JDEBM
Journal of Digital Engineering and Business Managemet

Volume-2, Issue-1,January-March 2026
ISSN: 3107-7897 | P.ID: JDEBM/11
DOI: https://doi.org/10.5281/zenodo.19083510

DATA-DRIVEN CYBER THREAT INTELLIGENCE USING NETWORK-BASED MODELING TECHNIQUES

[#1]K.RAVI, *Assistant Professor,*
[#2]JALI SRAVANI, *Assistant Professor,*
Department of Computer Science & Engineering,
MOTHER THERESSA COLLEGE OF ENGINEERING & TECHNOLOGY,
PEDDAPALLI, TELANGANA.

ABSTRACT: The increase in cyberattacks has made the cybersecurity problem worse. To navigate the ever-changing and intricate cyber landscape, cyber threat intelligence is essential. Since most cyber threat intelligence is unstructured, security analysts have a hard time keeping up with the massive amounts of data. Entity extraction, co-reference resolution, relation extraction, and knowledge graph building are the four cornerstones of the new approach to threat intelligence information extraction that is proposed in this research. A number of models are employed in this process. In order to extract the word dependence relationships for the entity extraction task, a multihead self-attention strategy is used. To enhance mention representation, the co-reference resolution method incorporates both mention embedding and contextual information. A convolutional neural network can retrieve features from multiple dimensions at once. The relation extraction task enhances the embedding representation by incorporating entity type, distance between entity pairs, part of speech, mention breadth, and relational distance. Lastly, a knowledge graph is created to formally define things and their interactions. Our model outperforms the baseline model in entity extraction (F1 score of 8.87), coreference resolution (F1 score of 9.82), and relation extraction (F1 score of 10.56). Our technology is demonstrated by Neo4j's knowledge graph.

KEYWORDS: *Cyber Threat Intelligence (CTI),Data-Driven Security, Network-Based Modeling, Threat Modeling,*

1. INTRODUCTION

Cyberattacks on contemporary information technology systems can result in varying degrees of damage. The only way to resolve this issue is to constantly monitor devices, gather and evaluate data, and document security incidents. Security professionals recommend Cyber Threat Intelligence (CTI) as a tool for consumers to better understand the threat landscape and prepare for unknown assaults.

Threat intelligence is defined as information on potential or actual risks to an asset, according to a Gartner researchfrom 2013. Various sources can be consulted for threat intelligence, including methodologies, indications, hypothetical scenarios, disclosures, and proposed solutions. Intrusion detection and monitoring systems and threat intelligence platforms differ primarily in two respects. In contrast, cyber threat intelligence is compiled following an assault and the subsequent reaction. Decisions about risk management could

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-1,January-March 2026
ISSN: 3107-7897 | P.ID: JDEBM/11
DOI: https://doi.org/10.5281/zenodo.19083510

be aided by providing access to expert information through its utilization. Intrusion detection systems will emit a sound when an intrusion has taken place. Conversely, security provider alerts, internet discussion forums, and social media can all give danger intelligence. It is helpful to know how hacks work and what triggers them.

Threat intelligence has grown substantially despite the fact that the Internet is complex, there are various ways to attack it, and there are many security measures available.

The use of plain English with strategically placed key words and complex hyperlinks is becoming increasingly common when disseminating cyber threat intelligence. This makes data collection, analysis, and sharing more challenging. Due to the high volume of warnings, security professionals are swamped with work. This leads many warnings to go unheeded and forgotten. Management and analysis of threat information must be addressed promptly.

Human analysis of threat data requires extensive training and experience in cyber defense. Because of this, fighting the increasingly frequent attacks becomes more difficult. Due to its critical importance, extracting structured knowledge from unstructured threat data has been the subject of much research. Four primary procedures comprise this extraction method: building knowledge graphs, relation extraction, entity extraction, and co-reference resolution.

