Ontology-based Text Convolution Neural Network (TextCNN) for Prediction of Construction Accidents

Author One[a], Author Two[b], Author Three[a,b]

[a]*Department One, Address One, City One, 00000, State One, Country One*
[b]*Department Two, Address Two, City Two, 22222, State Two, Country Two*

Abstract

The construction industry suffers from workplace accidents, including injuries and fatalities, which represent a significant economic and social burden for employers, workers, and society as a whole. The existing research on construction accidents heavily relies on expert evaluations, which often suffer from issues such as low efficiency, insufficient intelligence, and subjectivity. However, expert opinions provided in construction accident reports offer a valuable source of knowledge that can be extracted and utilized to enhance safety management. Today this valuable resource can be mined as the advent of artificial intelligence has opened up significant opportunities to advance construction site safety. Ontology represents an attractive representation scheme. Though ontology has been used in construction safety to solve the problem of information heterogeneity using formal conceptual specifications, the establishment and development of ontologies that utilize construction accident reports are currently in an early stage of development and require further improvements. Moreover, research on the exploration of incorporating deep learning methodologies into construction safety ontologies for predicting construction safety incidents is relatively limited. This paper describes a novel approach to improving the performance of accident prediction models by incorporating ontology into a deep learning model. First, a domain word discovery algorithm, based on mutual information and adjacency entropy, is used to analyze the causes of accidents mentioned in construction reports. This analysis is then combined with technical specifications and literature in the field of construction safety to build an ontology encompassing unsafe factors related to construction accidents. By employing a Translating on Hyperplane (TransH) model, the reports are transformed into conceptual vectors using the constructed ontology. Building on this foundation, we propose a

Text Convolutional Neural Network (TextCNN) model that incorporates the ontology specifically designed for construction accidents. We compared the performance of the TextCNN model against five traditional machine learning models, namely Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, and Multilayer Perceptron, using three different data sets: One-Hot encoding, word vector, and conceptual vectors. The results indicate that the TextCNN model integrated with the ontology, outperformed the other models in terms of performance achieving an impressive accuracy rate of 88% and AUC value of 0.92.

Keywords: Construction Safety, Deep Learning, Ontology, Text Mining, TransH, TextCNN

1. Introduction

The rapid development of the construction industry not only spurs economic growth but also promotes job creation. Construction safety though has remained a major concern. Factors that contribute to construction safety related problems include complex construction systems, harsh construction environments, high level of operational complexity, and high mobility of operators. Between 2010 and 2019, there were 6,005 fatal accidents in China's construction industry causing 7,275 deaths, averaging 1.2 deaths per accident. Furthermore, there were 258 major construction accidents causing 1,037 deaths, averaging 4.0 deaths per accident. This also accounted for 4.3% and 14.25% of the total number of construction accidents and deaths in this period, respectively (Xu and Xu, 2021). Therefore, research on construction safety is an important topic.

The use of artificial intelligence in improving construction safety has recently attracted significant attention as a research subject. A variety of systems have been introduced in this context, including social network platforms (Le et al., 2014), safety management systems (Park and Kim, 2013), and learning frameworks (Le et al., 2015). More recently, ontology technology, with its inherent capability to formulate a semantic model of domain-related data, has been used to address diverse challenges in the construction safety field. For example, there have been attempts to explore new ways of organizing, storing, and reusing construction safety knowledge (Zhang et al., 2015), crafting a meta-model for verifying construction safety (Lu et al., 2015), and establishing a conceptual information model for managing construction

safety (Farghaly et al., 2022).

Nevertheless, the incorporation of predictive models through the fusion of state-of-the-art machine learning and semantic techniques to address construction safety concerns remains relatively scarce in the literature. Such integration has demonstrated its efficacy in various other domains (Liu et al., 2015; Kastrati et al., 2019; Wu et al., 2019), yet its application in the context of safety has been limited.

This study attempts to fill this gap by using case information from construction safety accident investigation reports to extract and effectively use the domain knowledge and experience contained in these reports. To this end, we have constructed a comprehensive ontology tailored to construction accidents. Subsequently, we adopted the TransH model to convert the accident reports into a dataset comprised of ontology vectors. This transformed dataset was then subjected to analysis using the TextCNN deep learning model, facilitating the classification of distinct types of safety accidents. This novel approach, which combines ontology and TextCNN deep learning model, aims to improve the performance of prediction models for construction accidents. The results indicate a TextCNN model integrated with the ontology outperformed the other machine learning models in terms of performance achieving an impressive accuracy rate of 88% and AUC value of 0.92.

