scientific reports

Check for updates

OPEN

Automatic construction of risk transmission network about subway construction based on deep learning models

Yanxiang Liang[1], Na Xu[1]✉, Hong Chang[2], Shan Qian[1] & Yao Liu[1]

Safety risks management is a critical part during the subway construction. However, conventional methods for risk identification heavily rely on experience from experts and fail to effectively identify the relationship between risk factors and events embedded in accident texts, which fail to provide substantial guidance for subway safety risks management. With a dataset comprising 562 occurrences of subway construction accidents, this study devised a domain-specific entity recognition model for identifying safety hazards during the subway construction. The model was constructed by a Bidirectional Long Short-Term Memory Network with Conditional Random Fields (BiLSTM-CRF). Additionally, a domain-specific entity causal relation extraction model employing Convolutional Neural Networks (CNN) was also developed in thsi model. The constructed models automatically extract safety risk factors, safety events, and their causal relationships from the texts about subway accidents. The precision, recall, and F_1 scores of Metro Construction Safety Risk Named Entity Recognition Model (MCSR-NER-Model) all exceeded 77%. Its performance in the specialized domain named entity recognition (NER) with a limited volume of textual data is satisfactory. The Metro Construction Safety Risk Domain Entity Causal Relationship Extraction Model (MCSR-CE-Model) achieved an impressive accuracy, recall, and F_1 score of 98.96%, exhibiting excellent performance. Moreover, the extracted entities were normalized and domain dictionary was developed. Based on the processed entities and relationships processed by the domain dictionary, 533 domain entity causal relation triplets were obtained, facilitating the establishment of the directed and unweighted complex network and case database about the risks of subway construction. This research successfully converted accident texts into a causal chain structure of "safety risk factors to risk events," providing detailed categorization of safety risks and events. Concurrently, it revealed the interrelationships and historical statistical patterns among various safety risk factors and categories of risk events through the complex safety risks network. The construction of the database facilitated project managers in conducting management decisions about safety risks.

Keywords Metro construction, Safety risks management, BiLSTM-CRF, CNN

Metro infrastructure construction is a dynamic systems engineering with complex phenomena and chaotic characteristics[1]. A variety of elements such as intricate geological and hydrological settings can result in formidable challenges of construction and organizational coordination[2,3]. Safety hazards associated with subway construction are intricate, veiled, and dynamic, which may lead to financial losses, personal harm, ecological disruption, project setbacks and diminished structural integrity[4,5]. Traditionally, the identification and analysis of safety risks in context relied on the industry specialists, academics, and seasoned project leaders[6]. With the accumulation of historical data, the distinctive spatiotemporal attributes, highly nonlinear aspects, and intricate interconnections presented formidable hurdles for the comprehensive analysis[7,8]. The perception, analysis and inference of risks from security experts are also inevitably affected by cognitive bias and individual subjectivity[9]. To solve the problems mentioned above, this study conducted automatic risks identification of accident data through data mining technology based on the experience of experts, explored safety hazards and related rules in development and evolution. This study can are significant to make up for individual subjective limitations of industry staff, which can improve the safety risks management level.

[1]School of Mechanics and Civil Engineering, China University of Mining and Technology, Xuzhou 221116, China. [2]Shenzhen Urban Public Safety and Technology Institute, Shenzhen 518000, China. ✉email: xuna@cumt.edu.cn

In the realm of subway construction safety risks management, the analysis of accident cases has been widely utilized to facilitate the administration and enhancement of engineering safety, thereby serving as a potent resource for mitigating similar hazardous situations and risk incidents[10–12]. At present, researchers mostly focus on individual risk events or specific types of subway accidents. Relevant researches are all about the descriptive statistics for risk incidents in specific country or region[13–15]. However, risks do not arise suddenly and in isolation[16]. A multitude of interrelations exists among risk events, which is often overlooked by singular case analysis[17,18]. In a specific event, risk factors lead to the occurrence of original risk events, secondary events and derivative events in turn, forming a complete causal chain of risk transmission. The combination of multiple risk chains formed a risk network[19,20]. By mining and analyzing sets of causal chains in risk transmission, the interactions among various risk factors and events can be revealed in comprehensive, multi-category accident investigations. This approach effectively circumvents the limitations inherent in singular risk factor and event analysis, contributing to the continual improvement the level of a more systematic, comprehensive safety risks management.

Therefore, BiLSTM-CRF and CNN models were introduced in this study for entity recognition and relationship extraction in domain accident texts. BiLSTM was used to capture long-term dependencies in sentences. CRF play the role of addressing sequence labeling tasks and enhances named entity recognition. CNN can be used to reduce parameter count lowers computational costs of model, enabling parameter sharing and sparse connections. Consequently, the automatic extraction and transformation from accident reports of subway construction to causal chains with the structure of 'safety risk factors - risk events', addressing the deficiency of risk identification overly reliant on expert experience. This contributes enhanced the efficiency of domain-specific safety risks recognition. By constructing a set of causal chains for metro construction safety risks, it unveiled the intricate impact relationships and risk transmission pathways among various risk factors and events in Chinese subway construction spanning nearly two decades from a multi-case perspective. This approach resolved the problems of limited study cases and singular categories, mitigating the individual subjectivity and limitations associated with conventional analysis.

For accident texts of subway construction, we developed and trained a model for entity recognition in the domain of safety risks in subway construction by using a Bidirectional Long Short-Term Memory Network combined with Conditional Random Fields (BiLSTM-CRF). It can facilitate the automatic extraction of safety risk factors and domain entities in risk events from accident texts. We established and trained a causal relationship extraction model based on Convolutional Neural Networks (CNN) to extract causal relationships among domain entities. It also can automatically construct a causal chain of "subway construction safety risks factors-risk events," thereby revealing the universal laws of interaction among risk factors and risk events. In order to further clarify the research, the following assumptions were proposed: The BiLSTM-CRF and CNN models could accurately extract safety risk factors and causal relationships from textual data. The event text data represented the authentic safety risks scenarios in subway construction. Although all risk factors were not covered, its comprehensiveness was sufficient to support the objectives of this research.

Literature review

Construction safety risks identification based on text mining

The identification of safety risks constitutes is a complex system engineering task. Conventional approaches to safety risks identification primarily concentrated on individual risk factors and specific types of accidents. As for the safety risks identification in subway construction, the conventional focused on particular construction procedures and stages, employing expert surveys and interviews, brainstorming sessions and literature research as methods for risk recognition. These approaches heavily rely on experience, which are susceptible to individual cognitive limitations and subjective factors. Fang et al. inferred that nearby pipelines and existed buildings are the primary risk factors during the subway construction based on process control and situational surveys[21]. Meanwhile, Zhang et al. investigated the interaction between safety risks management performance and the perceived significance of each risk factor by conducting surveys and semi-structured interviews with subway construction workers in the Southeast region[22]. Shi et al. (2024) utilized text mining techniques and DEMATEL-ISM method to identify and evaluate safety risk factors[23]. Researchers proposed the subway construction safety risks identification and early warning systems based on construction drawings, Internet of Things, BIM, and other technological tools, progressively broadening the scope and categories of risk identification. Li et al. introduced the BIM-based subway construction safety risks identification and early warning system. It utilized engineering parameter information to achieve safety risks identification[24]. Guo et al. creatively combined BIM with D-S evidence theory to enhance risk management capabilities for complex underground projects[25]. However, the majority of risk identification data sources come from numerical data collected by construction machinery and image data obtained through optical equipment. It leads to the limited reports about emerging hazard patterns and subtle differences inferred from unstructured and semi-structured textual data such as subway accident reports and records. Furthermore, the constraint in data categories makes it difficult for risk identification to encompass all risk factors and events.

Subway construction safety management based on natural language processing

Natural Language Processing (NLP) is utilized to facilitate the comprehension of human language systems in computer. It is primarily applied in tasks such as text classification, information recommendation, and information extraction[26]. At present, NER mainly has three mainstream methods in safety management about subway construction: rule-based, statistical machine learning method, and deep learning method[27]. Tang et al. utilized text mining to extract risk data from texts, guiding on-site management of subway construction[28]. Huo et al. utilized text mining to extract key features related to subway accidents from raw data, developing a new causal path selection model[29]. By interrupting causal propagation along these paths, construction safety can be

enhanced. Rules are created by experts and scholars in professional fields to meet their own research needs. Li et al. performed entity recognition from human factors, management, and risks by BP neural network model[30]. The recognition results were used to predict potential accident types and propose safety management measures during the construction. Machine learning method requires high text standardization. In addition, it only can operate on limited data volumes and generally exhibit moderate effectiveness in entity recognition. Deep learning methods exhibited excellent performance in NER recognition, leading to further improvements in entity identification accuracy and efficiency[31]. Zhou et al. developed a double deep Q-network deep reinforcement learning model to predict subway construction safety risks, which is conducive to enhancing safety management at subway construction sites[32]. There remains a scarcity of research utilizing deep learning techniques for the NLP mining and analyzing of accident investigation reports, safety records, and other accident texts about subway construction in scholarly literature. With data mining remaining predominant, the entity recognition of safety risks in subway construction is still at an early stage.