In order to use automated threat intelligence, the following problems must be fixed: The threat intelligence field is distinct from others in important ways. Using a general subject entity extraction method could make it difficult to identify threat entities such as cyber organizations, attack strategies, malware, etc.Multiple mentions of the same company are possible in a threat intelligence report. Find out if the references are referring to the same item by looking at the situation holistically and getting semantic information.Sometimes it's hard to understand the whole meaning of a threat report. To grasp the interconnectedness of concepts, it usually requires more than a single paragraph.

Access to threat information datasets is limited.To improve the collection of threat intelligence, this research suggests a different method. Every knowledge graph needs four steps: "pulling entities," "resolving co-references," "extracting relationships," and "building." We have resolved these issues with this solution.

While Zhou et al. could only obtain related bits of APT threat intelligence, they did come up with a method to gather it all. Semantic connections between descriptive and static CTI data obtained from unstructured text were presumed by Vulcan. Nevertheless, they fail to provide a comprehensive overview of the interconnected networks and organizations uncovered by threat intelligence. A brief summary of the key findings from this research is shown here.

The utilization of a hybrid technique allows for a deeper understanding of threat intelligence. The model is able to construct a knowledge network and sift through huge amounts of unstructured threat intelligence data. A knowledge graph that shows the interconnected

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-1,January-March 2026
ISSN: 3107-7897 | P.ID: JDEBM/11
DOI: https://doi.org/10.5281/zenodo.19083510

components of threat intelligence is updated via the Neo4j graph database. Experts in security can use this data and the decisions it gives them to better understand risks and strengthen defenses.

The EEMAP method is a novel approach to entity extraction that makes use of POS and multiple attention colors. By including "multithread self-attention," the model may provide entity-related vector representations. The feature vectors of a model recurrent neural network are then enhanced with vector representations. Text entities are rediscovered after submitting the combined findings to a linear layer for sequence labeling.

Co-Reference Resolution with CNN and POS (CRCP) is a new approach to resolving co-reference issues. By incorporating contextual data into the embedding process, the method enhances mention representation. Classical co-reference resolution methods have a low recall rate, however a convolutional neural network can acquire some of the aforementioned features to compensate.

2. RELATEDWORK

The majority of this part comprises prior studies on our topic. A comprehensive examination of the most sophisticated techniques for co-reference resolution, entity extraction, and connection identification is presented.

Entity Extraction

The entity extraction method is essential to any NLP system. To convert unstructured text into semi-structured or structured data, it identifies references to individuals, locations, and enterprises within the text. This facilitates the examination of the text from multiple perspectives. The entity extraction process is predominantly dependent on name sequences.

Rules and dictionaries were crucial for the initial research on entity extraction. Experts developed numerous rule templates, with string matching as the primary method of utilization. This method is effective for datasets exhibiting patterns, however it is not applicable to all datasets and fails to accommodate novel occurrences. In the realm of cybersecurity, statistical machine learning can enhance entity extraction. This method may be applied universally, eliminating the necessity of manually generating rule files. Statistical machine learning employs models such as the Conditional Random Field (CRF), the Maximum Entropy Model (MEM), and the Hidden Markov Model (HMM). The CRF technique was proposed by Joshi et al. to develop a new cybersecurity ontology through the extraction of concepts from cybersecurity-related announcements and blogs.

Mulwad et al. employed support vector machines (SVMs) for prediction in order to gain a deeper understanding of the inner workings of assaults and the harm they do. To prevent the MEM from achieving excessive accuracy during training, Bridges et al. evaluated it on various security-related corpora. Although more accurate than rule-based approaches, the aforementioned strategies are challenging to implement with small datasets due to the extensive text feature extraction required.

Deep learning is heavily utilized, with neural networks capable of managing previously challenging tasks such as "entity extraction." For instance, the model developed by Chiu and Nichol employs CNN and BiLSTM to identify features between tokens and characters. They

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-1,January-March 2026
ISSN: 3107-7897 | P.ID: JDEBM/11
DOI: https://doi.org/10.5281/zenodo.19083510

developed an innovative method for word encoding and a corresponding matching mechanism.