In summary, the main contributions of this study are as follows:

1. We used a new word discovery algorithm, based on mutual information and adjacency entropy, to identify unsafe factors present in the safety accident reports. Furthermore, the correlation between these unsafe factors was analyzed using a linear correlation algorithm and clustering techniques using word vectors.
2. We conducted a comprehensive analysis of the relationship between unsafe factors, the correlation between unsafe factors and accident types, and references in the existing literature. Based on this analysis, we developed an ontology specifically focused on unsafe factors associated with construction accidents.
3. We performed a comparative analysis of machine learning models, namely Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest, Multilayer Perceptron, and TextCNN. The evaluation was conducted using three different datasets with One-Hot encoding, word vectors, and ontology.

The rest of the paper is structured as follows. In Section 2, we provide

the research background, followed by the related work. In Section 3, we present the methodology adopted to design an ontology, as well as the details regarding a computational model. In Section 4, we outline the process of data processing, in particular a resulting ontology. In Section 5, we demonstrate the results obtained from the experiments performed, followed by a discussion given in Section 6. In Section 7, we conclude the paper.

2. Background and Related Work

In this section we discuss the application of ontology across multiple domains, ontology-based risk prediction models in engineering, and ontology-based risk prediction models in construction safety management. We also summarize the literature, identify gaps in research on using ontology in construction safety, and then present our novel approach.

2.1. Application of ontology across multiple domains

Ontology, as a semantic technique, serves as a valuable tool for representing domain knowledge through formalized descriptions. By identifying core concepts, attributes, and relationships among these concepts, ontology facilitates knowledge exchange and sharing. This technique allows for the conversion of textual knowledge about a specific domain into a structured and understandable format, which enhances its usability and applicability (Studer et al., 1998).

Many important problems have been studied and solved using ontology models. For example, the application of ontology can enhance a language's ability to describe fuzzy information. Jiang and Tan (2009) proposed a semantic user ontology model to support personalized applications in the Semantic Web, enabling the learning and description of different users' needs. Pulido et al. (2006) introduced a method of using self-organizing mapping (SOM) to identify ontology components, facilitating the easier dynamic and intelligent operation of the semantic web.

In addition, ontology plays a crucial role in document preprocessing and improving the accuracy of search results. Luo (2009) introduced iMed, the first intelligent medical network search engine, which incorporates medical knowledge and questionnaires to ordinary users in searching for medical information. Similarly, Park et al. (2013) utilized ontology for building information queries. Zhao and Ichise (2014) developed a framework that integrates ontology and provides semi-automatic query methods related to on-

tology, thereby reducing ontology heterogeneity. Furthermore, Bashar and Li (2018) proposed a novel model that effectively discovers and interprets hidden meanings in text patterns within a collection of documents. By leveraging ontology concepts, the model provides pattern meanings, as well as generates and extracts features to describe relevant information.

Furthermore, ontology technology finds extensive applications in various disciplines and fields, including agriculture, education, environment, military, and e-commerce. It is utilized to establish domain-specific ontologies, facilitating the sharing and integration of professional knowledge, data, and information (Oberle, 2014).

2.2. Ontology-based risk prediction models in engineering

Ontology-based risk identification methods have found application in various practical engineering scenarios to enhance risk identification, assessment, and prevention as well as management processes. For example, Xing et al. (2019) applied ontology in a metro construction to improve effective knowledge sharing and reuse. The ontology categorizes safety risk knowledge into seven cohesive classes, including project, construction activity, risk factor, risk, risk grade, risk consequence, and risk prevention measures. These classes, alongside their respective properties and relationships, are structured using the Protégé platform. The effectiveness of their approach is assessed through theoretical and practical evaluations, demonstrating its suitability for knowledge sharing and reuse. The ontology is anticipated to have broad applicability in the realm of safety risk identification within metro construction.

Aziz et al. (2019) used the Semantic web-based Web Ontology Language (OWL) to capture knowledge about unwanted process events. The resulting knowledge model was then transformed into Probabilistic-OWL (PR-OWL) based Multi-Entity Bayesian Network (MEBN). Upon queries, the MEBNs produced Situation Specific Bayesian Networks (SSBN) to identify hazards and their pathways along with probabilities. Two open-source software programs, Protégé and UnBBayes, were used. The developed model was validated against 45 industrial accidental events extracted from the U.S. Chemical Safety Board's (CSB) database.

Wu et al. (2020) proposed an ontological method that integrates case-based reasoning (CBR), natural language processing (NLP) technologies, and rule-based reasoning (RBR) for subway accident case retrieval to support safety management decisions. By constructing ontology analysis, based on

railway domain knowledge and accident analysis reports. The ontology was effectively applied to formalize safety risk knowledge in subway construction.