Research methods

Research framework

The knowledge structure framework of safety risks in subway construction is depicted in Fig. 1. This framework can be segmented into four sections: corpus construction, entity recognition model, relationship extraction model, and the construction of a subway construction safety risks network. To ensure the representativeness of the results, knowledge extraction was conducted on 562 accident texts among 20 years. The BiLSTM-CRF model was utilized for entity recognition in safety risks about subway construction, while a convolutional neural network model was employed to extract causal relationships among domain entities. In order to get a more comprehensive analysis of safety risk factors and risk events, this research normalized safety risks factors and risk event entities based on the named entity recognition of domain entities, constructing a domain synonym dictionary. Ultimately, this research identified 533 causal relationship triplets in the domain of subway construction safety risks, which served as the basis for constructing a directed unweighted complex network and case database for safety risks in subway construction.

BiLSTM-CRF model

The research employed the BiLSTM-CRF as the deep learning framework for named entity recognition in safety risks of subway construction. BiLSTM is a type of recurrent neural network that processes information from both directions of a sequence to capture context effectively. The CRF is a discriminative probabilistic undirected graph model, which represents the conditional probability distribution of one set of random variables under another given distribution of random variables[33]. The BiLSTM model effectively captures long-range dependencies by processing sequence information bidirectionally, thereby enhancing the comprehensive utilization and understanding of contextual information. CRF is well-suited for sequence labeling tasks, ensuring the consistency and contextual relevance of predicted entity labels. The BiLSTM-CRF model is adept at handling text characterized by sequential patterns, necessitating the capture of long-range dependencies and ensuring coherent labeling within context. Its performance shines in named entity recognition, where contextual comprehension is pivotal for precise predictions. In this research, the output entities in accident texts were labeled and got the optimal global label sequence with the constraints of CRF. This method considers the influence of label results from other characters during the output of labels, effectively enhancing the recognition effectiveness of this model. Initially, the subway construction accident text was converted into a character vector representation on a per-character basis, denoted as $\{x_1, x_2, \ldots, x_n\}(x_t \in [1,n])$, serving as the input data for this model. The word vector features utilized the 100-dimensional "Chinese word vector library" trained from Wikipedia, encompassing 16,991 characters, to effectively express character features. Subsequently, the data was

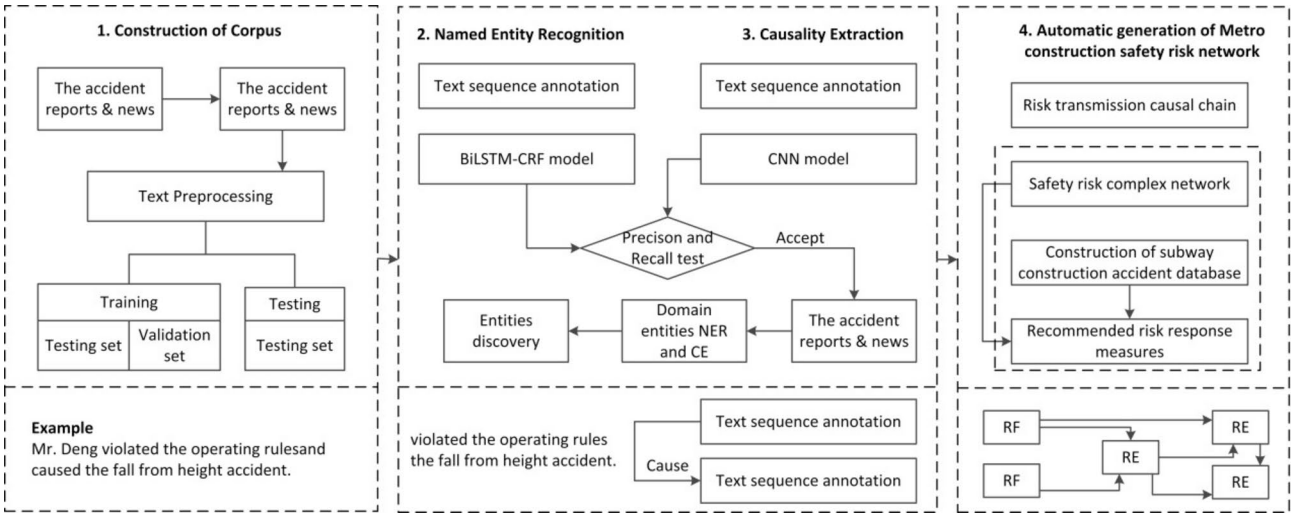

Fig. 1. Research framework.

input into the LSTM neural network in a forward and backward sequences to obtain forward and backward hidden vectors ($\overrightarrow{h_t}$ and $\overleftarrow{h_t}$) containing semantic information about the accidents in subway construction. The obtained two vectors were concatenated to form the final output vector h_t, serving as the input for the CRF layer. Finally, the CRF model was employed to obtain the predicted labels for named entities in safety risks of subway construction, which are displayed through the output layer.

Convolutional neural network model

The research utilized a deep learning framework based on CNN for extracting entity causal relationships in safety risks of subway construction. The CNN model consists of embedding layer, convolutional layers, pooling layer, and fully connected layer. The embedding layer represents word vectors through embedding word features and position features. The convolutional layer captures the overall semantic information of sentences[34]. The pooling layer compresses the results of the convolutional layer using max-pooling to extract significant features, control overfitting, and obtain the feature vector of a sentence[35]. The fully connected layer integrates highly abstracted features obtained through multiple convolutions to produce output probabilities for various classification, that is, relationship classification results[36]. The training process of this model is depicted in Fig. 2.

Model evaluation criteria

The recognition performance of the Metro Construction Safety Risk Named Entity Recognition Model (MCSR-NER-Model) and the Metro Construction Safety Risk Domain Entity Causal Relationship Extraction Model (MCSR-CE-Model) were evaluated based on three evaluation metrics: Precision (P), Recall (R), and F_1 Score[37]. The introduction and calculation formulas for each metric are displayed as follows.

Precision represents the proportion between the number of correctly identified entities and the total number of identified entities.

$$P = \frac{TP}{TP + FP} \times 100\%$$

Recall indicates the proportion between the number of correctly identified entities and the total number of pre-labeled entities.

$$R = \frac{TP}{TP + FN} \times 100\%$$

The F_1 Score is the weighted geometric mean of precision (P) and recall (R).

$$\mathrm{F_1} = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}} \times 100\%$$

In the above formulas, TP represents the number of samples predicted as positive class that are actually positive, FN represents the number of samples predicted as negative class that are actually positive, and FP represents the number of samples predicted as positive class that are actually negative.

Experiment and results

Data acquisition

The data set utilized in this research are the 562 accident texts collected from March 2001 to November 2021, including 130 reports of accidents about subway construction and 432 accident bulletins published by the Ministry of Housing and Urban-Rural Development and news media (Table 1). The dataset consists of encompassing 1821 informative sentences. In contrast to free text, these accident investigation reports contain extensive descriptions of safety risk factors, accident circumstances, outcomes, impacts, and responsibilities, which can facilitate the comprehensive analysis and causal chain delineation of accidents. It also can promote the risk event data mining, and structured documentation. The bulletins and notices published by the Ministry of Housing and Urban-Rural Development succinctly described subway accident causes, risk events name, risk outcomes, and their consequences. It can help to clarify the critical information about subway accidents, such as safety risk factors and accident outcomes.

In the accident texts of subway construction, the descriptions of safety risk factors and risk events are usually stated in the form of proprietary nouns and phrases, such as “violations in operations” and “overloaded vehicle cargo”. The descriptions of risk events are characterized by standardization and uniformity, such as “foundation pit collapse” and “objects striking.” The phrases composed of a small number of Chinese characters that can exist independently and convey a state, attribute, or explicit meaning are referred to as entities. With mining these entities, it can swiftly specifying key information such as the causes, names, and outcomes of risk events. The entities related to safety risks of subway construction studied in this research include two main categories, safety risk factors and risk event (RE). Safety risk factors represent the causes that may lead to risk events in subway construction, consisting of human factor (HF), material defect (MD), environmental factor (EF), and technical and management factor (TM)[38–41]. Risk events refers to the ultimate results caused by various safety risk factors, involving subway accidents, casualties, equipment losses, etc.