The inaugural application of CNN to analyze tweets for insights regarding IT system security was conducted by Dionisio et al. They subsequently employed BiLSTM to identify prominent organizations and receive alerts regarding potential risks. Through dependency analysis, Wu et al. identified methodologies, resources, and entities in e-commerce threat intelligence utilized in evolving assault patterns.

Gasmi et al. employed a BiLSTM-CRF model for the extraction of cybersecurity-related concepts. Subsequently, it was juxtaposed with three models based on LSTM. Zhao and colleagues employed a multigranular attention methodology to illustrate the functioning of inter-IOC interactions by constructing a network comprising various sorts of information. This enabled us to attain more precise results.

Co referenceResolution

The resolution of coreference will indicate whether the two reference pairs are connected. When extracting relationships from documents, an item may manifest many instances. In this instance, two references that denote the same entity are termed "coreferential." The virus clandestinely downloaded and installed ister59.apk, which is recognized as dangerous, as seen by the subsequent URL. Both Android.Reputation and Ister59.apk identified the identical bug. The ability to resolve coreferences is essential for effective communication, comprehending abstract concepts, and identifying causal relationships.

Relation Extraction

The initial research on "relationship extraction" focused on comprehending the links between two components at the phrase level. To accomplish document-level relation extraction, it is necessary to consolidate relationship information from many sources.

The two predominant techniques for relation extraction at the document level are transformation-based and graph-based methods. Graph-based methodologies generate graphs from textual data to provide a clearer elucidation of entity hierarchies. Zeng et al. created document graphs at the mention and entity levels and proposed an innovative path inference mechanism to determine the links among entities. Sun et al. introduced a dual-channel hierarchical graph convolutional neural network (DHGCN) to elucidate the intricate relationships among token-level, mention-level, and entity-level components inside a twenty-interpretation document. Transformer-based techniques assign a representation to each word in a document with pre-trained models (e.g., BERT, RoBERTa, ERNIE, etc.). Yuan et al. developed a gating function to integrate sentence-level information with document-level data, employing an inter-sentence attention mechanism to dynamically consolidate essential sentence elements. Zhang et al. developed a U-shaped segmentation module to aggregate local and global data for predicting entity-level association vectors. Zhou et al. proposed employing adaptive thresholding to minimize optimization expenses and local context pooling to improve entity embedding, thereby tackling multiunit and multilevel challenges in document-level relation extraction.

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-1,January-March 2026
ISSN: 3107-7897 | P.ID: JDEBM/11
DOI: https://doi.org/10.5281/zenodo.19083510

3. METHODOLOGY

Model Framework

The method we propose for the collection of threat intelligence data in this research is unique and is based on current knowledge. Named entity identification, coreference resolution, document-level connection extraction, and knowledge graph construction are all accomplished by the suggested system through the use of NLR, CLR, DLR, and KG. The system's complete configuration is shown in Figure 1. At first, the data is transformed into embeddings that improve part-of-speech labeling using the BERT and NLTK Python libraries. A named entity identification model, a coreference resolution model, and a document level relationship detector model all make use of the integrated data. Once sorting is complete, the results are added to the knowledge tree in three sets of three.

Encoding Layer

The traditional encoding layer in this method relies on randomly embedded words, however BERT provides a deep comprehension of word semantics, therefore it is replaced. The effectiveness of mention embedding in text description is improved by adding pieces of speech. At the beginning of the text, you'll see two designated tags: "[CLS]" and "[SEP]". As the encoder, we have the pretrained model BERT. Before and after every instance in the text, the identifying " "appear. D xt l 1 is the result of tokenizing the document, where xt is the word at position t (see Figure 1). The H representation of context is generated by encoding this area of the page using BERT-base.