Jiang et al. (2020) introduced an enhanced decision-making approach for construction safety risk management, combining ontology and improved case-based reasoning (CBR). This approach incorporated both a similarity algorithm and a correlation algorithm, elevating the reasoning process. Unlike conventional CBR, which solely relies on information similarity, this method ensures the inclusion of vital correlated data from multiple sources. The researchers subsequently applied this method to assess safety risks in subway construction.

Cao et al. (2020) developed an innovative integrated evolutionary model known as the Scenario-Risk-Accident Chain Ontology (SRAC). This model served to convert risk source data from the implementation process into actionable knowledge, empowering effective decision-making in the realm of risk assessment. Its dynamic capabilities enabled swift responses to challenging scenarios, particularly in managing grid infrastructure during severe hazards. To evaluate the risk source level, the authors used a long short-term memory (LSTM) neural network model and other neural network architectures to capture context and risk text features.

Additionally, Phengsuwan et al. (2022) proposed a data integration and analytics system that enables an interaction between natural hazard early warning system (EWS) and electrical grid EWS to contribute to electrical grid network monitoring and support decision-making for electrical grid infrastructure management. We prototype the system using landslides as an example natural hazard for the grid infrastructure monitoring. Essentially, the system consists of background knowledge about landslides as well as information about data sources to facilitate the process of data integration and analysis. Using the knowledge modeled, the prototype system can report the occurrence of landslides and suggest potential data sources for the electrical grid network monitoring.

Hai et al. (2021) introduced ontology technology and knowledge base construction into the risk management of underground integrated pipeline corridor, built an ontology-based knowledge base of integrated pipeline corridor risk, and constructed a Bayesian network based on the established risk knowledge base for risk evaluation of identified risk factors.

Macêdo et al. (2022) used text mining and fine-tuned trained bidirectional encoder representations from transformers (BERT) models to support and reduce the efforts required for completing the early stages of quantitative

risk analysis. They identified the potential consequences of accidents related to the operation of an oil refinery and classified each scenario in terms of the severity of the consequence and likelihood of occurrence. The potential consequences, severity, and likelihood categories were predicted with a mean accuracy of 97.42%, 86.44%, and 94.34% respectively.

Taking methanol production as an example, the concept of "ontology" is introduced to construct a safety knowledge ontology, and a safety information knowledge base is created with the help of the Protégé software. These can be used to efficiently handle the massive safety information data of dangerous chemical enterprises, associate all kinds of miscellaneous information, and improve the level of safety management. An accident tree reasoning model is designed to determine the cause of the accident using accident tree reasoning, and to mine the vast knowledge of safety information, according to safety information knowledge and accident tree analysis theory. Using these methods, the storage, processing, and reuse of safety information are realized, the efficiency of safety management can be improved, and the defects caused by incomplete personnel knowledge structure can be avoided (Liu et al., 2023).

Risk identification is a knowledge-based process that requires the time-consuming and laborious identification of project-specific risk factors. Current practices for risk identification in construction rely heavily on an expert's subjective knowledge of the current project and of similar historical projects to determine if a risk may affect the project under study. When quantitative risk-related data are available, they are often stored across multiple sources and in different types of documents complicating data sharing and reuse. The study by Mohamed et al. (2023) introduces an ontology-based approach for construction risk identification that maps and automates the representation of project context and risk information, thereby enhancing the storage, sharing, and reuse of knowledge for the purpose of risk identification. The study also presents a novel wind farm construction project risk ontology that has been validated by a group of industry experts. The resulting ontology-based risk identification approach is able to accommodate project context in the risk identification process and, through implementation of the proposed approach, has identified risk factors that affect the construction of onshore wind farm projects.

2.3. Ontology-based risk prediction models in construction safety management

In the construction industry, ontology-based risk identification methods have been employed to improve the understanding and identification of risks associated with construction projects. These methods involve creating ontologies that define different aspects of construction projects, such as materials, equipment, tasks, and environmental factors. In this subsection, we provide a comprehensive summary of the relevant research which focuses on the application of ontology in the domain of construction industry and safety management, with a particular emphasis on the research methods used. We then identify research gaps and highlight the innovations and contributions of this study.

Based on the survey of 342 safety managers, Wang et al. (2006) developed a risk assessment model designed to evaluate the hazard factors associated with the construction process. They further integrated this model with the project schedule tasks, enabling safety managers to prioritize their focus on tasks with high anticipated accident costs. This integrated approach effectively reduced the likelihood of construction accidents, thereby enhancing overall safety within the construction project.

Using the 4D CAD visual technology, Benjaoran and Bhokha (2010) proposed an integrated rule-based system for safety and construction management. This system incorporated rules that encompassed safety measures and requirements, along with design information related to building components, planning information regarding project activities, and control phases. The aim was to identify and address on-site hazards while also providing recommendations for safety measures.