Text preprocessing

First, the collected accident texts of subway construction in various formats, such as word, PDF, images, and web pages, were converted into TXT format text documents with UTF-8 encoding. Next, each document was

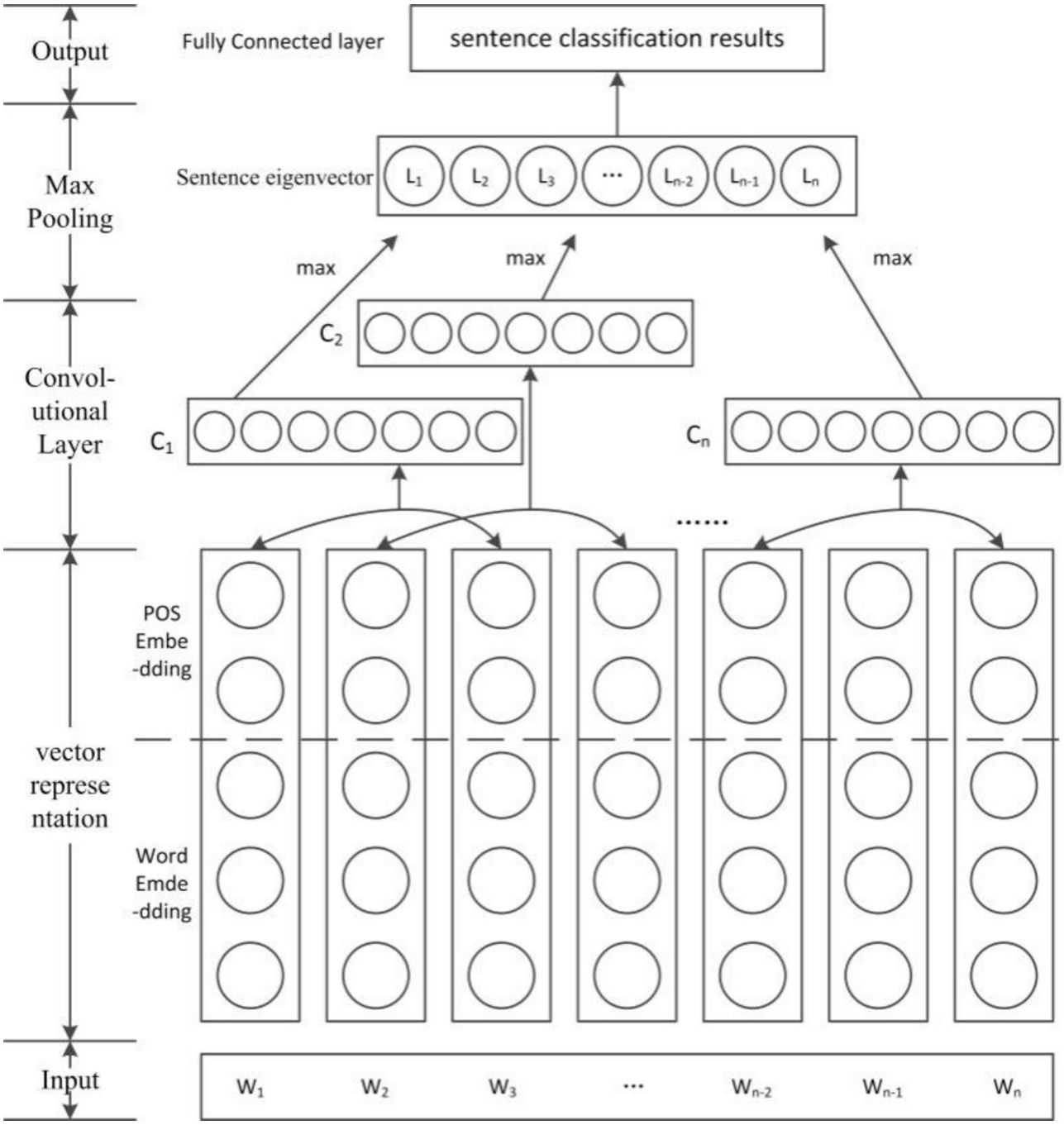

Fig. 2. CNN model.

Type	Sources	Texts	Sentence count
Subway construction accident reports	Survey of subway construction companies	130	421
Subway construction Safety accident briefing	Ministry of housing and urban-rural development	96	312
	News media	336	1088

Table 1. The information of data acquisition.

The types of stop words	Specific stop words
Functional words	And, however, have been
Quantitative word	One, two, three
Degree words	Significant, very, obvious, complete, large scale
Words represent the change of state	Occurrence, initiation, existence, appearance, therefore

Table 2. Stop words list in accident texts of subway construction.

Labels	Meaning	Labels	Meaning
B-HF	Human factor category initials	B-TM	Technical and management factors category initials
I-HF	Inside and end of human factor category	I-TM	Inside and end of technical and management factors category
B-EF	Environmental factor category initials	B-RE	Risk event category initials
I-EF	Inside and end of environmental factor category	I-RE	Inside and end of risk event category
B-MD	Material defect category initials	O	Non-entity characters
I-MD	Inside and end of material defect category		

Table 3. Definitions of label categories.

Types of domain entities	Number of entities	Number of annotation	Examples of entities
HF	141	287	Illegal operation, careless operation, smoking…
MD	123	134	Mechanical failure, cable short circuit, pile deformation, support system instability…
EF	106	191	Rain, strong typhoons, weak foundations, toxic gases, slippery ground…
TM	356	596	Ineffective hidden danger investigation, illegal subcontracting, chaotic safety management…
RE	613	2108	Water gushing, collapse, fall, lifting injury…
Total	1339	3316	

Table 4. The annotation results of entity sequence in accidents texts of subway construction.

reviewed to correct the misspellings, language expression errors, or semantic inaccuracies, ensuring precise and correct language usage[42]. Finally, the accident texts were segmented into individual sentences. A total of 1821 sentences related to subway construction accidents were obtained.

The accident texts of subway construction are unstructured texts that written in natural language. The colloquial expressions are inevitable. Stop words refer to high-frequency, low-value words lacking actual meaning, removal of which can effectively reduce text feature dimensions and enhance entity recognition. By manually analyzing accident texts of subway construction and reviewing domain-specific literature, the expression patterns for entities such as safety risk factors and risk events were identified. Finally, a stop words list for accident texts of subway construction were summarized and presented in the Table 2.

Python code was employed to remove stop words in the 1821 texts to mitigate the impact of stop words on subsequent text mining tasks and enhance the accuracy of entity recognition and relationship extraction.

Domain named entity identification

Text sequence annotation

The process of text sequences annotation related to safety accidents of subway construction involves the identification of individual risk factors, domain entities in risk events and attributes of these entities. This annotation process is documented for reference and analysis The text annotation for accident texts of subway construction follows the "BIO encoding format"[43]. The quality of the labeled corpus plays a decisive role in the performance of the training deep learning models[44]. Table 3 shows the definitions of label categories. The annotation was carried out by trained personnel from domain experts and the research team. Following the initial annotation, inter-annotator agreement was ensured through cross-validation among annotators.

The sequence annotation results were stored in TXT format documents with UTF-8 encoding. Annotation tasks on 1614 accident texts were completed with Python. Table 4 provided a statistical summary of the annotated quantities for each category of domain entities in texts. The example of entity names were also displayed.

Training data structure and environment configuration

The programming language and version utilized for constructing the Chinese named entity recognition model in safety risks of the subway construction is Python 3.6.5. The training framework of deep learning model is based on TensorFlow 1.13.1. The code was carried out on PyCharm Community Edition 2021.2.3. The operating system is a 64-bit Windows 10 system. Due to the relatively small scale of the experimental data, computations

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
Word_dim	100	Max_epoch	200	Dropout_keep	0.5	Learning-rate	0.001
Lstm_dim	100	Batch_size	20	Tag_schema	BIOES	Pre_emb	True
Optimizer	Adam	Clip	5.0				

Table 5. MCSR-NER-Model parameter settings.

Testing set	Processed batches: 14,084, pre-labeled phrases: 656 ; identify: 669; Correct: 520			
Entities types	Precision	Recall	F1 value	Number of identified entities
Total	77.73%	79.27%	78.49%	669
EF	53.85%	61.76%	57.53%	39
HF	77.78%	83.33%	80.46%	45
RE	84.58%	85.94%	85.26%	441
MD	68.18%	65.22%	66.67%	22
TM	62.30%	61.79%	62.04%	122

Table 6. Identification results of MCSR-NER-model.

New entities name	Type	New entities name	Type	New entities name	Type
Venture into	HF	Cable short circuit	MD	Pipe burst	RE
Hazardous area construction	HF	Relaxed quality Control	TM	Face of the palm Collapse	RE
Hurricane	EF	Steel frame collapse	RE	Gas line fracture	RE
Wet and slippery	EF	Vehicle rollover	RE	Building crack	RE

Table 7. Examples of the fresh domain-specific entities.

were performed by CPU, featuring an Intel Core i7-6700HQ CPU @ 2.60 GHz. The initial parameter settings for the MCSR-NER Model are outlined in Table 5.

