



OPEN

DATA DESCRIPTOR

A dataset for cyber threat intelligence modeling of connected autonomous vehicles

Yinghui Wang¹, Yilong Ren^{1,2,3}✉, Hongmao Qin^{4,5}, Zhiyong Cui¹, Yanan Zhao¹
& Haiyang Yu^{1,2,3}✉

Cyber attacks pose significant threats to connected autonomous vehicles in intelligent transportation systems. Cyber threat intelligence (CTI), which involves collecting and analyzing cyber threat information, offers a promising approach to addressing emerging vehicle cyber threats and enabling proactive security defenses. Obtaining valuable information from enormous cybersecurity data using knowledge extraction technologies to achieve CTI modeling is an effective means to ensure automotive cybersecurity. However, the lack of a specialized cybersecurity dataset for automotive CTI knowledge mining has hindered progress in this field. To address this gap, we present a novel corpus specifically designed for vehicle cybersecurity knowledge mining. This dataset, annotated using a joint labeling strategy, comprises 908 real automotive cybersecurity reports, 8195 security entities and 4852 semantic relations. In addition, we conduct a comprehensive analysis of CTI knowledge mining algorithms based on this corpus. Our work provides a valuable resource for enhancing CTI modeling and advancing automotive cybersecurity research.

Background & Summary

The emergence of connected autonomous vehicles (CAVs) is considered a significant technological breakthrough in the global transportation industry. It is expected that CAVs will greatly improve transportation safety by enhancing efficiency, reducing congestion and minimizing accidents, etc^{1–3}. Despite the progress made in advanced automation, enhanced connectivity, and the implementation of shared services, these trends have also brought about new cyber-risks and vulnerabilities. Hackers now have more opportunities to exploit these potential attack surfaces and even gain vehicle control^{4,5}. What's worse, various emerging features such as the remote control function, over-the-air (OTA) update, internet connectivity, external devices and charging infrastructure have transformed potential threats into an undeniable reality. Over the past decade, the frequency, magnitude and sophistication of cyber-attacks on CAVs have experienced an exponential increase⁶. These cyber-attacks may cause privacy disclosure, financial losses, human damage or even compromise national public safety⁷. The homologation of new vehicle types now mandates the monitoring and response to cyber-attacks on vehicles and their ecosystem, as per the recent UNECE legal regulation⁸. There are numerous publications that discuss various security measures in CAVs, including access control, firewall, intrusion detection and prevention system (IDPS), security operations center (SOC), as well as cyber resilient architectures, etc^{9–13}. Nonetheless, these methods have several limitations, such as passive protection, restricted capability of threat identification, and so forth. Fortunately, cyber threat intelligence (CTI) becomes an ideal way to realize proactive defense and timely response to unknown or emerging threats in the automotive field. However, the mining and analysis of CTI data requires a great deal of manual inspection from open-source unstructured texts, which is an exceedingly time-consuming task¹⁴. Automatically extracting CTI knowledge from massive amounts of unstructured data has become a pressing and critical topic in the automotive cybersecurity domain.

Research on CTI knowledge extraction or modeling is still in the early stages, typically relying on pipeline methods, which first perform named entity recognition and then extract relationships between entities^{15–17}.

¹Beihang University, School of Transportation Science and Engineering, Beijing, 102206, China. ²Beihang University, State Key Laboratory of Intelligent Transportation System, Beijing, 102206, China. ³Zhongguancun Laboratory, Beijing, 100094, China. ⁴Hunan University, College of Mechanical and Vehicle Engineering, Changsha, 410082, China. ⁵Wuxi Intelligent Control Research Institute of Hunan University, Wuxi, 214115, China. ✉e-mail: yilongren@buaa.edu.cn; hyyu@buaa.edu.cn

Nevertheless, these pipeline methods may result in error propagation in entity recognition, and disregard the correlation between entity recognition and relation extraction. Increasingly, the end-to-end joint extraction model is being recognized as a popular approach for CTI mining tasks. Li *et al.* developed an end-to-end entity-relation extraction method, incorporating Luo's joint labeling scheme^{18,19}. The researchers merged the BiLSTM-LSTM model with a dynamic attention mechanism, i.e., BiLSTM-dynamic-att-LSTM model, achieving the simultaneous extraction of both entities and their relations. Subsequently, Zuo *et al.* constructed an extraction framework based on BERT-BiLSTM-att-CRF²⁰. Similarly, this work has introduced a sequence label approach and combined extraction principles, enhancing the capability to recognize overlapping relationships. Moreover, the BERT model takes advantage of deep bidirectional transformers to form word embeddings by fusing contextual details. This capability empowers it to effectively capture intricate semantic features during the extraction process²¹. Guo *et al.* presented the CyberRel, a joint model for extracting entities and relations in cybersecurity concepts using the BERT-BiGRU-att-BiGRU-CRF scheme²². They modeled the information extraction problem as a multi-sequence tagging process, resulting in distinct label sequences for various relations. Nevertheless, the complexity of this joint model correspondingly increased due to the multiple-sequence labeling strategy. Data is an essential foundation for research in CTI modeling. Mittal *et al.* proposed a "CyberTweets" dataset, which was built through tweets collected from Twitter²³. The CyberTweets only contained 7 entity categories: vulnerability, ransomware, DDoS, data leak, general, 0 Day and botnet. Similarly, the "Cyberthreat" corpus presented by Dion sio *et al.* was also based on Twitter's data, which also only included 5 entity types due to limited training data²⁴. Zhao *et al.* collected CTI data from security blogs, hacker forums, etc., and labeled these texts by "B-I-O" method¹⁴. The dataset primarily focused on the regular indicator of compromise, overlooking other complex related entities and relationship types. In addition, Satyapanich *et al.* annotated a cybersecurity event dataset, which covered cyber attacks and vulnerability information²⁵. They defined 5 event types, including 14 semantic roles and 20 argument types. Despite the existence of several open-source datasets for CTI modeling, none are specifically focused on CTI entity and relation mining in the automotive domain. What's more, the CAVs are filled with numerous cybersecurity entities. These range from the electronic control unit (ECU), data, in-vehicle network, to its various functions, as well as the potential threats and vulnerabilities it faces. The composition of these cybersecurity entities is irregular, and they face challenges to semantic heterogeneity. In addition, there are many overlapping relational entities in the automotive CTI texts, i.e., one entity may be involved in several semantic relationships simultaneously. These overlapping relations might cause ambiguities, thus making it difficult to recognize such entities and relationships during CTI data mining. Our objective is to provide data support for CTI modeling by developing an automotive CTI ontology and simultaneously annotating security entities and their relationships. This research may provide a solution for effectively extracting security-safety knowledge hidden within vast amounts of cyber threat information, thereby enabling timely detection of potential cyber threats for CAVs.