Carbonari et al. (2011) used sensor technology to monitor construction safety management, aiming to enhance the intelligence and effectiveness of safety management practices. Their approach utilized sensors to gather real-time data on safety-related parameters, allowing for proactive identification and mitigation of potential safety hazards. Additionally, Cheng et al. (2012) surveyed 232 participants to assess the importance levels of 15 popular safety management practices (SMPs) and 5 project performance criteria. Through exploratory factor analysis, they found that the safety management process was perceived by the construction practitioners as the most crucial factor for reducing safety risks. Furthermore, they concluded that safety management information and safety management committees had a positive impact on overall project performance.

Aneziris et al. (2012) established an occupational risk assessment model and ranked the high-risk factors. Pinto (2014) built a fuzzy random-access memory model to reduce security risks by controlling security factors. Wachter and Yorio (2014) investigated the impact of emotions on construction accidents and highlighted the significant role emotions play in safety accidents. They recommended the inclusion of emotional factors in construction safety management to reduce the occurrence of accidents, emphasizing the need for a humanized management system. Leu and Chang (2013) combined fault tree transformation with a Bayesian network model to create a safety evaluation model. By analyzing risks, they derived safety management strategies aimed at reducing the probability of safety accidents at construction sites.

Zhang et al. (2014) undertook a thorough assessment and analysis of the implementation of ontologies within the construction industry. The authors concluded that the utilization of ontologies in the construction industry is still in its developmental phase and has not yet reached a state of full maturity. Wang and Yu (2014) as well as Chi et al. (2014) employed a domain-specific ontology for construction safety, specifically focusing on unsafe scenarios. Their objective was to decrease the amount of human effort needed in job hazard analysis (JHA) and enhance the range of potential solutions.

Zhang et al. (2015) employed ontology-based safety investigation techniques to develop an automated approach for safety planning in JHA. They utilized building information modeling (BIM) to create site models and define security settings as part of the process. Wu et al. (2021) presented a theoretical model that combines computer vision deep learning algorithms with a formal ontology to manage construction safety, specifically addressing hazard mitigation and inference. Computer vision techniques were employed to extract visual information from on-site construction images, while safety regulatory domain knowledge was represented by an ontology model and semantic web rule language rules.

Chen and Bria (2022) provided a review of the current trends in ontology-based safety management in construction. They also identified the gaps and potential opportunities for future research. The research primarily examines the life cycle of ontologies, including their development, integration, and application. The findings revealed a growing emphasis on ontology-based research during the maintenance phase of safety management. To optimize safety management in the future, there is a need for larger-scale case studies and increased automation within ontology-based systems. Pedro et al. (2022) proposed a new information sharing system that utilizes linked data,

ontologies, and knowledge graph technologies. This system is designed to semantically model safety information and formalize knowledge regarding construction accident cases.

2.4. Research Gaps and Contributions of This Study

In summary, the studies described above relied on existing safety measures and regulations, derived from experts or manuals, to mitigate the risk of accidents. However, a common characteristic among these studies is the limited utilization of expert knowledge contained in the existing corpus of construction accident reports and sporadic use of the-state-of-the-art artificial intelligence and soft computing methods, including deep learning methods in risk identification or prediction.

Ontologies have the potential to solve the problem of information heterogeneity in sharing and reusing information by means of developing precise formal conceptual specifications. The establishment and development of ontologies utilizing construction safety reports is currently at an immature stage and necessitates further enhancement. Furthermore, the incorporation of deep learning methodologies into construction safety ontologies for the prediction of safety incidents has been comparatively underexplored in research.

This study uses the case information in the construction safety accident investigation reports to extract and make full use of the domain knowledge and experience in the reports to improve the reuse of domain knowledge. We propose to apply ontology technology to extract the unsafe factors in construction safety accident reports and construct the prediction model using the deep learning model TextCNN to more efficiently conduct risk assessments.

We overcome a limitation on current methods of using a TextCNN model to create an ontology of factors within safety accident reports. By employing a Translating on Hyperplane (TransH) model, the accident reports are transformed into conceptual vectors using the constructed ontology. Since conceptual vector expressions have been effectively trained, the embedding layer is eliminated and followed directly by the convolution layer. Further, we test the effectiveness of the proposed method via a comparative analysis of machine learning models namely Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest and Multilayer Perceptron, and TextCNN.

