Cyber-threat intelligence for security decision-making: a review and research agenda for practice. Comput. Secur. 132, 103352. https://doi.org/10.1016/j.cose.2023.103352.

Alves-Pinto, A., Demus, C., Spranger, M., Labudde, D., Hobley, E., 2022. Iterative named entity recognition with conditional random fields. Appl. Sci. 12, 330.

Bayer, M., Kuehn, P., Shanehsaz, R., Reuter, C., 2022. Cysecbert: a domain-adapted language model for the cybersecurity domain. arXiv preprint. arXiv:2212.02974.

Bekoulis, G., Deleu, J., Demeester, T., Develder, C., 2018. Joint entity recognition and relation extraction as a multi-head selection problem. Expert Syst. Appl. 114, 34–45.

Chen, Z., Guo, C., 2022. A pattern-first pipeline approach for entity and relation extraction. Neurocomputing 494, 182–191.

Cheng, H., Liao, L., Hu, L., Nie, L., 2022. Multi-relation extraction via a global-local graph convolutional network. IEEE Trans. Big Data 8, 1716–1728.

Choo, K.K.R., Conti, M., Dehghantanha, A., 2019a. Special issue on big data applications in cyber security and threat intelligence–part 1. IEEE Trans. Big Data 5, 279–281.

Choo, K.K.R., Conti, M., Dehghantanha, A., 2019b. Special issue on big data applications in cyber security and threat intelligence–part 2. IEEE Trans. Big Data 5, 423–424.

Conti, M., Dargahi, T., Dehghantanha, A., 2018. Cyber Threat Intelligence: Challenges and Opportunities. Springer.

Dimitriadis, A., Prassas, C., Flores, J.L., Kulvatunyou, B., Ivezic, N., Gritzalis, D.A., Mavridis, I.K., 2021. Contextualized filtering for shared cyber threat information. Sensors 21, 4890.

Ge, W., Wang, J., 2022. Seqmask: behavior extraction over cyber threat intelligence via multi-instance learning. Comput. J.

Ghazanfar, I., Abbas, H., Iqbal, W., Rashid, I., 2021. Vulnerability assessment of Pakistan government websites. In: 2021 International Conference on Communication Technologies (ComTech). IEEE, pp. 115–119.

Guo, Y., Liu, Z., Huang, C., Liu, J., Jing, W., Wang, Z., Wang, Y., 2021. Cyberrel: joint entity and relation extraction for cybersecurity concepts. In: Information and Communications Security: 23rd International Conference. ICICS 2021, Chongqing, China, November 19-21, 2021, Proceedings, Part I 23. Springer, pp. 447–463.

Guo, Y., Liu, Z., Huang, C., Wang, N., Min, H., Guo, W., Liu, J., 2023. A framework for threat intelligence extraction and fusion. Comput. Secur. 132, 103371. https://doi.org/10.1016/j.cose.2023.103371.

Hanks, C., Maiden, M., Ranade, P., Finin, T., Joshi, A., et al., 2022. Recognizing and extracting cybersecurity entities from text. In: Workshop on Machine Learning for Cybersecurity, International Conference on Machine Learning.

Husari, G., Al-Shaer, E., Ahmed, M., Chu, B., Niu, X., 2017. Ttpdrill: automatic and accurate extraction of threat actions from unstructured text of CTI sources. In: Proceedings of the 33rd Annual Computer Security Applications Conference, pp. 103–115.

Jo, H., Lee, Y., Shin, S., 2022. Vulcan: automatic extraction and analysis of cyber threat intelligence from unstructured text. Comput. Secur. 120, 102763.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint. arXiv:1412.6980.

Kohli, H., Agarwal, J., Kumar, M., 2022. An improved method for text detection using Adam optimization algorithm. Glob. Trans. Proc. 3, 230–234.

Li, M., Wang, Y., Zhang, D., Jia, Y., Cheng, X., 2018. Link prediction in knowledge graphs: a hierarchy-constrained approach. IEEE Trans. Big Data 8, 630–643.

Li, Y., Guo, Y., Fang, C., Liu, Y., Chen, Q., et al., 2022a. A novel threat intelligence information extraction system combining multiple models. Secur. Commun. Netw. 2022.

Li, Z., Zeng, J., Chen, Y., Liang, Z., 2022b. Attackg: constructing technique knowledge graph from cyber threat intelligence reports. In: Computer Security–ESORICS 2022: 27th European Symposium on Research in Computer Security. Copenhagen, Denmark, September 26–30, 2022, Proceedings, Part I. Springer, pp. 589–609.

Liu, J., Yan, J., Jiang, J., He, Y., Wang, X., Jiang, Z., Yang, P., Li, N., 2022. Tricti: an actionable cyber threat intelligence discovery system via trigger-enhanced neural network. Cybersecurity 5, 8.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: a robustly optimized bert pretraining approach. arXiv preprint. arXiv:1907.11692.

Liu, Z., Li, H., Wang, H., Liao, Y., Liu, X., Wu, G., 2023. A novel pipelined end-to-end relation extraction framework with entity mentions and contextual semantic representation. Expert Syst. Appl. 228, 120435. https://doi.org/10.1016/j.eswa.2023.120435.

Lu, Y., Yang, R., Jiang, X., Zhou, D., Yin, C., Li, Z., 2021. Mre: a military relation extraction model based on bigru and multi-head attention. Symmetry 13, 1742.

Marchiori, F., 2021/2022. STIXnet: entity and relation extraction from unstructured CTI reports.