After preprocessing, 1614 sentences from accident texts of subway construction were used for model training and validation. The data was shuffled by the shuttle program to introduce randomness and improve the generalization performance of the neural network, thereby mitigating the effect of text input order on model training. The subway construction accident texts were partitioned into training, cross-validation, and test sets in a 7:1:2 ratio. The training set, containing 1130 sentences, was used for model training to learn text features. The cross-validation set, consisting of 161 sentences, was employed for automatic adjustment of model parameters during the training process.The testing set comprises 323 sentences and it was utilized to assess the performance of the trained model.

Analysis of training results

The prediction results of (MCSR-NER) model were assessed by the Python version of conlleval.pl. The evaluation results were presented in Table 6. The results encompasses the overall model performance and the statistical outcomes related to the recognition of individual entity types.

The trained MCSR-NER-Model achieved precision, recall, and F1 values all exceeding 77%. The entity recognition task was challenged by the moderate scale of processed text in this research and the diversity of expressions encountered in texts. As for the recognition task of named entity in specialized domain based on a relatively limited amount of textual data, the performance MCSR-NER-Model is considered quite favorable. The MCSR-NER-Model effectively captured the language features and semantic expressions in accident texts of subway construction. Subsequently, the remained 207 accident-related texts were input into the trained MCSR-NER-Model to identify safety risk factors and risk event entities. A total of 370 entities in safety risks of subway construction were identified.

During the process of entity recognition in the safety risks of subway construction, it was observed that some specific nouns unique to the subway construction were not identified completely, such as “shield machine is flooded” and “face of the palm collapsed”. It is imperative to continue enriching domain-specific vocabulary and improving the recognition capabilities of this model by embedding domain-specific dictionaries.

In terms of various entity recognition, the risk event (RE) category exhibited the most promising results, with the highest F_1 value reaching 85.26%. Several factors contribute to this result. (1) In training text data set, the number of annotation about risk event types is the largest, and the description of accidents is more accurate and unified than other types. (2) The text position of the risk event category is more fixed. It generally exists at the beginning or the end of the whole sentence. For example, “the object strike accident is caused by the fatigue of the construction personnel”, and the accident entity is the object strike. “Construction personnel did not

hang the safety rope according to the regulations, violate the operating procedures during the process of high edge operation, resulting in a high fall accident." (3) The risk event entity is often accompanied by words such as "reason", "cause", "lead" and "trigger". These high-frequency words promote the location of entity about risk event. Therefore, risk event can be easily identified by the model.

It was noted that the environmental factor (EF) exhibited the lowest recognition effectiveness. It can be attributed to the scarcity of environment factor entities and their corresponding category labels in training text. The word "gas" appeared only twice in training text. In addition, some longer expressions about environment factor are failed to be identified. For instance, "Coupled with the early rain, the soil is soaked and loosened, and once the soil layer above the pipeline is not compacted, it is easy to cause the earth to collapse", the "rain" was correctly identified, while "soil is soaked and loosened" was not recognized.

Domain entity discovery

The MCSR-NER-Model has assimilated fresh domain-specific entities, encompassing safety risk factors and accidents of subway construction that extracted from the 207 accident-related texts. They surpassed the initial 1614 texts utilized for model inception. The fresh domain-specific entities are delineated in Table 7. Considering the limited textual data of accidents about subway construction, and the number of new domain entities is relatively small. With the continuous expansion and enrichment of accident texts, more domain entities will be identified, and the recognition performance of domain entities will be enhanced.

With the consolidation of domain entities based on pre-annotated and the outcomes furnished by the trained MCSR-NER-Model, 1361 entities of safety risks about subway construction were obtained. The risk factors of subway construction encompass 147 entities characterizing human-centric factor, 124 entities emblematic of material defect, 107 entities reflective of environmental influence and 358 entities epitomizing technological and management problem. A total of 625 entities about risk accidents were acquired.

Causality extraction about domain entities

Relationship types identification

A supervised learning approach was utilized for relation extraction to delineate three distinct causal relationship categories among domain entities: causes, effects, and co-occurrences. The extraction outcomes were represented as triplets (entity1, entity2, relation). These relationships are defined as follows: cause: the occurrence of entity1 is caused by entity 2; effect: the occurrence of entity1 leads to the occurrence of entity 2; accompany: entity 1 and entity 2 frequently occur together. With the manual examination of accident texts about subway construction, it was observed that only partial texts containing domain entities exhibit causal relationships. They primarily encompass the following cases. Firstly, the sentences of accident texts only contain one domain entity. Next, sentences encompass two entities, but a causal relationship does not exist in entities. Finally, sentences comprise three or more entities, the causal relationships may exist among the domain entities. Considering these issues, a set of rules was formulated to filter out the sentences without causality, as depicted at Fig. 3.

Text sequence annotation

The text annotation format for relation extraction differs from that of the recognition of entities. In the training corpus of relation extraction, each individual sentence occupies a distinct line and includes the following components: [sentence, relation, head, head_type, head_offset, tail, tail_type, tail_offset]. The specific meanings of each component is elaborated in the Table 8.

The domain entity and its location information in training and testing sentences were obtained by regular expression and implemented by RE library of Python. The annotation results were stored in csv format.

Training data structure and hardware construction

This research collected and organized 996 preprocessed sentences containing causal relationships among domain entities from accidents of subway construction for model training and validation. The shuttle program was used to increase randomness and enhance the generalization performance of neural network. The accident text data of subway construction was divided into training set and testing set in an 8:2 ratio, where the training set and testing set comprises 798 and 198 sentences, respectively.

In order to improve the effectiveness of relationship extraction, a supervised learning approach based on CNN deep learning model for relation extraction was utilized. The MCSR-CE-Model was constructed. The programming language and version used for constructing the extraction model of causal relationship among domain entities is Python 3.6.5. The training framework of neural network model is PyTorch. The experimental code utilized PyCharm Community Edition 2021.2.3, and the operating system employed a 64-bit Windows 10. GPU was utilized in this research, with the NVIDIA GeForce GTX 960 M graphics card. The parameters of MCSR-CE-Model are presented in Table 9.

Data name	Specific meaning	Data name	Specific meaning
Sentence	Statements to be trained and tested	Head_offset	Location information of the head domain entity
Relation	Causality type	Tail	Tail domain entity
Head	Head domain entity	Tail_type	Entity type of the tail domain entity
Head_type	Entity type of the head domain entity	Tail_offset	Location information of the tail domain entity

Table 8. Annotation structure description of domain entity causality extraction text.

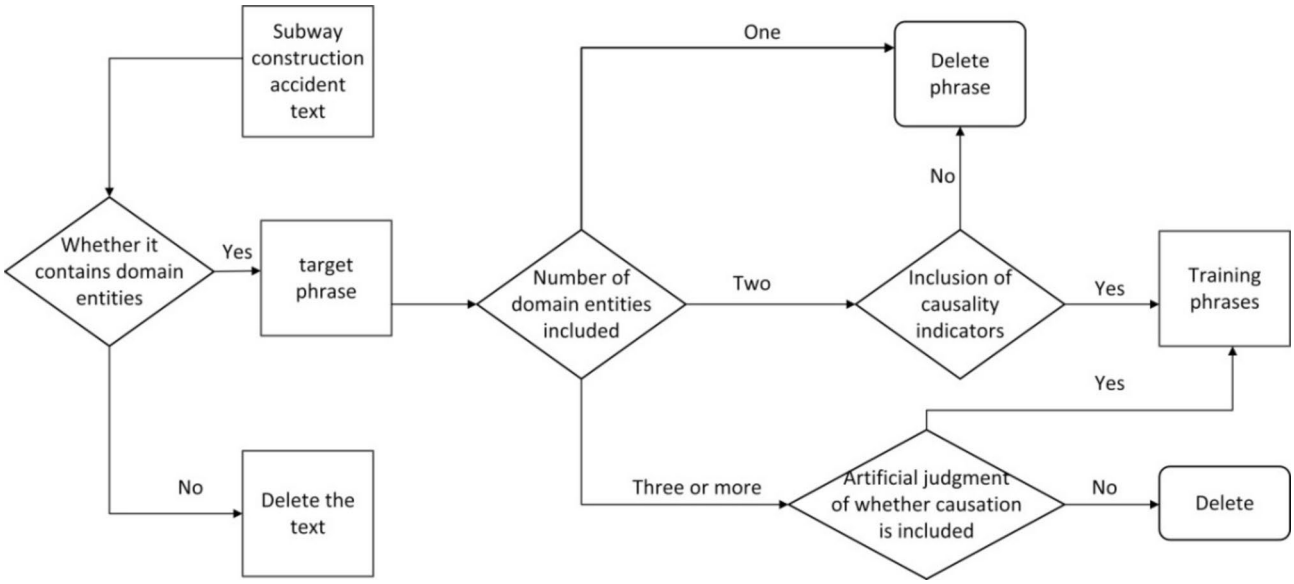

Fig. 3. Text selection process of entity causality extraction in safety risks of subway construction.