3. Methodology

3.1. Accident causing theory

To improve the prediction of construction accidents, it is crucial to understand the underlying mechanism behind such accidents. Theories of accident causation are developed by studying and analyzing the causes of significant historical accidents, and then refining and summarizing them. These theories provide an explanation for the accident process, offering a starting point for analyzing the causes of accidents and providing theoretical support for accident prevention and prediction. One notable model within the theories of accident causation is the Trajectory Intersecting Model (TIM), which considers accidents as the outcome of the intersection of human and physical factors (Ge et al., 2022). This perspective posits that events with interactions contribute to the occurrence of accidents following a specific sequence of causality. In a safe state, the two trajectories of human and physical factors do not intersect with each other. However, under the combined influence of various unsafe factors, people and objects are brought into an unsafe state, resulting in the convergence and the two trajectories and leading to an accident. Based on this theory, we analyzed the relevant literature to identify and summarize the unsafe factors. We then determined the presence of these unsafe factors in our data, explored their interrelationships, and established an ontology to represent and organize this knowledge.

3.2. Natural Language Processing Technology

Natural language technologies encompass various components such as word segmentation, lexical analysis, syntactic analysis, and semantic analysis. In our approach, we begin by utilizing a domain information discovery algorithm based on mutual information and adjacency entropy in natural language processing. This algorithm aids in mining unsafe factors and domain-specific terms from construction safety accident reports. Subsequently, we employ Chinese segmentation techniques to obtain word vectors from the reports, enabling us to represent the textual data in a structured format. Finally, these word vectors are used as input variables for machine learning models, facilitating the prediction and analysis of construction accidents.

In natural language processing tasks, the new word discovery algorithm is often used to automatically identify vocabulary in a specific field. The most commonly used new word discovery algorithm is based on mutual information and adjacency entropy. Initially, candidate words are generated

based on their mutual information. Subsequently, these candidate words undergo filtering using adjacency entropy, which is a measure borrowed from information theory. In the current study, this method is used to discover domain-specific words within the section of accident reports that provides causal analysis. The underlying principles of the new word discovery algorithm, specifically the neighborhood word discovery algorithm based on mutual information and left and right entropy, are outlined as follows.

The Pointwise Mutual Information (PMI) is defined as in equation 1:

$$PMI\,(x,y) = log_2 \frac{p\,(x,y)}{p\,(x)\,p\,(y)} \tag{1}$$

where $p(x,y)$ represents the probability of two words, x and y, appearing simultaneously, whereas $p(x)$ and $p(y)$ are the probabilities of a single word appearing.

One can measure the degree of freedom of preselected words using the adjacency entropy (Huang and Powers, 2003). The greater the left and right adjacency entropy, the more flexible the combination of the word and the left and right words, then the word is likely to be a single word. The left and right entropy is defined as (taking left entropy as an example) as shown in equation 2:

$$H_L(w) = -\sum_{w_l \in s_l} p(w_l|w) log p(w_l|w) \tag{2}$$

where s_l is the set of left adjacent words of the candidate word w, and the candidate word $p(w_l|w)$ represents the probability that w_l is the candidate word w.

3.3. Ontology construction methods

There are various existing methods for constructing ontologies, each with its own approach and methodology. Some of these methods include: the Skeleton method, TOVE method (Toronto Virtual Enterprise) , IDEF5 method, Methontology, KACTUS, SENSUS, and seven-step method, which is the most widely used method (Fu et al.). Below, we provide a brief description for each of the methods.

The Skeletal Methodology (Skeleton), also known as the Enterprise Methodology, is specifically designed for creating enterprise ontologies, which are collections of defined terms related to business entities. The basic process

includes identifying the purpose and scope of the ontology application as well as constructing and evaluating the ontology (Noy et al., 2000).

The TOVE (Toronto Virtual Enterprise) quality enterprise ontology serves as a structured portrayal (utilizing first-order logic) of concepts, connections, and fundamental principles related to quality that possess a generality surpassing any particular quality domain (Gruninger, 1995; Kim et al., 1995).

IDEF5, known as the Integrated Definition for Ontology Description Capture Method, represents a software engineering approach aimed at creating and upkeeping functional and precise domain ontologies. Within the IDEF5 approach, an ontology is developed by capturing the substance of specific statements concerning real-world entities, their attributes, and how they are interconnected, and then expressing this substance in an instinctive and organic manner (Peraketh et al., 1994; Ye et al., 2008).

Methontology is a method for systematically designing high-quality ontologies (formal representations of domain knowledge). It emphasizes clear phases, reusability, formalization using logic, and rigorous verification to create ontologies that effectively capture concepts and relationships in various domains for applications in fields like information systems and artificial intelligence (Fernández-López et al., 1997).

KACTUS stands as a European ESPRIT initiative with the goal of establishing a systematic process for repurposing knowledge related to technical systems throughout their entire lifecycle. This approach enables the utilization of a consistent knowledge repository for tasks such as design, diagnosis, operation, maintenance, redesign, and instruction (Uschold and Gruninger, 1996; Eardley et al., 2016).