To address the lack of data on automotive CTI knowledge extraction and promote research in CTI modeling, we built an automotive CTI dataset, Acti, focusing on mining CTI entities and their associations. This dataset can be used for CTI modeling, enabling the timely identification and analysis of potential cyber threats to vehicles. This dataset contains data from 908 real cybersecurity incident texts collected across three cyber threat information sources. The dataset includes 10 entity concepts related to cyber and physical world, along with 10 semantic relationship categories. These entities and relationships are derived from the definition of the automotive CTI ontology. The CTI ontology is a way used to formally express the concepts and their semantic relationships in the automotive cyber threat intelligence domain. We have adjusted our data to be the "BIOES" - "entity type" - "relation type" - "entity role" joint annotation schema, and made the data available at Figshare database under the <https://doi.org/10.6084/m9.figshare.27916758.v1>. This dataset includes 3678 sentences, covering 8195 security entity instances and 4852 entity-relation triples. We train the CTI mining models using entity-relation joint extraction techniques to validate the reliability of the CTI dataset. Besides, the data was categorized into cyber and physical elements in accordance with the CTI ontology, further depicting and analyzing the interrelation between security and safety. This CTI dataset is expected to facilitate the collaborative analysis of functional safety and cybersecurity, enabling supporting further vehicle cybersecurity research work. In this study, we aim to widely introduce this CTI dataset and make it accessible for public use, enabling more people to conduct meaningful investigations.

Methods

In this section, we outline the methodology for creating the Acti corpus, comprising two main components: (1) data collection, focusing on identifying and documenting cybersecurity attack incidents related to automobiles; and (2) data processing, including automotive CTI ontology modeling and a joint annotation scheme to transform unstructured cybersecurity incidents data into "BIOES" - "entity type" - "relation type" - "entity role" annotation format.

Data collection. We collect vehicle cybersecurity data primarily through two channels: (1) retrieving published vehicle-related cybersecurity vulnerability information from the national vulnerability database (NVD), and (2) gathering cybersecurity data from specialized vehicle threat intelligence platforms, cybersecurity conferences, reports, literature, and other sources. The NVD (<https://nvd.nist.gov/>) contains some cybersecurity vulnerability data specific to CAVs, stored alongside over 200,000 vulnerability entries from various industries. To identify relevant vehicle vulnerabilities from this extensive dataset, we apply a keyword retrieval approach. Based on our previous research⁷, these keywords are classified into three categories, as shown in Table 1.

In addition, we collect vehicle cybersecurity data from the auto-threat intelligence cyber incident repository of Upstream Security and automotive attack database (AAD)²⁶. The Upstream Security's repository (<https://upstream.auto/research/automotive-cybersecurity>) includes over 1,300 publicly reported cyber

Keyword type	Keyword
Vehicle term	Vehicle, Car, Automotive
Vehicle components and networks	Bluetooth, Braking system, Engine control, Infotainment, Keyless entry, CAN, LIN, MOST, OBD-II, T-BOX, Gateway, TPMS, etc.
Original equipment manufacturer (OEM)	Bmw, Audi, Toyota, Jeep, Mercedes-benz, Ford, Mazda, Subaru, Great wall, Lexus, Chrysler, Tesla, etc.

Table 1. Vehicle cybersecurity data retrieval keywords.

Data source	Count of CTI	Count of sentences
Automotive CVE vulnerability (NVD)	198	380
Upstream's cybersecurity incident	360	1219
Automotive attack database	350	2079
Total	908	3678

Table 2. The Overview of Acti corpus.

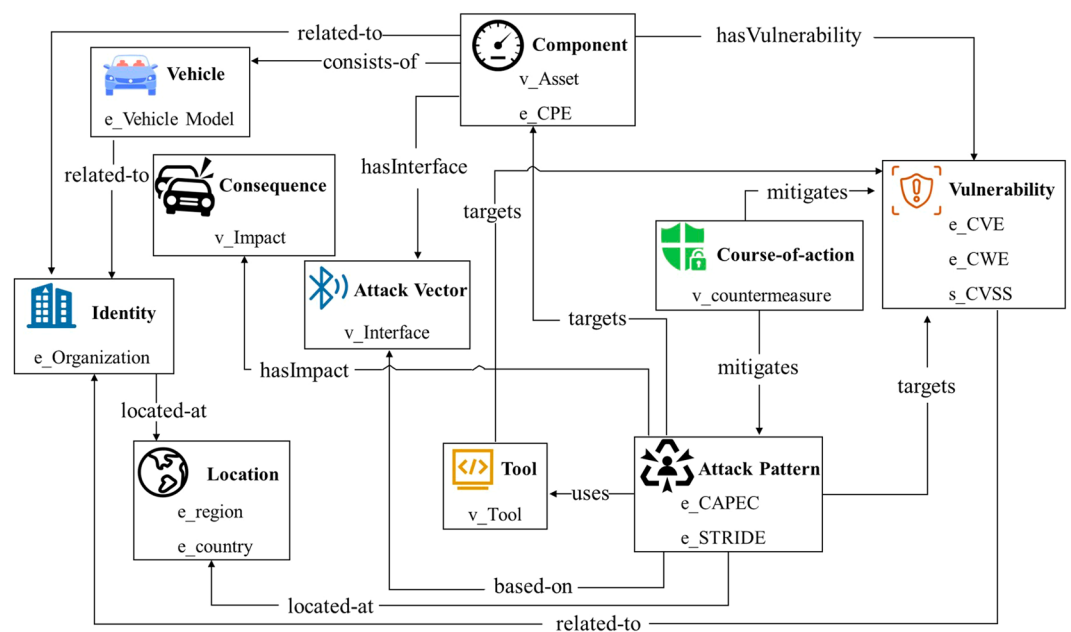


Fig. 1 Automotive CTI Ontology Model.

incidents targeting the smart mobility ecosystem. In this study, we focus exclusively on threats and vulnerabilities that directly or indirectly impact vehicles. Some cybersecurity incidents such as ransomware attacks, information leakages or other security breaches affecting backend platforms of manufacturers or car rental services are excluded. Meanwhile, we only gather the description information concerning vehicle vulnerabilities and auto-threat intelligence cyber incidents, without considering content from other fields. Besides, the AAD (<https://github.com/IEEM-HsKA/AAD>) contains 23 fields mapped to an attack taxonomy, including year, description, attack class, attack type and vulnerability, etc. From this dataset, we select three columns of unstructured text: description, attack path and consequence information. In total, we collect 908 real automotive cybersecurity data to create the Acti corpus. An overview of the essential data contained in the corpus is presented in Table 2.

Data processing. The data processing aims to label the entity categories and relation types from the unstructured cybersecurity data, and convert them into the “BIOES” - “entity type” - “relation type” - “entity role” joint annotation format. The first step involves defining a automotive CTI ontology to describe the entity categories and their interrelations. Then, the manual joint annotation process is carried out using the Brat tool. The Brat is an open-source, web-based text annotation tool for adding entity and relation notes to existing text documents. This tool can be downloaded from <https://github.com/nlplab/brat/archive/refs/tags/v1.3p1.tar.gz>.