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
Hidden_size	100	Dropout	0.5	Epoch	50	Word_dim	300
Batch_size	32	Learning_rate	0.001	Pos_dim	10	Out_channels	100
Decay_rate	0.3	Decay_patience	5	Kernel_size	[3, 5]		

Table 9. MCSR-CE-Model parameters setting.

Evaluation criteria	Precision	Recall	F_1-Score
Value	98.96%	98.96%	98.96%

Table 10. Extraction results of MCSR-CE-Model.

Training results analysis

Extraction results conducted by MCSR-CE-Model are shown in Table 10.

MCSR-CE-Model achieved high accuracy, recall, and an F1 score of 98.96%, indicating the excellent performance. It can be attributed to several factors. The research aimed to explore the influence between risk factors and risk events in subway safety. Therefore, the relationship was constrained to causality, and initial text filtering was conducted based on this rule. It can significantly reduce the interference of sentences without causality. In addition, the extraction of relationship only focused on "cause, effect, and co-occurrence".

In this section, 996 out of 1614 accident texts containing entity causal relationships in subway construction were utilized for training and testing the MCSR-CE-Model. The remained 207 accident texts were extracted the causal relationships of entities, 163 sentences containing causality were obtained. During the causality extraction in the safety risks of subway construction, the majority of prediction errors are associated with the "co-occurrence" relationship. For instance, "The mishandling from worker caused the scaffold to fall onto a gas cylinder, leading to the fire and explosion," the manually labeled relationship between "fire" and "explosion" is "co-occurrence". However, this relationship was predicted as "effect" by this model. The primary reason lies in the limited training data for the "co-occurrence" relationships, which only contain 25 instances out of 798 training texts. The scarcity of training data hinders the model to effectively learn the structural characteristics of this kind of text. Additionally, the manual analysis and review of text containing "co-occurrence" revealed the minimum differences in comparison to the other two relationships, increasing the difficulties of learning for the model.

In conclusion, the causality results obtained from the relational extraction model and the relationships manually marked during the model training and testing were obtained. Finally, 1159 sets of causal triad structure were obtained, including 319 sets of "cause" relationships, 811 sets of "effect" relationships, and 29 sets of "co-occurrence" relationships. The results of causality extraction are basically consistent with the results of extraction by industry experts.

Number	Safety risk factor	Number	Safety risk factor	Number	Safety risk factor
HF1	Fatigue operation	HF6	Illegal operation	HF11	Violation of labor discipline
HF2	Misoperation	HF7	Insufficient safety distance	HF12	Lack of awareness of risks
HF3	Careless operation	HF8	Rest in construction area	HF13	Safety equipment is not equipped or properly used
HF4	Distraction	HF9	Smoking in the workplace		
HF5	Vacant or absent	HF10	Professional skills are not proficient		

Table 11. Risk factors of subway construction - human factor(part).

Number	Safety risk factor	Number	Safety risk factor	Number	Safety risk factor
MD1	Mechanical failure	MD6	Material quality is not qualified	MD11	Improper storage of on-site materials
MD2	Vehicle overloading	MD7	Tools aging or defect	MD12	Deformation or displacement of supporting system
MD3	Irregular stacking of materials and machinery	MD8	Pipe aging or broken	MD13	The support system is unstable or defective
MD4	Extrusion deformation of material	MD9	Wire circuit or leakage		
MD5	Insufficient structural strength	MD10	Safety protection deficiency		

Table 12. Risk factors of subway construction-material defect (part).

Number	Safety risk factor	Number	Safety risk factor	Number	Safety risk factor
EF1	Typhoon	EF6	Uneven ground	EF11	Toxic and harmful gas
EF2	Rain	EF7	Soil disturbance	EF12	Lack of soil stability
EF3	Soft soil	EF8	Slippery road surface	EF13	Complicated hydrogeological conditions
EF4	Weak foundation	EF9	Groundwater abundance	EF14	Long exposure time of foundation pit
EF5	Geological cavity	EF10	Poor geological condition		

Table 13. Risk factors of subway construction-environmental factor (part).

Number	Safety risk factor	Number	Safety risk factor	Number	Safety risk factor
TM1	Rush to finish work	TM9	Construction quality defect	TM17	Not strictly perform the duties of the post
TM2	Unlicensed operation	TM10	Construction process error	TM18	Insufficient on-site supervision and inspection
TM3	Illegal contracting	TM11	Chaotic site management	TM19	Insufficient supervision
TM4	Unqualified	TM12	Construction plan preparation defect	TM20	Insufficient regulatory responsibility of the industry
TM5	Poor engineering exploration	TM13	Inadequate Emergency response measures	TM21	Inadequate safety precautions
TM6	Poor engineering monitoring	TM14	Illegal organize operations	TM22	Inadequate safety management
TM7	Engineering design defect	TM15	Inadequate construction methods and measures	TM23	Inadequate safety training and education
TM8	Construction technical defect	TM16	Construction organization and management omission	TM24	Failure to investigate and eliminate security risks in time

Table 14. Risk factors of subway construction-technical and management factor (part).

Construction of domain dictionary

Normalization of domain entities

The diverse origins of accident information lead to the distinct formatting standards, resulting significant disparities in the description of similar or related risk factors and events about subway construction across various accident texts[45]. This research systematically organized all safety risk factors and events, meticulously analyzed the linguistic expressions of each entity, integrated and summarized similar expressions referencing national standards about accidents classification. Subsequently, the preliminary summary was improved by subway project managers, construction technical leaders, university researchers, and graduate students. Ultimately, 1361 entities were divided into four major classes of risk factors about subway construction and 56 types of risk events. Specifically, the risk factors of subway construction encompass human factors (HF) with 13 categories totaling 147 entities, material defect factors (MD) with 13 categories totaling 124 entities, environmental factors (EF) with 14 categories totaling 107 entities, technical and management factors (TM) with 24 categories totaling 358 entities, risk events (RE) with 56 categories totaling 625 entities. The classification results and codes for the risk factors of subway construction and risk events are exhibited in Tables 11, 12, 13, 14 and 15.

Number	Safety risk factor	Number	Safety risk factor	Number	Safety risk factor
RE1	Hydrops	RE20	Pit collapse	RE39	Gas leakage
RE2	Water burst	RE21	Trench collapse	RE40	Explosive deflagration
RE3	Water inrush	RE22	Trench collapse	RE41	Geological mutation
RE4	Water leakage	RE23	Structural collapse	RE42	Traffic jam
RE5	Sand burst	RE24	Top off the sheet	RE43	Injury from Vehicle machinery
RE6	Fire	RE25	Pavement collapse	RE44	Lifting injury
RE7	Drowning	RE26	Station collapse	RE45	Object strike
RE8	Poisoned	RE27	Interval collapse	RE46	Object fall
RE9	Electric shock	RE28	Secondary collapse	RE47	Fall from high places
RE10	Power cut	RE29	Frame collapse	RE48	Structural crack of component
RE11	Water cut	RE30	Earth collapse	RE49	Support system failure
RE12	Gas cut	RE31	Wall topple down	RE50	Tools fail to stabilize and overturn
RE13	Bury	RE32	Materials topple down	RE51	Structure is unstable and overturns
RE14	Soil erosion	RE33	Structural damage of component	RE52	Other injuries
RE15	Soil landslide	RE34	Building damage	RE53	Injury
RE16	Pit Subsidence	RE35	Pipeline leakage	RE54	Coma
RE17	Pavement subsidence or cracks	RE36	Water pipe break	RE55	Asphyxia
RE18	Tunnel depression	RE37	Pipeline explosion	RE56	Death
RE19	Secondary depression	RE38	Tracheal rupture		

Table 15. Types of risk events of subway construction.

Domain entity	Types	Synonym representation
Illegal operation	Human factor	Violation of rules and regulations, illegal construction, violation of legal provisions, violation of safety management provisions…
Mechanical fault	Material defect	Equipment failure, pile mechanism dynamic lock failure, device failure, Gantry crane failure, engineering equipment failure, crane failure…
Weak foundation	Environmental factors	Soft foundation, poor foundation stability, weak formation, lack of stable subgrade…
Chaotic site management	Technical and management factors	Temporary labor management is chaotic, the site management is ineffective, the site management structure is not sound, the site management is out of control, the construction management is loose…
Earth collapse	Risk events	Soil collapse, sediment collapse, rock collapse, slope collapse…

Table 16. Dictionary of subway construction safety risks (part).

Construction of domain dictionary

A domain dictionary for safety risks of subway construction was constructed based on the classification results. This dictionary encompasses four major classes of risk factors about subway construction and 56 types of risk events, totaling 1361 entity synonymous expressions. Due to the space limitation, a partial display of the domain dictionary is provided in the Table 16.