The SENSUS ontology contains a vast collection of 70,000 domain-independent concepts. In order to establish connections, representative concepts from the specific domain are meticulously chosen and linked to the SENSUS ontology through manual efforts. Subsequently, all concepts situated directly along the path from the specific terms to the root are incorporated into the ontology (Knight and Luk, 1994; Knight et al., 1995; Swartout et al., 1996).

The seven-step method is similar to the software development life cycle model which applies these steps iteratively and in a cyclical manner. The basic process for the construction of domain ontologies is as follows: determine the professional domain and scope of the ontology; assess the possibility of reusing existing ontologies; enumerate the significant terms of the ontology; define classes and their hierarchical relationships; define properties of the classes; define facets of the properties; and create instances (Kamel et al.,

2007).

The seven-step method is based on the ontology construction tool Protégé (Musen, 2015), which was also used in this study for ontology construction.

3.4. Translating Embeddings and Translating On Hyperplane Algorithms

Knowledge representation involves expressing entities and relationships in a knowledge base as low-dimensional vectors. The core idea of a TransE (Translating Embeddings) algorithm is that for a triple $(h, r, t) \in \triangle$ ($\triangle$ indicating the correct triplet set, $\triangle'$ indicating the incorrect triplet set), each triplet can be expressed as (h, r, t), where h is the vector representation of the head entity; r is the vector representation of the relationship, and t is the vector representation of the tail entity. There is $h + r = t$. Two conclusions can be drawn from the TransE model:

1. If $(h, r, t) \in \triangle$ and $(t, r, h) \in \triangle$, the relationship r is a reflexive mapping, then $r = 0$ and $h = 0$;
2. If $\forall i = \{0, 1, 2, \ldots, m\}$, $(h_i, r, t) \in \triangle_i$, r is a $m - 1$ mapping, $h_0 = \cdots = h_m$. Similarly, $\forall i = \{0, 1, 2, \ldots, m\}$, $(h, r, t_i) \in \triangle_i$, r is a $m - 1$ mapping, $t_0 = \cdots = t_m$.

Based on the analysis provided, it is evident that the TransE algorithm treats different entities as having the same value when handling reflexive relationships, one-to-many, many-to-one, and many-to-many relationships. This occurs because the model algorithm does not differentiate between entities based on their relationships. In other words, entities with different relationships are assigned the same values in the TransE algorithm.

The TransE model algorithm is not without shortcomings. Wang et al. (2014) proposed the TransH (Translating On Hyperplane) model, which defines a hyperplane W_r and a relationship vector d_r for each relationship $h_\perp$ $t_\perp$, respectively represent the projection of h, t on the hyperplane W_r, and the correct triplet must satisfy $h_r + d_r = t_r$. Such a representation method can make the entity vectors different in different relationships and make the entity vectors the same in the same relationship.

3.5. Text-CNN

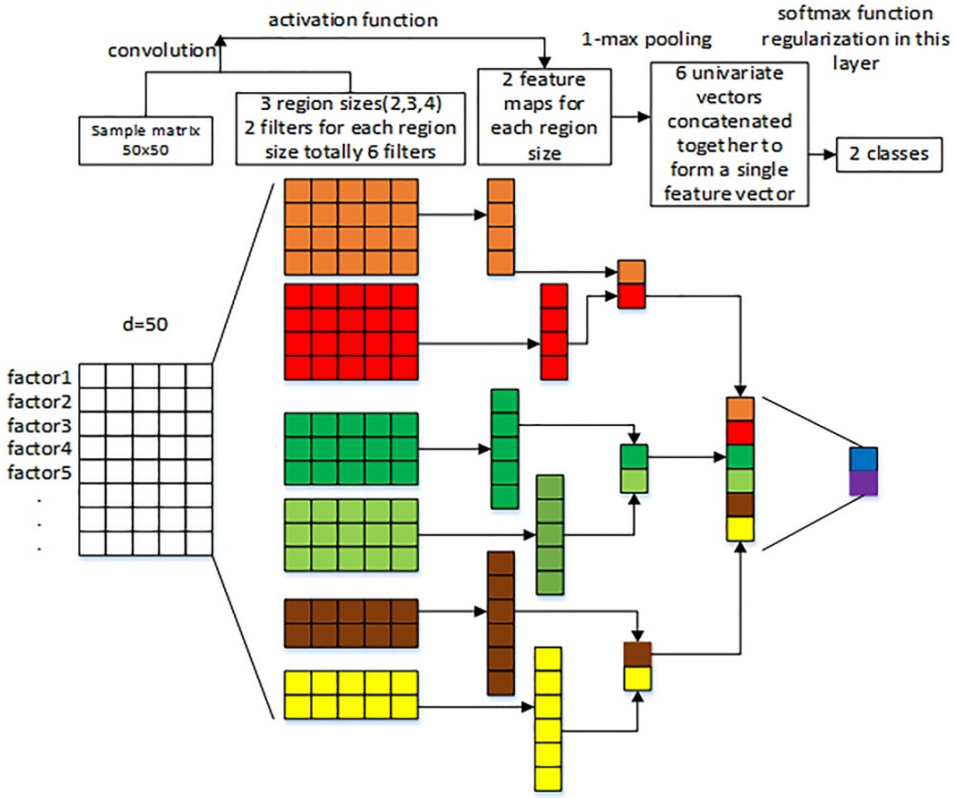
Chen (2015) introduced the widely recognized Text-CNN model for text classification, which incorporates convolutional neural networks into the task of text classification. The architecture of the TextCNN model is shown in Figure 1. The model comprises four main components:

1. The input layer, which includes the original model and an embedding layer.
2. The convolutional layer, responsible for applying operations to capture local patterns in the input text.
3. The pooling layer, which performs pooling operations to extract the most relevant features from the convolutional outputs.
4. The fully connected layer, which connects the extracted features to the output layer for classification.

The TextCNN model, specifically the TextCNN model mentioned in this paper, is commonly used for text classification tasks. However, it is worth noting that TextCNN is limited in its ability to capture the sequential order and positional information of words within the text. This limitation can diminish the effectiveness of the model for certain text classification tasks.

However, this shortcoming of the TextCNN model does not affect our study. In our study, the input variables consist of a set of unsafe factors corresponding to each sample, where the order of the factors is considered irrelevant information. Therefore, the original TextCNN model's inability to capture word order does not impact the effectiveness of our approach. In our study, three datasets are used: one with ontology using unsafe factor vectors as input variables for classification, another with One-Hot encoding transforming categorical variables into binary variables, and a third with word vectors as input variables. For the datasets involving word vectors and unsafe factor vectors, the embedding layer is removed since the word vectors and unsafe factor vectors have already undergone suitable training. The remaining model structures follow the original TextCNN model structure. To prevent overfitting, we employ two main methods: the 10-fold cross-validation method on the data and the combination of DropOut with Early Stopping.

Figure 1: The Structure of TextCNN Model



3.6. The prediction model for construction accidents

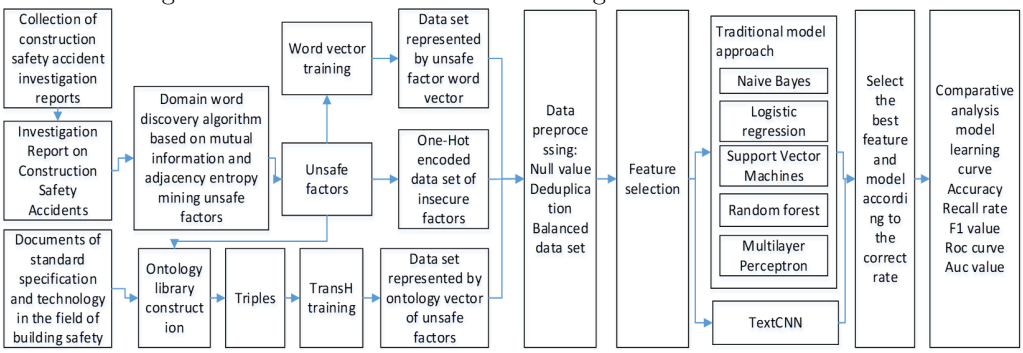
The framework for the prediction of construction accidents, as illustrated in Figure 2, involves several steps.

1. Data Collection. A search engine and Scrapy crawler are used to gather construction accident reports from the website of MHURD (Ministry of Housing and Urban-Rural Development of the People's Republic of China).
2. Domain Word Discovery. The accident cause section of the construction safety accident reports is analyzed using a domain word discovery algorithm based on mutual information and adjacency entropy. This algorithm extracts unsafe factors relevant to the accidents.
3. Knowledge Extraction. The extracted unsafe factors are used to extract and utilize domain knowledge and experience from the accident reports.

Subsequently, the core concepts of the construction safety field are obtained using the domain word discovery algorithm applied to the standard specifications and technical documents within the construction safety domain. Finally, the relationships between unsafe factors and their corresponding types are established through correlation analysis. By conducting this analysis on the unsafe factors and types, we combined standard specifications and technical documents in the field of constructor safety to construct an ontology specifically dedicated for unsafe factors.