Automotive CTI ontology modeling. Most CTI is unstructured or encoded in diverse formats, characterized by dispersed data sources, numerous concepts and complex semantic relationships. The ontology concept provides a feasible solution to the challenges posed by heterogeneous data and intricate characteristics in the CTI field²⁷. In constructing a CTI corpus, it is crucial to establish a well-defined ontology model tailored to automotive CTI. To provide a comprehensive depiction of CTI in the automotive domain, we define 10 entity

Relation type	Property	Triple
hasVulnerability	a specific component has a vulnerability instance	(Component, hasVulnerability, Vulnerability)
hasInterface	a component instance has an attack vector instance	(Component, hasInterface, Attack vector)
hasImpact	the attack pattern or vulnerability can cause some impact directly or indirectly	(Attack pattern, hasImpact, Consequence) (Vulnerability, hasImpact, Consequence)
targets	the attack pattern/tool targets the component, or vulnerability	(Attack pattern, targets, Component) (Attack pattern, targets, Vulnerability) (Tool, targets, Vulnerability)
uses	the tool is employed to carry out the behavior identified in the attack pattern	(Attack pattern, uses, Tool)
mitigates	the course of action can mitigate the related attack pattern or vulnerability	(Course of action, mitigates, Attack pattern) (Course of action, mitigates, Vulnerability)
related-to	two entity classes have a related relationship	(Component, related-to, Identity) (Vehicle, related-to, Identity) (Vulnerability, related-to, Identity)
located-at	the identity is located at the related location, or an attack occurred in a region	(Identity, located-at, Location) (Attack pattern, located-at, Location)
based-on	the attack pattern instance was carried out through the attack vector	(Attack pattern, based-on, Attack vector)
consists-of	two entity classes have a containment relationship	(Component, consists-of, Vehicle) (Component, consists-of, Component)

Table 3. ACTI entity-relation triples.

and 10 relationship types. These elements integrate industry-leading outcomes, such as unified cybersecurity ontology (UCO)²⁸, structured threat information eXpression (STIX), the classification model for automotive cyber-attacks²⁶ and the CTI data model²⁷. The automotive CTI ontology model, shown in Fig. 1, defines the following 10 entity types:

- **Component:** The component (or asset) class refers to any physical or logical element of a vehicle system or network, including hardware, software, firmware, data and interfaces, etc.
- **Consequence:** This class describes the potential impact of an attack, such as vehicle theft, control vehicle systems, data breach, service/business disruption, location tracking, fraud, and so on.
- **Identity:** The identity class represents individuals, organizations or groups, e.g., Acura, BMW, Chrysler, Ford, etc.
- **Vehicle:** The class refers to the vehicle model, which is a distinctive identifier given by the vehicle manufacturer to a particular type of vehicle. e.g., the Tesla Model S.
- **Location:** The region where the attack occurred or the target organization/company belongs to.
- **Attack vector:** The different points where an attacker might enter or retrieve data in a system, often referred to as the attack surface. This class consists of sub-classes like cellular, Bluetooth, CAN-bus, and so on.
- **Attack pattern:** This class describes various ways attackers employ to compromise targets and consists of sub-classes like brute force, replay, eavesdropping, buffer overflow, reverse engineering, etc.
- **Tool:** This class represents the equipment that could be employed by attackers to execute an attack. The tools include vehicle diagnostic tools, debuggers, sniffing tools, and so forth.
- **Vulnerability:** A Vulnerability is an internal mistake that enables an external threat to compromise the system causing security consequences. For example, missing input validation fault, buffer overflow vulnerability, etc.
- **Course of action:** The course of action is an action that is taken to prevent or respond to an attack, such as access control, encryption, patch, firewall, and so forth.

Subsequently, based on the specific attributes of automotive CTI data and insights from prior research, the ontology model incorporates 10 relation types to depict the connections among the predefined threat entities. Specifically, the relation types are: “hasVulnerability”, “hasInterface”, “hasImpact”, “targets”, “uses”, “mitigates”, “related-to”, “located-at”, “based-on” and “consists-of”. The semantic relations between entity classes are reflected through object properties, as depicted in Table 3. It provides a clearer understanding of how the predefined entities are interconnected. Furthermore, considering the intricate and professional nature of the automotive cybersecurity domain, we developed a comprehensive lexicon encompassing domain-specific vocabulary and public enumeration. This lexicon resource serves as a valuable tool for accurately annotating automotive CTI data. The following is an overview of the lexicon:

- **Domain-specific Vocabulary:** The vocabulary represents the list of candidate content for pre-defined entities that is supplied with automotive CTI ontology, namely internal enumeration. The vocabulary list mainly includes “asset”, “impact”, “interface”, “countermeasure” and “tool” categories. For example, the asset vocabulary contains more than 1,000 terms, such as central gateway, brake system ECU, advanced driver assistance system (ADAS), etc. The tool vocabulary list mainly consists of hardware, software, security, sensing, measurement, and wireless tool, with previous reference to the research of Sommer *et al.*²⁶.
- **Public Enumeration:** The public enumeration refers to publicly available expressions or databases that contain valuable data related to CTI. This includes critical details such as the configuration, weakness, vulnerability, vehicle type, location and among other aspects. The exemplary enumerations involve CVE number, common weakness enumeration (CWE), common attack pattern enumeration and classification (CAPEC), and so forth.

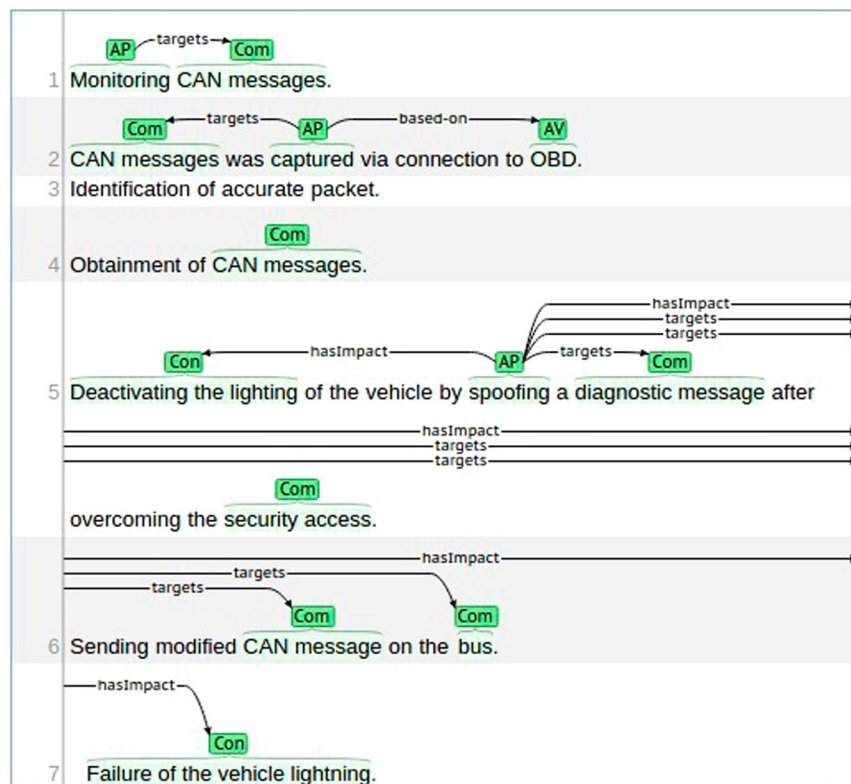


Fig. 2 Brat data annotation.

data: Monitoring CAN messages.
label: S AP targets 1 B_Com_targets_2 E_Com_targets_2 O
data: CAN messages was captured via connection to OBD.
label: B_Com_targets_2 E_Com_targets_2 O S AP_M_1 O O O S_AV_based-on_2 O

Fig. 3 Automotive CTI annotation data.