$$H = \mathrm{BERT}([x_1, \ldots, x_l]) = [h_1, \ldots, h_l], \qquad (1)$$

The hidden size, denoted by d1, is added to Rl to give H. The order of the parts of speech in the text is determined using NLTK. The process of creating the POS embedding matrix P is described in the following steps:

$$P = \mathrm{Pos}([x_1, \ldots, x_l]) = [p_1, \ldots, p_l], \qquad (2)$$

Depending on the situation, contextual and POS embedding improve the word representation of each token. Each token's point-of-sale embedding has dimensions d2 and P Rld2.

$$C = [h_1 \circ p_1, \ldots, h_l \circ p_l] = [c_1, \ldots, c_l], \qquad (3)$$

where $C \in R^{l \times (d_1 + d_2)}$, and $\circ$ indicates the linking operation.

Entity Extraction

Vector representations of relevant items are obtained using our entity extraction methodology's multihead self-attention mechanism. By giving each token representation a unique weight, this method may find out how important the relationships are between every pair of words. One way to speed up model training is to use several attention heads to find features in different representational subspaces. This method gives the attention layer a set of token representations that are enriched with point-of-sale information, which it can use to correctly embed the current term.

$$\mathrm{head}_i = \mathrm{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V,$$

$$A = \mathrm{MultiHead}_i(Q, K, V) \qquad (4)$$

$$= \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_{H_1}) W_A,$$

.

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-1,January-March 2026
ISSN: 3107-7897 | P.ID: JDEBM/11
DOI: https://doi.org/10.5281/zenodo.19083510

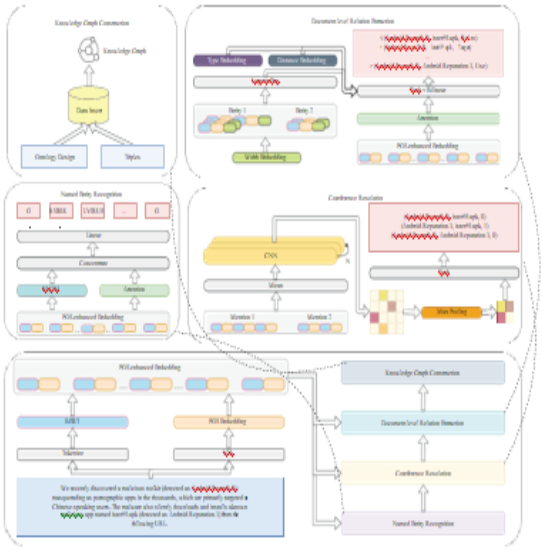


Figure 1:Theproposedthreatintelligenceinformationextractionsystem.

A sequence of questions is represented by Q, a sequence of keys by K, and a sequence of values by V. V is the value sequence notation and dk is the key sequence dimensionality.

BiLSTM has been demonstrated to be effective in capturing contextual semantic information, and this research recommends its use to enhance comprehension of the current word's previous and future applications. Layers that are interconnected and have both forward and backward LSTMs make up a BiLSTM network. Subnetworks, which are collections of memory nodes, make up each LSTM. An LSTM storage module is built at each moment using the current input word embedding, the hidden vector from the moment before, and the storage cell vector from the moment before that.

Once the BiLSTM has been fed a succession of token representations augmented with POS, it will generate a feature vector.

$$B = \mathrm{BiLSTM}[c_1, \ldots, c_1] = \left[\vec{b}_1 \circ \overleftarrow{b}_1, \ldots, \vec{b}_1 \circ \overleftarrow{b}_l\right]. \quad (5)$$

The sequence labels are assigned using a linear classifier that processes the aggregated feature vectors and relevant contextual embeddings.Simply said,

$$y_{ner} = \arg\max\left[W_1 (A \circ B) + b_1'\right]. \quad (6)$$

Coreference Resolution.