Results and applications

Subway construction safety risks complex network construction

In this section, 1159 causal relationship triplets concerning domain entities of safety risks about subway construction were normalized based on the domain dictionary constructed in "Construction of domain dictionary". This process got 533 causal relationship triplets in safety risks of subway construction and constructed a directed unweighted complex network, termed the Metro Construction Safety Risk Complex Network (MCSRCN). The MCSRCN model comprises 120 nodes and 533 directed arcs, encompassing all identified risk factors and event types of subway construction. The MCSRCN model is visualized by Pajek (Fig. 4). Human factors (HF) are represented by yellow nodes, green nodes denote defect factors (MD), environmental factors (EF) are labeled by red nodes, blue and pink nodes indicate the technical and management factors (TM) and risk events (RE), respectively.

The distribution of nodes in MCSRCN was generated based on the random node degree indicator (Fig. 4). The closer a node to the central position, the greater influence it is to the entire network. It is apparent that the cluster of technical and management factor nodes (blue nodes) and human factor nodes (yellow nodes) are relatively closer to the center position of the network. It can be attributed to the fact that human factors are the direct causes of most safety accidents of subway construction, while technical and management factors are frequently the crucial indirect causes of safety accidents. These two types of risk factors are extensively described in accident texts of subway construction. Conversely, the material defect nodes (green nodes) and environmental factor nodes (red nodes) are relatively farther from the center in the network. The possible reason is associated with the enhancement of quality management about subway construction. The construction materials with quality issues

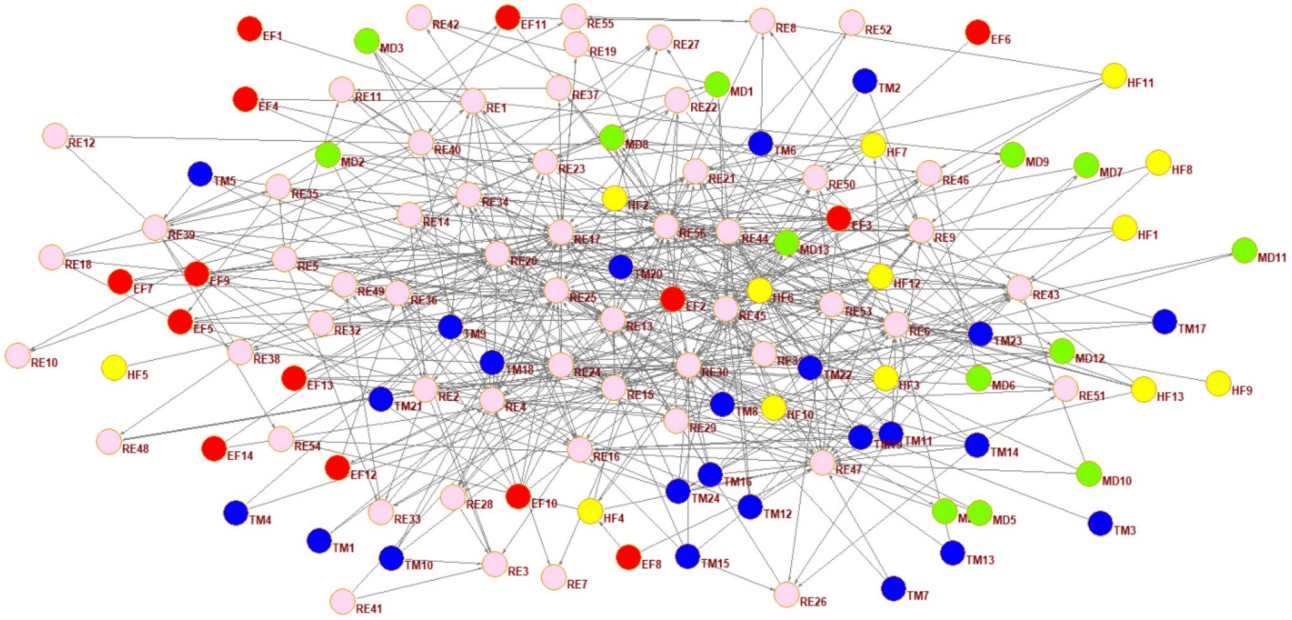

Fig. 4. MCSRCN model.

Database entry	Introduction and examples
Accident number	Record the type of the historical event, numbered RE1-RE56
Risk event type	The type of the accident corresponds to the accident number. For example, RE2 indicates a water inrush accident
Accident case number	The case of the accident type. For example, RE2-1 is the accident case 1 of the water inrush accident
Accident name	The name information of historical accidents. For example: RE2-2 is the water inrush accident of 2# inclined shaft in Tongluoshan Tunnel in Chongqing
Date	The date of occurrence of specific accident cases. For example, the date of RE2-2 accident: 20,120,719
Solution	The solution measures and countermeasures of specific accident cases
Warning education	Warning teaching information after the specific accident case solved, including the corresponding accident prevention information

Table 17. Database of subway construction about accident cases.

are often rejected before entering the site, thereby effectively mitigating these risk factors. Environmental factors are significantly influenced by regional conditions during subway construction. For example, "rain (EF2)" and "weak soil conditions (EF3)" have the significant impact on projects of subway construction in central and southern regions, whereas the impact on projects in northern regions is relatively smaller.

Construction of subway construction accident database

Based on 146 collected accident reports of subway construction, detailed information including project overview, accident types, accident resolution measures, and accident warnings were documented to establish a accident database about subway construction. The database is presented in Table 17.

During subway construction, project management personnel and construction workers can utilize the MCSRCN to identify potential risk events associated with the risk factors observed at the construction site. With the safety response measures of risk events and accident warning database, they can get the historical information regarding the specific accident type, thus learning and researching the resolution and educative warnings from the past accidents. Subsequently, the tailored measures for risk prevention and control with feasibility can be proposed based on the actual engineering circumstances. The establishment of risk database provides the basis for assisting decision-making in safety management. However, the on-site safety management is a complex process, which requires the integration of on-site monitoring data and various methods to ensure comprehensive control.

Recommendations for mitigating subway construction risks

With the database, recommendations for risk response were provided for the 2008 Xianghu Station foundation pit collapse accident in Hangzhou Metro Line 1. After entity recognition, the risk factors of this project include complex geological conditions (EF13), rainfall (EF2), rushing work (TM1), over-excavation of foundation pits (TM8), and support system defects (MD13). Based on historical experience and the complex risks network of subway construction, five risk factors were incorporated into a complex network to generate the risk network diagram specific to subway projects. The risk network diagram displayed nodes of varying sizes based

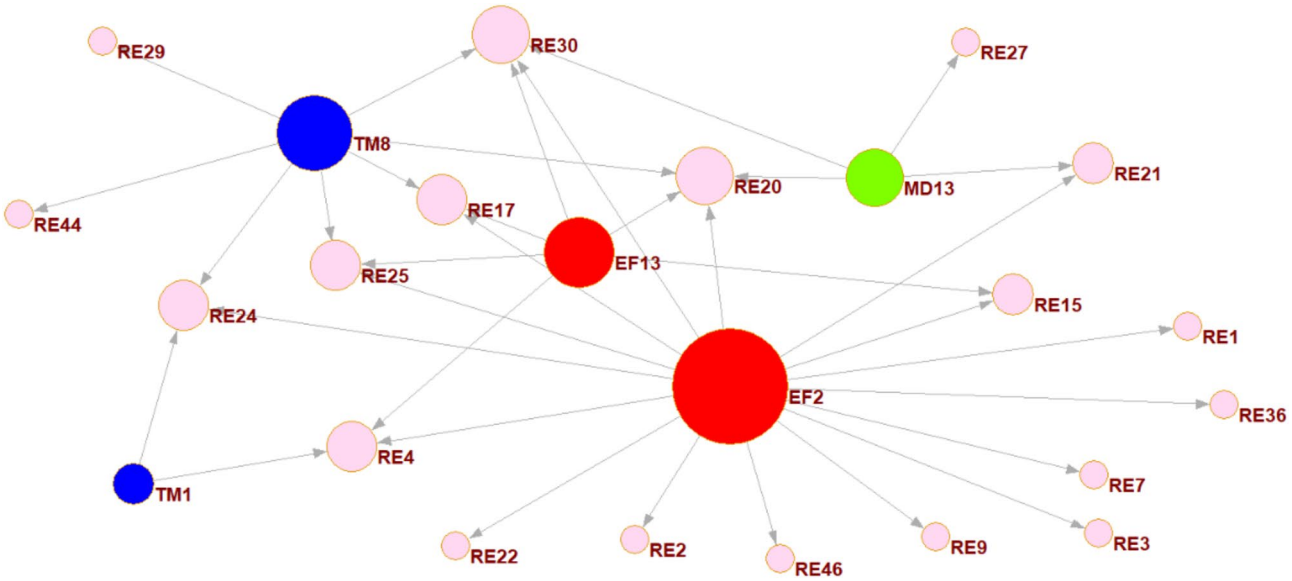

Fig. 5. The risks network diagram of the Hangzhou Metro project.