Automotive CTI corpus joint annotation. We integrate entity and relation labels into a joint annotation scheme¹⁸, enabling comprehensive exploitation of the information they represent. This scheme consists of four components: entity boundary, entity type, relation type and entity role. Tokens in the annotation process are classified into three categories: (1) tokens unrelated to entities or relations, (2) tokens explicitly associated with entities, and (3) tokens that encompass both entities and relations. The detailed scheme is outlined as follows:

- **Entity boundary:** The “BIOES” (Begin, Inside, Other, End, Single) scheme is adopted to indicate the token’s position within the entity. The “B”, “I”, and “E” tags denote the position of the word, i.e., the beginning, middle and tail of the entity, respectively. The tag “S” represents that the word is a single entity. Additionally, the tag “O” means that the word isn’t associated with any pre-defined entities.
- **Entity type:** The entity type information is pre-defined according to the automotive CTI ontology model. We classify entities into ten categories: component (Com), consequence (Con), identity (Ide), vehicle (Veh), location (Loc), attack vector (AV), attack pattern (AP), tool (Tool), vulnerability (Vul) and course of action (CoA).
- **Relation type:** The relation type is also acquired from the pre-defined set: {“hasVulnerability”, “hasInterface”, “hasImpact”, “targets”, “uses”, “mitigates”, “related-to”, “located-at”, “based-on”, “consists-of”}. Furthermore, the tag “M” is added to denote the entity exits multiple relations, i.e., overlapping relations.
- **Entity role:** The label indicates the role of the entity in the relation and is defined using the tags “1”, “2” and “m”. Concretely, the labels “1” and “2” represent the word of the first entity and second entity in the relation, respectively. The “m” represents the word of distinct role entities in the overlapping relation.

Data name	Description	Data format
Raw data	Raw unstructured cybersecurity data that has been cleaned, filtered and preprocessed.	txt
Brat annotation data	The data (following the BIO schema) was automatically generated based on manual annotation of entity and relationship types using the Brat tool.	ann
BIOES	The corpus is a collection of annotated data in the “BIOES” - “entity type” - “relation type” - “entity role” joint annotation format.	txt

Table 4. Dataset description.

Corpus	Entity	Relation	Scale
CASIE ²⁵	20	14	10384
Cybertweets ²³	8	—	21000
Cyberthreat ²⁴	5	—	11073
HINTI ¹⁴	6	—	30000
Acti ²⁹	10	10	3678

Table 5. Cybersecurity knowledge extraction open source corpora.

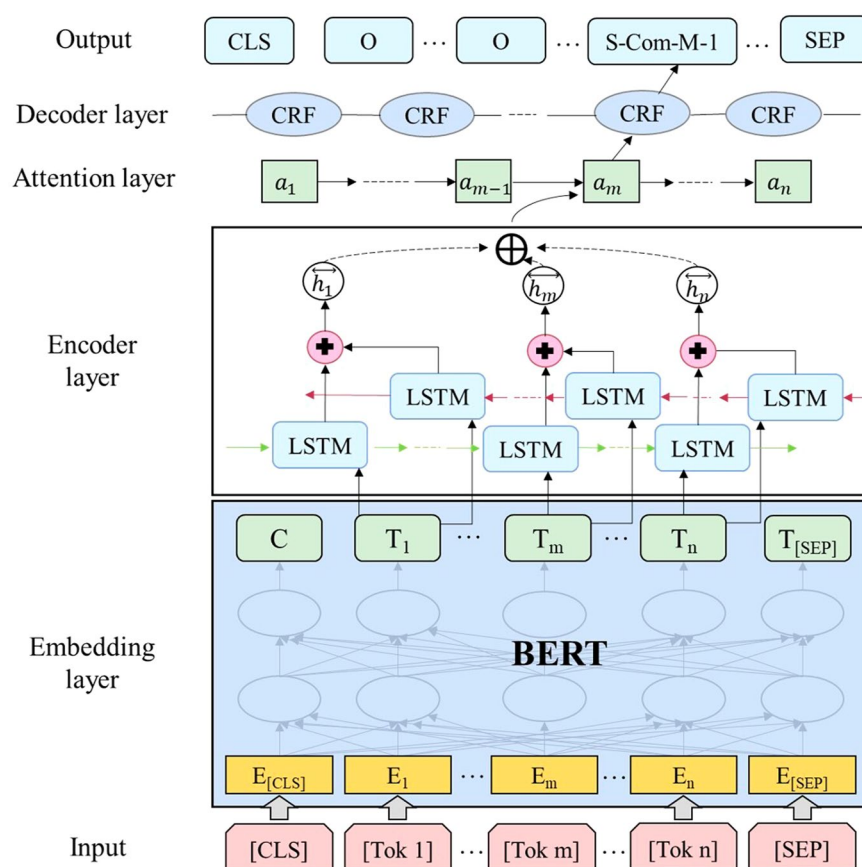


Fig. 4 BERT-BiLSTM-att-CRF model.

After identifying the entity and relation types to be annotated, the data annotation process begins. This process is known as a sequence labeling task, where each element in the open-source unstructured data is annotated with a label by treating each sentence as a sequence, with each word serving as an individual element. We leverage the above annotation strategy and the Brat tool to label entity and relation types in automotive cybersecurity data. Before data annotation, the Brat tool requires a predefined annotation format, which involves manually configuring entity and relationship types in its configuration file. Subsequently, the Brat tool is used to label various types of entities and establish connections between entities to annotate relation types. The entity and relation annotation is performed manually, as shown in Fig. 2.

Once the manual annotation process is completed, the Brat tool automatically generate “.ann” and “.conll” files. However, the “.conll” file is limited to the standard BIO format. To address this, a Python script is developed to convert these “.ann” files into the joint annotation format described above. For the sentence “Monitoring CAN