"Resolving a coreference" means checking to see if two references to the same thing really mean the same thing. This research takes a classification problem method to solving cross-references.For each mention token, the model calculates an average vector, and then it generates variations with POS enhancements.

$$\begin{aligned} \mathrm{mention}_1 &= \mathrm{Mean}\left(c_{1_B}, \ldots, c_{1_I}\right), \\ \mathrm{mention}_2 &= \mathrm{Mean}\left(c_{2_B}, \ldots, c_{2_I}\right). \end{aligned} \quad (7)$$

CNN offers assistance to those who rely on remote sources by acquiring depth information through a sliding window. In order to do convolutional operations on word vectors, a convolutional layer usually just has one layer that uses a convolution kernel.After representations are compressed and reduced, a pooling layer in the convolutional architecture removes superfluous data and prevents overfitting. This new technique ignores all but the most prominent features and instead uses the feature values produced by each layer after the convolution layer to make a selection.

$$\begin{aligned} \mathrm{Mention-Pair}_i &= \mathrm{Conv}_i\left(\mathrm{mention}_1 \cdot \mathrm{mention}_2\right), \\ M &= \mathrm{Concat}\left(\ldots, \ldots, \mathrm{Mention-Pair}_N\right) W_M, \\ \mathrm{MP} &= \mathrm{MaxPooling}(M). \end{aligned} \quad (8)$$

Once the combined feature vector of mention pairs has been obtained, the tanh activation function is used to compute the label probability, which indicates whether two mentions pertain to the same entity.

$$y_{\mathrm{CR}} = \tanh\left(W_2 \cdot \mathrm{MP} + b_2'\right). \quad (9)$$

The proper occurrences are identified by

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-1,January-March 2026
ISSN: 3107-7897 | P.ID: JDEBM/11
DOI: https://doi.org/10.5281/zenodo.19083510

using the sequence labels generated by the entity extraction model from the section. The mentions are checked to see if they relate to the same thing using the coreference resolution procedure.

4. RESULTANALYSIS

Performance on Entity Extraction

Ablation tests were performed to evaluate the effect of each module on the overall performance of the model. Table 1 presents the comparative efficacy of our EEMAP-BERT model relative to the predominant baselines for the entity extraction task. EEMAP-WE employs a random word embedding as its encoder, in contrast to the conventional BERT approach. Our model markedly surpassed the BiLSTM and BiLSTM-CRF baselines in precision, memory efficiency, F1 score, and exact-match accuracy. The overall F1 score rose by 9.94 points and 8.87 points, respectively.

Upon the removal of POS embedding (designated as NoPOS), the F1 score decreased by 0.67 points, and the exact-match accuracy diminished by 0.74 points. Threat intelligence indicates that these two embedding types significantly impact the model's performance, as entities predominantly comprise words and verbs.

The subsequent layer to be removed was the multihead self-attention layer. Reports indicate that the F1 score and exact-match precision diminished by 0.56 and 0.42, respectively. The experiment's results indicate that the multihead self-attention mechanism enhances the comprehension of remote, interrelated things and facilitates the identification of significant context.

The multihead self-attention layer and POS integration were subsequently disabled (labeled No POS + No Attention). The 2.76-point reduction in the F1 score and the 3.45-point reduction in the exact-match accuracy score illustrate a substantial decrease in performance. We also investigated the efficacy of BERT and random word embeddings to understand the impact of the model encoder on performance.

Arbitrarily incorporating words into models was discovered to markedly diminish their accuracy. A significant decrease of 9.65 and 13.01 points occurred in the F1 and exact-match scores, respectively. The trials validated that a machine trained on a substantial corpus could identify generic language representations. Consequently, it was better suited to multitask and achieve its objectives more rapidly.

Figure 4 illustrates that the utilization of five attention heads results in an enhancement of both the F1 score and the exact-match accuracy. We have H2 rated at five due to this reason. Figure 5 delineates the nuanced performance for a certain category. This enables the examination of the efficacy of the POS embedding and attention mechanism across various data kinds.