Identification of accident cases	Incident title
RE20-1	Collapse of shaft foundation pit for ventilation well 2# in phase 2 of Shanghai Mingzhu line
RE20-2	Collapse accident of Haizhu square metro station foundation pit in Guangzhou metro
RE20-3	Collapse accident of the panda roundabout foundation pit in line 10 of Beijing metro
RE20-4	Investigation Report on the Collapse Accident of the Foundation Pit at Xianghu Station on Line 1 of Hangzhou Metro on "11.15"
RE20-5	Investigation report on the collapse accident at section 5108 "7.29" of Chongqing rail transit line 5
RE20-6	Collapse accident in Shanghai metro foundation pit

Table 18. Foundation pit collapse accident cases.

on the node degree index, with larger radii corresponding to higher node degrees. Figure 5 indicated that rainfall (EF2) exhibited the highest node degree index, highlighting it as the risk factor deserving the most attention. By adopting strategies of risk response for rainfall, the impact of rainfall on construction sites was timely reduced and eliminated. The occurrence of eight risk events (RE1, RE2, RE3, RE7, RE9, RE22, RE36, RE46) was effectively prevented. In response to the risk of rainfall, strategies including optimizing drainage around foundation pits, reinforcing monitoring in regions with elevated groundwater levels and challenging geological conditions, applying plastic film on slopes to prevent erosion during heavy rainfall, and intensifying construction safety inspections could be undertaken. Among all potential risk events, there were four safety risk factors pointing towards foundation pit collapse (RE20) and earth-rock collapse (RE30). It can be anticipated that there might be some risk events related to foundation pit collapse and earth-rock collapse within these construction circumstances. Based on the prototype case, the eventual occurrence of the foundation pit collapse accident was indeed confirmed. The risk network of subway construction developed in this study holds certain practical guidance significance.

With collapse risk events as the specific cases, this study proposed risk response measures based on database analysis. Project managers can refer to the database to search for historical cases of foundation pit collapse accidents that occurred during subway projects (Table 18). Through the analysis of resolution measures in the cases database about subway accidents, preventive and contingency measures for foundation pit collapse incidents in the Hangzhou Metro project were extracted.

Conclusion

The safety risks of subway construction have the characters of complex, covert, and dynamic. Combining with the spatiotemporal features, nonlinear characteristics, and coupling effects about the continual accumulation of historical data, the comprehensive analyses exhibited significant challenges[1,46]. In order to solve the problem, this research focused on text mining and natural language processing in the safety risks of subway construction. The model MCSR-NER-Model, based on BiLSTM-CRF, was developed and trained specifically for entity recognition with Chinese in the safety risks of subway construction. The collection and analysis of textual data related to accidents of subway construction were conducted by this model. Firstly, entities and their types associated with the risk factors and risk events about subway construction were labeled by this model. Subsequently, the automatic extraction of entities from accident text of subway construction was completed.

The task of risk identification within the subway construction was achieved by this model with specific train. Additionally, the normalization of entities was achieved by constructing a dictionary for safety risks of subway construction. Ultimately, a total of 736 entity expressions related to risk factors of subway construction was compiled, including 147 human factor entities, 124 material defect factor entities, 107 environmental factor entities, and 358 technical and management factor entities. In addition, 625 risk event entities were obtained. Obtained 1361 domain entities basically covered the key vocabulary from 562 instances of accident texts about subway construction. This research refined the types of risk factors and risk events, thus provided the data support for the assessment and response research of safety risks about subway construction.

This research also constructed and trained the MCSR-CE-Model based on CNN within the safety risks of subway construction for extracting causal relationships among domain entities. 1159 causal relationship triplets among domain entities were obtained. These triplets essentially covered the interrelationships among risk factors, risk events and the mutual influences of subway construction within 562 instances of accident texts. This model clarified the transmission pathways of risk factors during subway construction, providing the better data support for risk response measures.

With the developed MCSR-NER-Model and MCSR-CE-Model, a chain-like structure of causal relationships was automatically transformed from free-text about subway construction accidents, which bridged the gap about the heavily relies on expert experience in conventional identification about safety risks. With the continuous collection and expansion of textual data, a safety risk database about subway construction were established, containing numerous expressions of risk factors and event entities in subway construction, as well as the causal relationships among these entities. This database can combine the extensive historical accident cases with the experience from domain experts to analyze and resolve management issues of safety risk during subway construction from a data-driven perspective.

The primary sources of textual data involve subway accident reports and subway accident notices. Because of the limitations of the text types for recording risk events, the text carries of accident information will be obtained from internal enterprise construction logs, accident hazard inspection forms, and records of attempted accidents in the future. Based on the model architecture, automatic extraction of more relationship types such as synonymy relationship, membership relationship and coupling relationship will be explored, and the interaction relationship between risk factors and risk event types of subway construction will be further mined. In terms of data mining samples, the current accident texts of subway construction used for model training are limited. Consequently, it is challenging to encompass all risk factors and risk event types along with their causalities about subway construction. This experiment segregated the entity recognition and causal relationship extraction into two independent processes, inevitably resulting in relationship extraction being reliant on the results of named entity recognition. As for relationship extraction, the textual data sources mainly involve reports and bulletins of subway accident. The limitations of text type for recording risk events still exist. In the future, more text carriers of accident information such as construction logs, accident hidden danger investigation tables, and attempted accident records inside enterprises can be obtained. Based on the model architecture, the automatic extraction of relationships such as synonymy, membership and coupling relationship should be explored. In addition, the further investigations about the mining of the interaction between risk factors and risk event types of subway safety are supposed to be conducted. The effect of time is crucial in risk research, and the time series will be introduced to simulate the evolution of risks at different construction stages in the future.

Data availability

Some or all data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 8 August 2024; Accepted: 21 April 2025
Published online: 11 May 2025

References

1. Zhou, Z. & Guo, W. Applications of item response theory to measuring the safety response competency of workers in subway construction projects. *Saf. Sci.* **127**, 104704. https://doi.org/10.1016/j.ssci.2020.104704 (2020).
2. Ye, Z. et al. A digital twin approach for tunnel construction safety early warning and management. *Comput. Ind.* **144**, 103783. https://doi.org/10.1016/j.compind.2022.103783 (2023).
3. Zhang, Y. Application of risk management plan to technical risks in metro construction: case study of the grand Paris express project. *Tunn. Undergr. Sp Tech.* **147**, 105716. https://doi.org/10.1016/j.tust.2024.105716 (2024).
4. Wang, X., Xia, N., Zhang, Z., Wu, C. & Liu, B. Human safety risks and their interactions in China's subways: stakeholder perspectives. *J. Manag. Eng.* **33** (5), 05017004. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000544 (2017).
5. Zhang, S. et al. Assessing safety risk management performance in Chinese subway construction projects: A multistakeholder perspective. *J. Manag. Eng.* **38** (4), 05022009. https://doi.org/10.1061/(ASCE)ME.1943-5479.000106 (2022).
6. Zhang, S. et al. Identifying critical factors influencing the safety of Chinese subway construction projects. *Eng. Constr. Archit. Manag.* **28** (7), 1863–1886. https://doi.org/10.1108/ECAM-07-2020-0525 (2021).
7. Liu, X., Shao, C., Ma, H. & Liu, R. Optimal Earth pressure balance control for shield tunneling based on LS-SVM and PSO. *Autom. Constr.* **20** (4), 321–327. https://doi.org/10.1016/j.autcon.2010.11.002 (2011).
8. Zheng, D., Huang, J., Li, D. Q., Kelly, R. & Sloan, S. W. Embankment prediction using testing data and monitored behaviour: A bayesian updating approach. *Comput. Geotech.* **93**, 150–162. https://doi.org/10.1016/j.compgeo.2017.05.003 (2018).
9. Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B. & Bowman, D. Automated content analysis for construction safety: A natural Language processing system to extract precursors and outcomes from unstructured injury reports. *Autom. Constr.* **62**, 45–56. https://doi.org/10.1016/j.autcon.2015.11.001 (2016).
10. Gu, S., Li, K., Feng, T., Yan, D. & Liu, Y. The prediction of potential risk path in railway traffic events. *Reliab. Eng. Syst. Saf.* **222**, 108409. https://doi.org/10.1016/j.ress.2022.108409 (2022).
11. Xu, N., Liu, M. A. L., Li, Q., Deng, Y. & W.A.N.G. and An improved text mining approach to extract safety risk factors from construction accident reports. *Saf. Sci.* **138**, 105216. https://doi.org/10.1016/j.ssci.2021.105216 (2021).