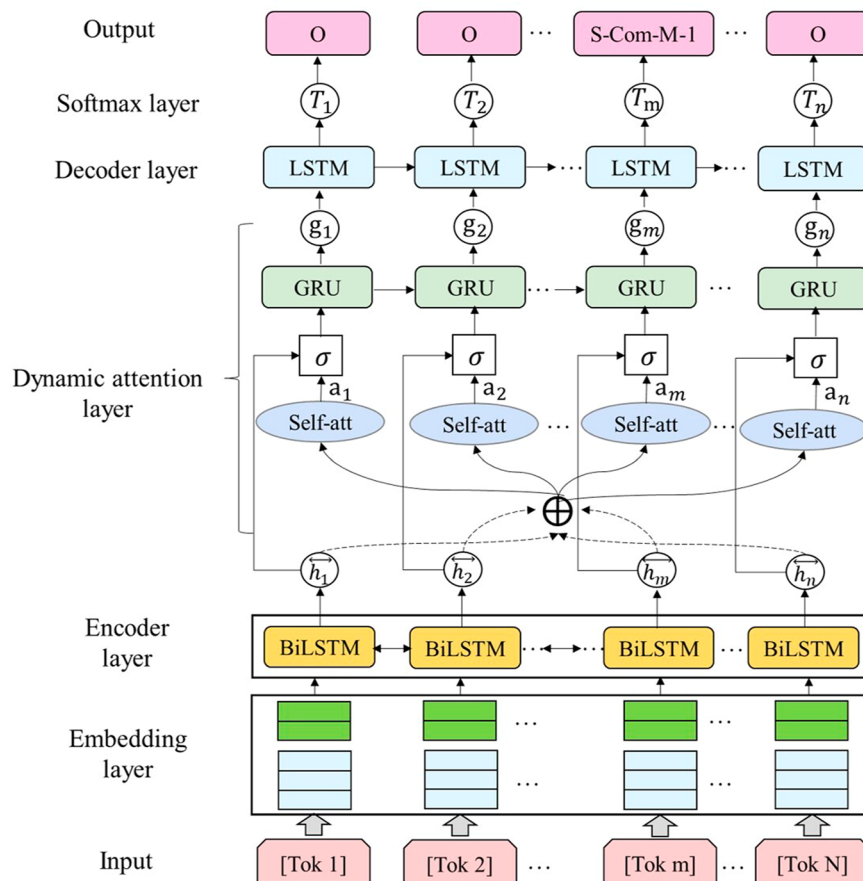


Fig. 5 BiLSTM-dynamic-att-LSTM model.

message”, the BIO format is represented as “S-AP B-Com E-Com O”. The corresponding joint annotation label can be represented as “S-AP-targets-1 B-Com-targets-2 E-Com-targets-2 O”, as shown in Fig. 3.

Data Records

The datasets, automotive cyber threat intelligence corpus (Acti)²⁹, has been uploaded to the Figshare and is available at <https://doi.org/10.6084/m9.figshare.27916758.v1>. The Acti dataset consists of two main components: (1) data (which includes raw data and brat annotation data), and (2) BIOES (the corpus in a joint annotation format). The details are shown in Table 4. The raw unstructured cybersecurity data is collected from NVD, auto-threat intelligence cyber incident repository of Upstream Security and automotive attack database. It includes 908 files in “.txt” format, with names such as “CVE-2022-23126”, “2022-11-1” and “ID2019NxumaloSSA1”, respectively corresponding to the aforementioned data sources. After manual annotation using the Brat tool and automatic format conversion, “.ann” files in BIO format and “.txt” files in the joint annotation format were generated sequentially, i.e. brat annotation data and the BIOES. We randomly divided the entire Acti dataset into training and test sets in a ratio of 8:2, which can be found in the “model” folder. In addition, the source code to convert these “.ann” files into the joint annotation format and preprocess the annotation data could be found in the <https://doi.org/10.6084/m9.figshare.27916758.v1>. Meanwhile, the source code for training several CTI knowledge mining deep learning models is also available and will be provided in the Code Availability section.

Technical Validation

To evaluate the performance of this Acti corpus, we compared various open source corpora, and conducted a comprehensive analysis of deep learning algorithms for CTI knowledge extraction. Specifically, we evaluated the performance of the end-to-end CTI joint extraction models on the Acti dataset. These methods avoid the error propagation inherent in the classical pipeline model and have increasingly become a key direction in CIT modeling research.

Comparison of corpora. We carefully investigated various open-source corpora for CTI knowledge mining tasks in the cybersecurity domain, as shown in Table 5. The “Cybertweets”, “Cyberthreat” and “HINTI” corpora only provide the annotation for entity types, regardless of the semantic relations among security entities. Furthermore, the entity types labeled in these datasets are so restricted that they inadequately capture the comprehensive landscape of CTI information. The “CASIE” dataset contains annotations for both entities and relations. However, it simply annotated the entity type while ignoring the labeling of entity boundaries. And

Model	Parameter	Value
BERT-BiLSTM-att-CRF	Transformer layers	12
	hidden size	768
	activation function	ReLU
	max position embedding	128
	BiLSTM_dim	800
	epoch	125
	batch	16
	learning rate	5e-5
	dropout	0.1
BiLSTM-dynamic-att-LSTM	word embedding_dim	300
	char embedding_dim	15
	CNN filter	15
	CNN kernel size	5
	BiLSTM_dim	300
	LSTM_dim	600
	learning rate	0.001
	bias weight	10

Table 6. Hyperparameter setting.

Model	Metric		
	P%	R%	F1%
BERT-BiLSTM-att-CRF	47.39	47.66	47.52
BiLSTM-dynamic-att-LSTM	45.59	39.55	42.36
BiLSTM-att-LSTM	44.91	40.3	42.33
BiLSTM-att-CRF	45.38	41.9	43.57
BERT-BiGRU-att-BiGRU-CRF	45.74	45.12	45.43

Table 7. Comparison of experimental results.

the problem of overlapping relational entities was also not considered. What's more, none of the above datasets cover information related to automobile cybersecurity. In this paper, we employ the sequence tagging strategy to annotate entities and relations within the automotive CTI data, effectively alleviating the problem of overlapping relational entities. The Acti corpus is a valuable source of automotive CTI information mining tasks, offering a comprehensive view of CTI entities and relations. Besides fundamental security-related elements, the Acti also incorporates the marking of vehicle components, physical impact and other entities. These elements significantly facilitate the study of the interrelation between automobile cybersecurity and functional safety.

CTI knowledge mining models. To verify the reliability of the Acti corpus, several models, including “BERT-BiLSTM-att-CRF”²⁰, “BiLSTM-dynamic-att-LSTM”¹⁹, “BERT-BiGRU-att-BiGRU-CRF”²², “BiLSTM-att-LSTM” and “BiLSTM-att-CRF”, are applied in automotive CTI knowledge mining tasks. Specifically, the architecture of the BERT-BiLSTM-att-CRF model is illustrated in Fig. 4. This model consists of four layers: embedding, BiLSTM, attention mechanism and conditional random fields (CRF)²⁰. The embedding layer transforms words into vector representations, and then the BiLSTM layer calculates the probability distribution of the word vector. Subsequently, the attention layer reflects the relationships between words, and the CRF layer predicts the globally optimal labeling sequence.

- **Embedding layer.** The BERT model encodes the inputs on multiple transformer layers, incorporating multi-head attention and feedforward neural networks to effectively extract deep semantic information.
- **Encoder layer.** The BiLSTM, or bidirectional long short-term memory (LSTM), is a highly potent network that utilizes both forward and backward LSTM to effectively capture semantic information in both directions of each sequence. It could address the issue of gradient disappearance and enhance the generation of comprehensive semantic features.
- **Attention mechanism layer.** The self-attention mechanism is introduced to focus on the key information from the cybersecurity dataset. It initializes an attention matrix to represent the relative significance of words to their vectorization representations, and effectively reflects the influence between words.
- **Decoder layer.** Labels exhibit strong interconnections rather than being independent. The CRF model is commonly used in sequence labeling tasks. It effectively calculates the transfer probability between labels in a tagging sequence, allowing it to select a global optimal sequence label.