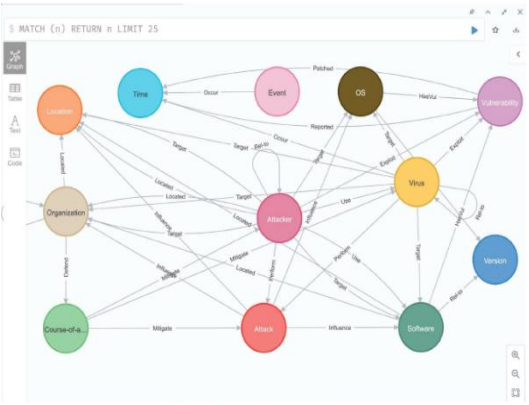
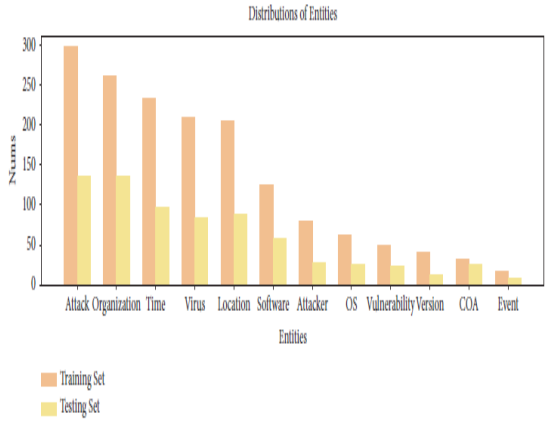


FIGURE 2: Threat intelligence ontology.

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-1,January-March 2026
ISSN: 3107-7897 | P.ID: JDEBM/11
DOI: https://doi.org/10.5281/zenodo.19083510



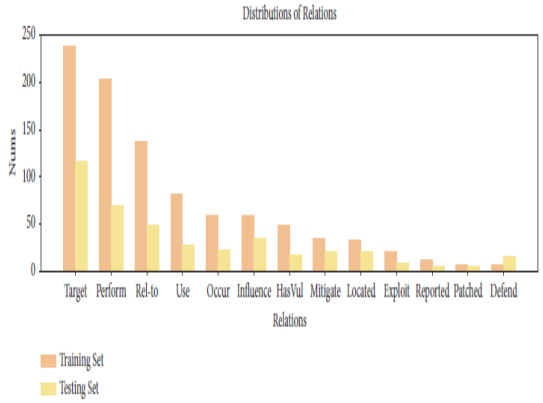


FIGURE 3: The distribution of entities and relationships.

TABLE I: The performance of the entity extraction task.

Model	Precision	Recall	F1	Accuracy
EEMAP-WE	69.23	67.71	68.46	61.12
BiLSTM [11]	69.80	66.62	68.17	62.11
BiLSTM-CRF [13]	70.10	68.40	69.24	62.77
EEMAP-BERT (our model)	79.02	77.22	78.11	74.13
No POS	78.43	76.46	77.44	73.39
No attention	78.56	76.84	77.69	73.57
No POS+no attention	75.97	74.74	75.35	70.68

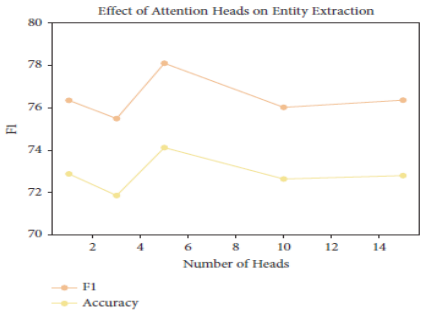
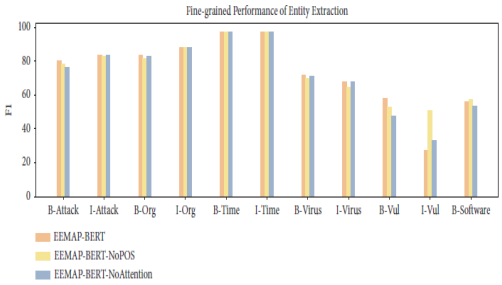


FIGURE 4: The effect of attention heads on entity extraction.



(a)

FIGURE 5: Continued.