12. Shen, J., Liu, S. & Zhang, J. Using text mining and bayesian network to identify key risk factors for safety accidents in metro construction. *J. Constr. Eng. M.* **150** (6), 04024052. https://doi.org/10.1061/JCEMD4.COENG-14114 (2024).
13. Lin, S. S., Shen, S. L., Zhou, A. & Xu, Y. S. Risk assessment and management of excavation system based on fuzzy set theory and machine learning methods. *Autom. Constr.* **122**, 103490. https://doi.org/10.1016/j.autcon.2020.103490 (2021).
14. Liu, K., Zhu, J. & Wang, M. An event-based probabilistic model of disruption risk to urban metro networks. *Transp. Res. Part. Policy Pract.* **147**, 93–105. https://doi.org/10.1016/j.tra.2021.03.010 (2021).
15. Tang, B., Guo, S., Li, J. & Lu, W. Exploring the risk transmission characteristics among unsafe behaviors within urban railway construction accidents. *J. Constr. Eng. M.* **28** (6), 443–456. https://doi.org/10.3846/jcem.2022.16924 (2022).
16. Jin, R., Wang, F. & Liu, D. Dynamic probabilistic analysis of accidents in construction projects by combining precursor data and expert judgments. *Adv. Eng. Inf.* **44**, 101062. https://doi.org/10.1016/j.aei.2020.101062 (2020).
17. Wang, J., He, Z. & Weng, W. A review of the research into the relations between hazards in multi-hazard risk analysis. *Nat. Hazards (Dordr).* **104** (3), 2003–2026. https://doi.org/10.1007/s11069-020-04259-3 (2020).
18. Zhou, Z., Liu, S. & Qi, H. Mitigating subway construction collapse risk using bayesian network modeling. *Automat Constr.* **143**, 104541. https://doi.org/10.1016/j.autcon.2022.104541 (2022).
19. Zhou, Z., Irizarry, J. & Li, Q. Using network theory to explore the complexity of subway construction accident network (SCAN) for promoting safety management. *Saf. Sci.* **64**, 127–136. https://doi.org/10.1016/j.ssci.2013.11.029 (2014).
20. Zhuang, J. et al. Scenario simulation of the geohazard dynamic process of large-scale landslides: a case study of the Xiaomojiu landslide along the Jinsha river. *Nat. Hazards (Dordr).* **112** (2), 1337–1357. https://doi.org/10.1007/s11069-022-05229-7 (2022).
21. Fang, Q., Zhang, D. & Wong, L. N. Y. Environmental risk management for a cross interchange subway station construction in China. *Tunn. Undergr. Sp Tech.* **26** (6), 750–763. https://doi.org/10.1016/j.tust.2011.05.003 (2011).
22. Zhang, S. et al. Assessing safety risk management performance in Chinese subway construction projects: A multistakeholder perspective. *J. Manag. Eng.* **38** (4), 05022009. https://doi.org/10.1061/(ASCE)ME.1943-5479.0001062 (2022).
23. Shi, X. et al. Evaluation of risk factors affecting the safety of coal mine construction projects using an integrated DEMATEL-ISM approach. *Eng. Constr. Archit. Manag.* https://doi.org/10.1108/ECAM-02-2023-0103 (2024).
24. Li, M., Yu, H., Jin, H. & Liu, P. Methodologies of safety risk control for China's metro construction based on BIM. *Saf. Sci.* **110**, 418–426. https://doi.org/10.1016/j.ssci.2018.03.026 (2018).
25. Guo, K. & Zhang, L. Multi-source information fusion for safety risk assessment in underground tunnels. *Knowl. Based Syst.* **227**, 107210. https://doi.org/10.1016/j.knosys.2021.107210 (2021).
26. Pais, S., Cordeiro, J. & Jamil, M. L. NLP-based platform as a service: a brief review. *J. Big Data.* **9** (1), 54. https://doi.org/10.1186/s40537-022-00603-5 (2022).
27. Li, J., Sun, A. X., Han, J. L. & Li, C. L. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl.* **34** (1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314 (2022).
28. Tang, C., Shen, C., Zhang, J. & Guo, Z. Identification of safety risk factors in metro shield construction. *Buildings* **14** (2), 492. https://doi.org/10.3390/buildings14020492 (2024).
29. Huo, X., Du, S. & Jiao, L. Critical causal path analysis of subway construction safety accidents based on text mining. *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part. Civ. Eng.* **11** (1), 04024075. https://doi.org/10.1061/AJRUA6.RUENG-1284 (2025).
30. Li, M. & Wang, J. Intelligent recognition of safety risk in metro engineering construction based on BP neural network. *Math. Probl. Eng.* **2021**(1), 5587027. https://doi.org/10.1155/2021/5587027 (2021).
31. Nasar, Z., Jaffry, S. W. & Malik, M. K. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.* **54** (1), 1–39. https://doi.org/10.1145/3445965 (2021).
32. Zhou, Z. et al. Developing a deep reinforcement learning model for safety risk prediction at subway construction sites. *Reliab. Eng. Syst. Saf.* 110885. https://doi.org/10.1016/j.ress.2025.110885 (2025).
33. Liu, J. et al. A hybrid deep-learning approach for complex biochemical named entity recognition. *Knowl. Based Syst.* **221**, 106958. https://doi.org/10.1016/j.knosys.2021.106958 (2021).
34. Gaba, S. et al. A federated calibration scheme for convolutional neural networks: models, applications and challenges. *Comput. Commun.* **192**, 144–162. https://doi.org/10.1016/j.comcom.2022.05.035 (2022).
35. Wu, Z., Pan, S., Long, G., Jiang, J. & Zhang, C. Beyond low-pass filtering: graph convolutional networks with automatic filtering. *IEEE Trans. Knowl.* **35** (7), 6687–6697. https://doi.org/10.1109/TKDE.2022.3186016 (2022).
36. Yasin, M., Sarıgül, M. & Avci, M. Logarithmic learning differential convolutional neural network. *Neural Netw.* **172**, 106114. https://doi.org/10.1016/j.neunet.2024.106114 (2024).
37. Geng, Z., Zhang, Y. & Han, Y. Joint entity and relation extraction model based on rich semantics. *Neurocomputing.* **429**, 132–140. https://doi.org/10.1016/j.neucom.2020.12.037 (2021).
38. Deng, Y., Liu, Z., Song, L., Ni, G. & Xu, N. Exploring the metro construction accidents and causations for improving safety management based on data mining and network theory. *Eng. Constr. Archit. Manag.* https://doi.org/10.1108/ECAM-06-2022-0603 (2023).
39. Zhou, H., Tang, S., Huang, W. & Zhao, X. Generating risk response measures for subway construction by fusion of knowledge and deep learning. *Autom. Constr.* **152**, 104951. https://doi.org/10.1016/j.autcon.2023.104951 (2023).
40. Zhou, M., Tang, Y., Jin, H., Zhang, B. & Sang, X. A BIM-based identification and classification method of environmental risks in the design of Beijing subway. *J. Civ. Eng. Manag.* **27** (7), 500–514. https://doi.org/10.3846/jcem.2021.15602 (2021).
41. Zhou, Z., Goh, Y. M., Shi, Q., Qi, H. & Liu, S. Data-driven determination of collapse accident patterns for the mitigation of safety risks at metro construction sites. *Tunn. Undergr. Sp Tech.* **127**, 104616. https://doi.org/10.1016/j.tust.2022.104616 (2022).
42. Qi, H., Zhou, Z., Yuan, J., Li, N. & Zhou, J. Accident pattern recognition in subway construction for the provision of customized safety measures. *Tunn. Undergr. Sp Tech.* **137**, 105157. https://doi.org/10.1016/j.tust.2023.105157 (2023).
43. Ke, J. et al. Medical entity recognition and knowledge map relationship analysis of Chinese EMRs based on improved BiLSTM-CRF. *Comput. Electr. Eng.* **108**, 108709. https://doi.org/10.1016/j.compeleceng.2023.108709 (2023).
44. Xu, N. et al. Entity recognition in the field of coal mine construction safety based on a pre-training language model. *Eng. Constr. Archit. Manag.* https://doi.org/10.1108/ECAM-05-2023-0512 (2025).
45. Huang, M. S. et al. Biomedical named entity recognition and linking datasets: survey and our recent development. *Brief. Bioinform.* **21** (6), 2219–2238. https://doi.org/10.1093/bib/bbaa054 (2020).
46. Zhou, Z., Irizarry, J. & Zhou, J. Development of a database exclusively for subway construction accidents and corresponding analyses. *Tunn. Undergr. Sp Tech.* **111**, 103852. https://doi.org/10.1016/j.tust.2021.103852 (2021).

Author contributions

All authors contributed to the study conception and design. Writing, original draft preparation and methodology were performed by Y.L. Methodology, writing and review were performed by N.X. Validation and data curation were performed by H.C. Resources and data curation were performed by S.Q. Models calculation was performed by Y.L. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by National Social Science Fund of China (23BGL277).

Declarations

Competing interests
The authors declare no competing interests.

Ethical statement
The submitted work is original and it not published elsewhere in any form or language. It is not submitted to any other journal for simultaneous consideration.

Additional information
Correspondence and requests for materials should be addressed to N.X.

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