Entity	P%	R%	F1%	Relation	P%	R%	F1%
Component	65.9	46.03	54.2	hasVulnerability	40.08	18.97	44
Consequence	63.98	42.77	51.27	hasInterface	54.39	12.3	25.75
Identity	58.84	21.56	31.56	hasImpact	60.07	51.63	55.53
Vehicle	66.96	44.42	53.41	targets	54.28	43.30	48.17
Location	39.12	14.86	21.54	uses	46.64	27.92	34.93
Attack vector	33.75	12.5	18.24	mitigates	30.88	7.78	12.43
Attack pattern	80.21	45.26	57.86	related-to	44.74	17.25	24.9
Tool	62.68	30.01	40.59	located-at	31.94	8.52	13.45
Vulnerability	72.42	33.48	45.79	based-on	19.2	6.35	9.54
Course of action	6.67	0.93	1.63	consists-of	55.16	38.65	45.45

Table 8. Experimental results on entities and relations.

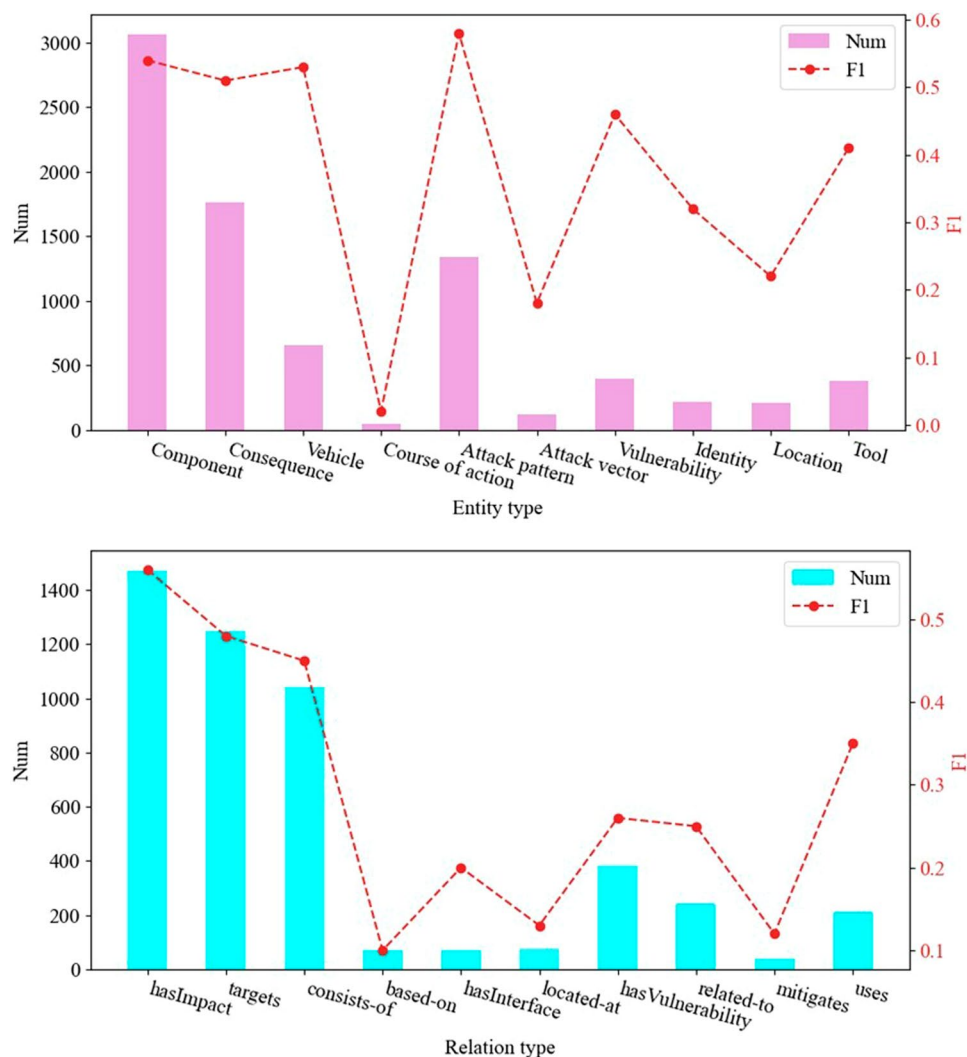


Fig. 6 Distribution and F1 of entity and relation extraction.

Similarly, in the “BERT-BiGRU-att-BiGRU-CRF” model, the bidirectional gated recurrent unit (BiGRU) encoder captures contextual information from both forward and backward directions within a sequence. Compared to BiLSTM, BiGRU employs a simpler architecture with fewer parameters, making it computationally more efficient in certain scenarios. In addition, the structure of the BiLSTM-dynamic-att-LSTM model is shown in Fig. 5. This model includes the embedding layer, encoder layer, dynamic attention layer, decoder layer and softmax layer.

- **Embedding layer.** The word2vec is used to obtain word-level features of the sentence sequence. Then, the convolutional neural network (CNN) module is adopted to extract the character features of the input sentence sequence. Finally, these two features are spliced together as the input of the encoding layer.
- **Encoder layer.** The encoder layer adopts the same method as the BERT-BiLSTM-att-CRF model, i.e., BiLSTM, to obtain the feature encoding of input sequences.
- **Dynamic Attention Mechanism layer.** The dynamic attention mechanism takes into account the semantics of words and considers their variations in different contexts¹⁹. The feature encoding h_t generated by the BiLSTM layer is spliced with the output a_t of the self-attention mechanism. The splicing result $[h_t, a_t]$ is filtered by the sigmoid function to obtain γ_t , then a dot-product operation is performed to calculate ξ_t . The ξ_t is the input for the gated recurrent unit (GRU). The detailed process is as follows:

$$\gamma_t = \text{sigmoid}(W_s[h_t, a_t]) \quad (1)$$

$$\xi_t = \gamma_t[h_t, a_t] \quad (2)$$

$$g_t = \text{GRU}(g_{t-1}, \xi_t, \theta) \quad (3)$$

where W_s and θ are hyper-parameter matrices. Let $G = \{g_1, g_2, \dots, g_t\}$ represent the final output of the dynamic attention layer, and subsequently input to the lower layer for decoding.

- **Decoder layer.** The decoding layer employs a LSTM network to generate the vector representation of label sequences. The LSTM tag decoder can significantly speed up model training and promises to achieve comparable performance to the CRF decoder. Finally, the label sequence is normalized using a softmax layer.

Pre-processing and experimental setting. This specialized cybersecurity dataset for automotive cyber threat intelligence modeling comprises 3678 sentences, 8195 labeled entities and 4852 relationships. we randomly divided the entire Acti dataset into training and test sets in a ratio of 8:2. The experiments are carried out on a high-performance server running the Windows 10 operating system. And the server is equipped with an Intel Core i5-13400F CPU@2.50 GHz, 32GB RAM and the powerful NVIDIA GeForce RTX 3090 GPU. The software environment comprised Python 3.7, CUDA 11.2, PaddlePaddle-GPU 2.3.2 and paddlenlp 2.1.1, etc. Some main hyperparameters of the “BERT-BiLSTM-att-CRF” and “BiLSTM-dynamic-att-LSTM” models are shown in Table 6.