Miwa, M., Bansal, M., 2016. End-to-end relation extraction using lstms on sequences and tree structures. arXiv preprint. arXiv:1601.00770.

Noravesh, F., 2023. Semantic tagging with lstm-crf. arXiv preprint. arXiv:2301.12206.

OpenCTI, 2019. GitHub - OpenCTI-platform/opencti: open cyber threat intelligence platform — github.com. https://github.com/OpenCTI-Platform/opencti. (Accessed 1 March 2023).

Pingle, A., Piplai, A., Mittal, S., Joshi, A., Holt, J., Zak, R., 2019. Relext: relation extraction using deep learning approaches for cybersecurity knowledge graph improvement. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 879–886.

Poostchi, H., Piccardi, M., 2019. Bilstm-ssvm: training the bilstm with a structured hinge loss for named-entity recognition. IEEE Trans. Big Data 8, 203–212.

Ren, Y., Xiao, Y., Zhou, Y., Zhang, Z., Tian, Z., 2022. Cskg4apt: a cybersecurity knowledge graph for advanced persistent threat organization attribution. IEEE Trans. Knowl. Data Eng.

Sarhan, I., Spruit, M., 2021. Open-cykg: an open cyber threat intelligence knowledge graph. Knowl.-Based Syst. 233, 107524.

Srivastava, S., Paul, B., Gupta, D., 2023. Study of word embeddings for enhanced cyber security named entity recognition. Proc. Comput. Sci. 218, 449–460. https://doi.org/10.1016/j.procs.2023.01.027, International Conference on Machine Learning and Data Engineering.

Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y., Zhang, J., 2023. Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives. IEEE Commun. Surv. Tutor. 25 (3), 1748–1774. https://doi.org/10.1109/COMST.2023.3273282.

Tang, M., Alazab, M., Luo, Y., 2017. Big data for cybersecurity: vulnerability disclosure trends and dependencies. IEEE Trans. Big Data 5, 317–329.

UBIAI, 2020. Easy to use text annotation tool | upload documents, start annotating, and create advanced NLP model in a few hours. – ubiai.tools. https://ubiai.tools/. (Accessed 13 March 2023).

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

Wang, X., Liu, J., 2023. A novel feature integration and entity boundary detection for named entity recognition in cybersecurity. Knowl.-Based Syst. 260, 110114. https://doi.org/10.1016/j.knosys.2022.110114.

Wang, X., Liu, R., Yang, J., Chen, R., Ling, Z., Yang, P., Zhang, K., 2022b. Cyber threat intelligence entity extraction based on deep learning and field knowledge engineering. In: 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, pp. 406–413.

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

Wang, X., Xiong, M., Luo, Y., Li, N., Jiang, Z., Xiong, Z., 2020b. Joint learning for document-level threat intelligence relation extraction and coreference resolution based on gcn. In: 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, pp. 584–591.

Watters, P.A., 2023. Counterintelligence Theory. Springer International Publishing, Cham, pp. 1–17.

Yang, J., Zhang, Y., Li, L., Li, X., 2017. Yedda: a lightweight collaborative text span annotation tool. arXiv preprint. arXiv:1711.03759.

Yang, X., Xiao, Y., 2022. Named entity recognition based on bert-mbigru-crf and multi-head self-attention mechanism. In: 2022 4th International Conference on Natural Language Processing (ICNLP). IEEE, pp. 178–183.

Yi, F., Jiang, B., Wang, L., Wu, J., 2020. Cybersecurity named entity recognition using multi-modal ensemble learning. IEEE Access 8, 63214–63224.

Zhang, Z., Hamadi, H.A., Damiani, E., Yeun, C.Y., Taher, F., 2022. Explainable artificial intelligence applications in cyber security: state-of-the-art in research. arXiv preprint. arXiv:2208.14937.

Zhang, Z., Sind, X., Liu, T., Fang, Z., Li, Q., 2020. Joint entity linking and relation extraction with neural networks for knowledge base population. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.

Zhao, F., Gui, X., Huang, Y., Jin, H., Yang, L.T., 2020a. Dynamic entity-based named entity recognition under unconstrained tagging schemes. IEEE Trans. Big Data 8, 1059–1072.

Zhao, J., Yan, Q., Liu, X., Li, B., Zuo, G., 2020b. Cyber threat intelligence modeling based on heterogeneous graph convolutional network. In: RAID, pp. 241–256.

Zhong, L., Wu, J., Li, Q., Peng, H., Wu, X., 2023. A comprehensive survey on automatic knowledge graph construction. arXiv preprint. arXiv:2302.05019.

Zhou, Y., Ren, Y., Yi, M., Xiao, Y., Tan, Z., Moustafa, N., Tian, Z., 2023. Cdtier: a Chinese dataset of threat intelligence entity relationships. IEEE Trans. Sustain. Comput.

Zhou, Y., Tang, Y., Yi, M., Xi, C., Lu, H., 2022. CTI view: Apt threat intelligence analysis system. Secur. Commun. Netw. 2022, 1–15.

Zongxun, L., Yujun, L., Haojie, Z., Juan, L., 2021. Construction of ttps from apt reports using bert. In: 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE, pp. 260–263.

Zuo, J., Gao, Y., Li, X., Yuan, J., 2022. An end-to-end entity and relation joint extraction model for cyber threat intelligence. In: 2022 7th International Conference on Big Data Analytics (ICBDA). IEEE, pp. 204–209.