5. CONCLUSION

The present investigation presents a hybrid method for extracting threat intelligence information that takes into account the interactions between knowledge graph construction, entity extraction, coreference resolution, and connection extraction.

The multihead self-attention mechanism is integrated into the entity extraction model to facilitate the identification of pertinent contextual vectors. Coreference resolution employs environmental information and reference encoding. A convolutional neural network integrates information from multiple layers to identify features. Subsequent aspects, such as entity pair identification and the segmentation of parts of speech and reference width, are incorporated into the relation extraction model. Our model enhances entity extraction by 8.87 points, coreference resolution by 9.82 points, and relation extraction by 10.56 points in F1 score relative to baseline comparisons, as demonstrated by studies.

The subsequent phase involves constructing a threat intelligence knowledge graph that will illustrate potential semantic links among components. Finally, our methodology can rapidly extract information from numerous publications and identify connections between significant elements. It provides a robust basis for recognizing hazards and

JDEBM
Journal of Digital Engineering and Business Managemet
Volume-2, Issue-1,January-March 2026
ISSN: 3107-7897 | P.ID: JDEBM/11
DOI: https://doi.org/10.5281/zenodo.19083510

understanding contemporary occurrences. The proposed initiative will yield further instances for our dataset and enhancements to our entity and relationship classification system. A knowledge graph will also be utilized in cognitive processes. This will furnish us with supplementary information.

REFERENCES

[1] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, S. Garg, and M. M. Hassan, "A distributed intrusion detection system to detect DDoS attacks in blockchain-enabled IoT network," Journal of Parallel and Distributed Computing, vol. 164, pp. 55–68, 2022.

[2] P. Kumar, G. P. Gupta, and R. Tripathi, "Design of anomalybased intrusion detection system using fog computing for IoT network," Automatic Control and Computer Sciences, vol. 55, no. 2, pp. 137–147, 2021.

[3] P. Kumar, R. Tripathi, and P. G. Gupta, "P2IDF: a privacypreserving based intrusion detection framework for software de?ned Internet of ,ings-fog (SDIoT-Fog)," in Proceedingsof the 2021 International Conference on Distributed Computingand Networking, pp. 37–42, Nara Japan, January 2021.

[4] Y. Zhou, Y. Tang, M. Yi, C. Xi, and H. Lu, "CTI view: APT threat intelligence analysis system," Security and CommunicationNetworks, vol. 202215 pages, Article ID 9875199, 2022.

[5] H. Jo, Y. Lee, and S. Shin, "Vulcan: automatic extraction and analysis of cyber threat intelligence from unstructured text," Computers & Security, vol. 120, Article ID 102763, 2022.

[6] X. Liu and J. Li, "Key-based method for extracting entities from XML data," Journal of Computer Research and Development, vol. 51, no. 1, pp. 64–75, 2014.

[7] A. Joshi, R. Lal, T. Finin, and A. Joshi, "Extracting cybersecurity related linked data from text," in Proceedings of the 2013IEEE Seventh International Conference on Semantic Computing, pp. 252–259, IEEE, Irvine, CA, USA, September 2013.

[8] V. Mulwad, W. Li, A. Joshi, T. Finin, and K. Viswanathan, "Extracting information about security vulnerabilities from web text," in Proceedings of the 2011 IEEE/WIC/ACM InternationalConferences on Web Intelligence and IntelligentAgent Technology, pp. 257–260, IEEE, Lyon, France, August 2011.

[9] R. A. Bridges, K. M. Hu4er, C. L. Jones, M. D. Iannacone, and J. R. Goodall, "Cybersecurity automated information extraction techniques: drawbacks of current methods, and enhanced extractors," in Proceedings of the 2017 16th IEEEInternational Conference on Machine Learning and Applications (ICMLA), pp. 437–442, IEEE, Cancun, Mexico, December 2017.

[10] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," Transactions of the associationfor computational linguistics, vol. 4, pp. 357–370, 2016.