Experimental results. We performed a comprehensive comparison of the Acti dataset for the CTI knowledge mining task. We employ the general evaluation metrics for information extraction tasks, namely precision (P), recall (R) and F1-score, to evaluate the performance of CTI knowledge mining models.

$$P = \frac{TP}{TP + FP} \quad (4)$$

$$R = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = \frac{2 * P * R}{P + R} \quad (6)$$

Where TP represents the number of positive instances correctly recognized; FP is the count of positive samples incorrectly identified; and FN stands for the number of negative examples incorrectly classified. The overall performance of these CTI knowledge mining models is evaluated using the aforementioned evaluation indicators. The precision, recall and F1 score are shown in Table 7. We compare the experiment results of the above CTI knowledge extraction models on the Acti corpus. The aforementioned models achieve joint extraction of security entities and their relations, utilizing the semantic relationship between entity recognition and relation extraction tasks. Observably, the BERT-based model, especially “BERT-BiLSTM-att-CRF”, demonstrates better performance compared to other models with an F1 score of 47.52%, which also proves the exceptional capability of BERT in the CTI knowledge mining tasks. The reason behind this is that BERT word vector better captures grammatical and semantic information across various contexts, enhancing the model’s ability to generalize. Thus, it is capable of effectively handling the complex semantic characteristics of automotive CTI data. In addition, the experimental results of the “BiLSTM-att-LSTM” and “BiLSTM-att-CRF” models show that, at the decoder layer, the performance of the LSTM model is nearly equivalent to CRF.

Subsequently, we evaluate the performance metrics for each class of entity and relation extraction in the “BERT-att-BiLSTM-CRF” model. Specifically, as for entity recognition, a predicted entity is deemed accurate when its boundary and entity category are accurately labeled. Similarly, when the boundary, entity category and relation type are all correct, the relation extraction result is deemed exact. The experiment statistical results are as shown in Table 8. Meanwhile, we further conducted a detailed analysis of the data characteristics within the Acti corpus. The distribution of entity and relation instances is shown in Fig. 6. It can see that there is a clear

imbalance in the entity and relation instances of the corpus. This is because the description of original text data mainly focuses on security elements such as component, consequence and attack pattern, etc. The instances of these corresponding relationships also constitute a significant portion in the automotive CTI data.

Based on these, we further analyze the relationship between the experimental results of each entity and relation extraction and their corresponding instance distribution. As shown in Table 8, the F1 scores for entity recognition of “Attack vector”, “Location” and “Course of action” are relatively low, with “Course of action” entities scoring as low as 1.63%. This is primarily due to the limited number of instances for these entity types, which restricts the model’s recognition capability. Consistently, the F1 scores for the “based-on”, “located-at” and “mitigates” relations corresponding to these entities are also relatively low. Additionally, it can be observed that the F1 values for each entity and relation extraction performance generally align with their instance distributions, as shown in Fig. 6.

Furthermore, there are many cross-sentence relation entities in the Acti corpus, as shown in Fig. 2. Unfortunately, these existing CTI knowledge mining models only consider the entity-relation extraction at the sentence level, ignoring the problem of cross-sentence entities. These may have a certain influence on the performance of CTI knowledge extraction experiments. Despite this, these experiments have also sufficiently demonstrated the reliability of the Acti corpus in extracting entities and their relations of automobile CTI data. Researchers can select or improve these approaches according to the specific requirements and restrictions of their projects.

Discussion

The Acti corpus currently faces limitations in terms of data sources, relying solely on publicly available automotive cybersecurity reports. This approach may fail to comprehensively cover real-world attack types and threat scenarios. Additionally, significant disparities in the distribution of entity descriptions within automotive cybersecurity data have resulted in imbalanced entity and relationship representations in the corpus. To address these limitations, future work will expand the corpus coverage by incorporating diverse data sources, such as real-time updates from the NVD and open-source threat intelligence platforms, as well as data from dark web, security forums, anonymized internal threat intelligence and testing reports. This approach is expected to mitigate the imbalance in entity and relationship distributions. Currently, the data collection process is exclusively focused on CTI text data. However, threat intelligence may also exist in multimodal forms, such as images and videos. These multimodal data types require cleaning and preprocessing before effective integration into the corpus. Future research will focus on the integration and application of such multimodal data. In addition, due to the limited size of the currently corpus, this study did not incorporate a validation set during data splitting. As the dataset expands, a separate validation set could be incorporated into future data splits to evaluate the model’s generalization ability and mitigate overfitting.

In the data annotation process, the subjectivity and consistency issues are inevitable. Although this study constructed a lexicon containing domain-specific vocabulary and public enumerations, some entities may exhibit polysemy in different contexts. Annotators may interpret the same data differently, leading to inconsistencies in the annotation results. To mitigate this, future work could establish detailed annotation guidelines and standards to clearly define the annotation criteria, such as granularity, annotation methods, and the use of annotation symbols. Additionally, the cross-annotator validation mechanisms could be implemented to minimize subjective bias and enhance the quality and consistency of the corpus. Furthermore, annotators should participate in systematic training to enhance their annotation skills. Regarding overlapping and cross-sentence entity issues within the corpus, the following improvements are proposed: for overlapping entities, a multi-sequence labeling strategy can be adopted, generating separate label sequences for each relationship in a sentence. Cross-sentence entities, however, are inherently constrained by the original data description and may not be addressed through annotation strategies. For the automotive CTI knowledge mining task, exploring document-level entity-relation joint extraction models for cross-sentence relationships may offer a viable solution. As automotive CTI data grows, developing automated annotation models for entities and relationships could further reduce reliance on manual annotation.

In practical security analysis scenarios, CTI can be mapped to the vehicle hardware bill of materials and software bill of materials (i.e., HBOM and SBOM), then stored in the Neo4j graph database to support multidimensional correlation analysis. This approach facilitates tight integration between vehicle assets, software components, known vulnerabilities, and potential threats, providing a comprehensive visualization of a vehicle’s overall security posture. Moreover, automotive CTI can provide the indicators of compromises, such as malicious IP addresses and port numbers, to optimize security information and event management rule sets within the vehicle security operations center (VSOC), thereby enhancing threat detection capabilities. By analyzing threat behavior patterns in CTI data, such as advanced persistent threat attack techniques, the VSOC can identify anomalies and detect novel, unknown attacks. During security incident response, CTI offers valuable contextual information about threats, including targets and propagation methods, enabling the VSOC to rapidly formulate effective response strategies. Furthermore, through CTI sharing platforms, analyzed and processed threat intelligence can be shared with regulatory authorities, supply chain partners, development teams, and other stakeholders, fostering collaborative threat mitigation. In summary, the findings of this study contribute to advancing the detection, response and defense capabilities for automotive cybersecurity threats. These efforts provide significant support for achieving proactive defense and dynamic security management.

Usage Notes

Our proposed Acti corpus supplies a comprehensive collection of real cybersecurity data related to vehicles, encompassing a wide range of security entities, safety entities and semantic relationships. That is to say, the Acti dataset has the potential to serve as a valuable resource for CTI modeling and for enhancing security analysis in the automotive domain. The Acti dataset can be employed to train deep learning models which can

automatically extract CTI knowledge—specifically, security and safety entities and their relationships—from massive amounts of unstructured data, thereby enhancing the ability to timely identify potential threats, and aiding in the formulation of appropriate security measures for automobiles. By supplying researchers with access to an automotive CTI entity-relation joint annotation corpus, the Acti corpus facilitates the development and evaluation of more effective CTI modeling or knowledge mining algorithms. The Acti dataset contains both security-related and physical elements. It is expected to facilitate the collaborative analysis of functional safety and cybersecurity, enabling supporting further cybersecurity research work for CAVs.

Code availability

All source code for processing the Acti dataset and deep learning training have been uploaded to GitHub at <https://github.com/AutoCS-wyh/Automotive-cyber-threat-intelligence-corpus>. The code is openly accessible and can be used freely with appropriate attribution.

Received: 8 July 2024; Accepted: 8 January 2025;

Published online: 01 March 2025

References

- Gupta, B. B., Gaurav, A., Marín, E. C. & Alhalabi, W. Novel graph-based machine learning technique to secure smart vehicles in intelligent transportation systems. *IEEE transactions on intelligent transportation systems* **24**, 8483–8491 (2022).
- Bendiab, G., Hameurlaine, A., Germanos, G., Kolokotronis, N. & Shiaele, S. Autonomous vehicles security: Challenges and solutions using blockchain and artificial intelligence. *IEEE Transactions on Intelligent Transportation Systems* **24**, 3614–3637 (2023).
- Mansourian, P., Zhang, N., Jaekel, A. & Kneppers, M. Deep learning-based anomaly detection for connected autonomous vehicles using spatiotemporal information. *IEEE Transactions on Intelligent Transportation Systems* **24**, 16006–16017 (2023).
- Maple, C., Bradbury, M., Le, A. T. & Ghirardello, K. A connected and autonomous vehicle reference architecture for attack surface analysis. *Applied Sciences* **9**, 5101 (2019).
- Pandey, M. *et al.* A review of factors impacting cybersecurity in connected and autonomous vehicles (cavs). In *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*, vol. 1, 1218–1224 (IEEE, 2022).
- Upstream. Upstream's 2023 global automotive cybersecurity report. Tech. Rep., Upstream Security (2023).
- Wang, Y. *et al.* Automotive cybersecurity vulnerability assessment using the common vulnerability scoring system and bayesian network model. *IEEE Systems Journal* **17**, 2880–2891 (2022).
- Burkacky, O. *et al.* Cybersecurity in automotive. Tech. Rep., McKinsey (2020).
- Wu, W. *et al.* A survey of intrusion detection for in-vehicle networks. *IEEE Transactions on Intelligent Transportation Systems* **21**, 919–933 (2019).
- Luo, F. & Hou, S. Cyberattacks and countermeasures for intelligent and connected vehicles. *SAE International Journal of Passenger Cars-Electronic and Electrical Systems* **12**, 55–67 (2019).
- Ring, M., Frkat, D. & Schmiedecker, M. Cybersecurity evaluation of automotive e/e architectures. In *ACM Computer Science In Cars Symposium (CSCS 2018)*, vol. 92 (2018).
- Rosenstatter, T. & Tuma, K. A state-of-the-art investigation: Cyrev deliverable d1.1. cyber resilience for v ehicles—cyrev, cyrev consortium. Tech. Rep., Gothenburg University (2019).
- Grimm, D., Stang, M. & Sax, E. Context-aware security for vehicles and fleets: a survey. *IEEE Access* **9**, 101809–101846 (2021).
- Zhao, J., Yan, Q., Liu, X., Li, B. & Zuo, G. Cyber threat intelligence modeling based on heterogeneous graph convolutional network. In *RAID*, 241–256 (2020).
- Wu, H., Li, X. & Gao, Y. An effective approach of named entity recognition for cyber threat intelligence. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, 1370–1374 (IEEE, 2020).
- Qin, Y. *et al.* A network security entity recognition method based on feature template and cnn-bilstm-crf. *Frontiers of Information Technology & Electronic Engineering* **20**, 872–884 (2019).
- Wang, X. *et al.* A method for extracting unstructured threat intelligence based on dictionary template and reinforcement learning. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 262–267 (IEEE, 2021).
- Luo, L. *et al.* A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of biomedical informatics* **103**, 103384 (2020).
- Li, T., Guo, Y. & Ju, A. Knowledge triple extraction in cybersecurity with adversarial active learning. *J. Commun* **41**, 80–91 (2020).
- Zuo, J., Gao, Y., Li, X. & Yuan, J. An end-to-end entity and relation joint extraction model for cyber threat intelligence. In *2022 7th International Conference on Big Data Analytics (ICBDA)*, 204–209 (IEEE, 2022).
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (2019).
- Guo, Y. *et al.* 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**, 447–463 (Springer, 2021).
- Mittal, S., Das, P. K., Mulwad, V., Joshi, A. & Finin, T. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 860–867 (IEEE, 2016).
- Dionísio, N., Alves, F., Ferreira, P. M. & Bessani, A. Cyberthreat detection from twitter using deep neural networks. In *2019 international joint conference on neural networks (IJCNN)*, 1–8 (IEEE, 2019).
- Satyapanich, T., Ferraro, F. & Finin, T. Casie: Extracting cybersecurity event information from text. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 8749–8757 (2020).
- Sommer, F., Dürrwang, J. & Kriesten, R. Survey and classification of automotive security attacks. *Information* **10**, 148 (2019).
- Menges, F., Sperl, C. & Pernul, G. Unifying cyber threat intelligence. In *Trust, Privacy and Security in Digital Business: 16th International Conference, TrustBus 2019, Linz, Austria, August 26–29, 2019, Proceedings* **16**, 161–175 (Springer, 2019).
- Syed, Z., Padia, A., Finin, T., Mathews, L. & Joshi, A. Uco: A unified cybersecurity ontology. In *Workshops at the thirtieth AAAI conference on artificial intelligence* (2016).
- Wang, Y. *et al.* Automotive cyber threat intelligence corpus. *Figshare* <https://doi.org/10.6084/m9.figshare.27916758.v1> (2024).

Acknowledgements

This work was supported by the Industrial Internet Innovation and Development Project in 2023 under Grant 0747-2361SCCZA196, the Research on Vehicle-Road Integration Intelligent Transportation Data Model and Cybersecurity Technology Project of the Shandong High-speed Innovation Research Institute, as well as the State Key Laboratory of Intelligent Transportation System.

Author contributions

Y.W., Y.R., H.Q., Z.C., Y.Z. and H.Y. prepared the manuscript. Y.W. established the data collection and processing methodology, implemented the data processing, and performed data validation experiments. Y.R. and H.Y. provided financial support. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.R. or H.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025