The TransH model is utilized to generate knowledge vectors from the ontology represented by the triples. In order to incorporate the concepts derived from construction safety accident reports, we create a dataset with ontology for unsafe factors. To compare its performance, we compare the method using conceptual vectors on different datasets. Keywords are extracted from the construction safety accident reports to generate three data sets: One-Hot encoding, word vectors, and conceptual vectors. After performing data preprocessing and feature selection on these datasets, we proceed to compare the TextCNN model with five traditional models individually for each of the three datasets. Please refer to Figure, 2 for a more detailed illustration of the process.

Figure 2: The Framework for Predicting Construction Accidents



4. Data Set

4.1. Data collection and preprocessing for construction safety accident reports

The construction safety accident reports generally include:

- Basic Information.

- Accident occurrence and accident rescue situation,
- Casualties and direct economic losses,
- Cause and nature of the accident,
- Suggestions for handling of accident-related personnel and units, and
- Accident prevention and rectification measures.

For a detailed illustration of the contents of the construction safety accident reports, refer to Listing 1.

Listing 1: An Example Report of a Construction Safety Accident

```
xx month xx day xx hour xx, xxxx year, x city x phase III
   construction site...after the accident,...
1. Basic Information
(1) Project overview
The xxx project is a residential construction project, a total
   of 6 residential buildings...
(2) Construction unit
(3) Machinery leasing unit
(4) Supervision unit
(5) Accident tower crane situation
(6) Relevant personnel holding certificates
2. Accident occurrence and accident rescue situation
(1) The accident happened
  Xxxx year xx month xx day xx hour xx minute, ...
(2) Emergency rescue situation
   After the accident, ... organize the rescue work of the
      accident.
(3) Opinions on emergency assessment
3. Accident casualties and direct economic losses
(1) The situation of the deceased
   1.xxx, male, 31 years old, home address, xxx. xxx county
      forensics...
(2) The wounded
  Xxx, male, 45 years old, home address, xxx, currently
     hospitalized in the first hospital of xxx Medical
     University...
(3) Direct economic loss of the accident
   The direct economic loss of the accident was about RMB 3.5
      million.
4. The cause and nature of the accident
(1) Direct cause
```

```
  The demolition personnel are removing the connecting bolts of
     the standard section of the tower body......the tower top
     part is unbalanced and falls to the ground.
(2) Indirect reasons
   1.xxx Equipment Installation Co., Ltd., in violation of the
      Safety Supervision and Management Regulations for
      Construction Cranes,...
(3) Nature of the accident
  According to the investigation team, the nature of this
     accident is a production safety responsibility accident,
     ..... the accident level is a major accident.
5. Suggestions on the handling of accident-related personnel and
    units
(1) Suggestions for handling the responsible personnel
```

The keywords related to accidents, such as accident time, accident area, and accident type were used to conduct searches through a search engine. To ensure the reliability of the accident reports, only reports from government websites were selected. These reports primarily included information disclosure, accident announcements, or accident investigation sections on the government websites of the cities, districts, and counties where the accidents occurred, as well as on the websites of Work Safety Supervision Administration or Emergency Management departments. The Scrapy crawler was used to collect construction safety accident reports. Initially, the URLs of the relevant web pages were obtained and stored for further processing.

Next, we extracted the content from the web pages and then checked whether it contained relevant information such as accident time, accident type, construction units involved, accident causes, nature of accidents, and keywords like "Handling suggestions". This allowed us to determine whether the web page was a construction safety accident report. The collected data from the web pages contained a considerable amount of noise, so it was necessary to remove irrelevant and useless information in order to extract the relevant content from these reports. We utilized regular expressions to extract accident time, accident location, accident type, and construction unit information from the construction safety accident investigation reports. To avoid duplicate reports, we named the files using the accident time, location, type, and construction unit as unique identifiers for each accident report. In the end, the dataset consisted of 280 construction safety accident reports.

4.2. Unsafe factors in construction accidents based on mutual information and adjacency entropy

The paragraph in the construction safety accident report provides information about the accident causes identified by experts. These causes are categorized into direct causes and indirect causes. Direct causes refer to factors that directly contribute to the accidents and include both human factors and physical factors. Examples of direct causes include workers' illegal operations, physical discomfort, and equipment that does not meet safety standards. On the other hand, indirect causes are primarily related to management factors. These can include incorrect commands from managers or evidence of insufficient safety training. In this study, we employed the domain word discovery algorithm based on mutual information and adjacency entropy to identify domain-specific terms in the cause analysis section of the construction safety accident reports. This allowed us to extract relevant cause-related information. Through this algorithm, we were able to extract numerous unsafe factors such as: "paralysis", "tall formwork support system", "foggy weather", "not wearing protective equipment", "protective fence", "steel canopy", and "weak awareness." Following manual screening, we obtained a total of 538 keywords representing unsafe factors.

4.3. Correlation analysis of unsafe factors in construction accidents

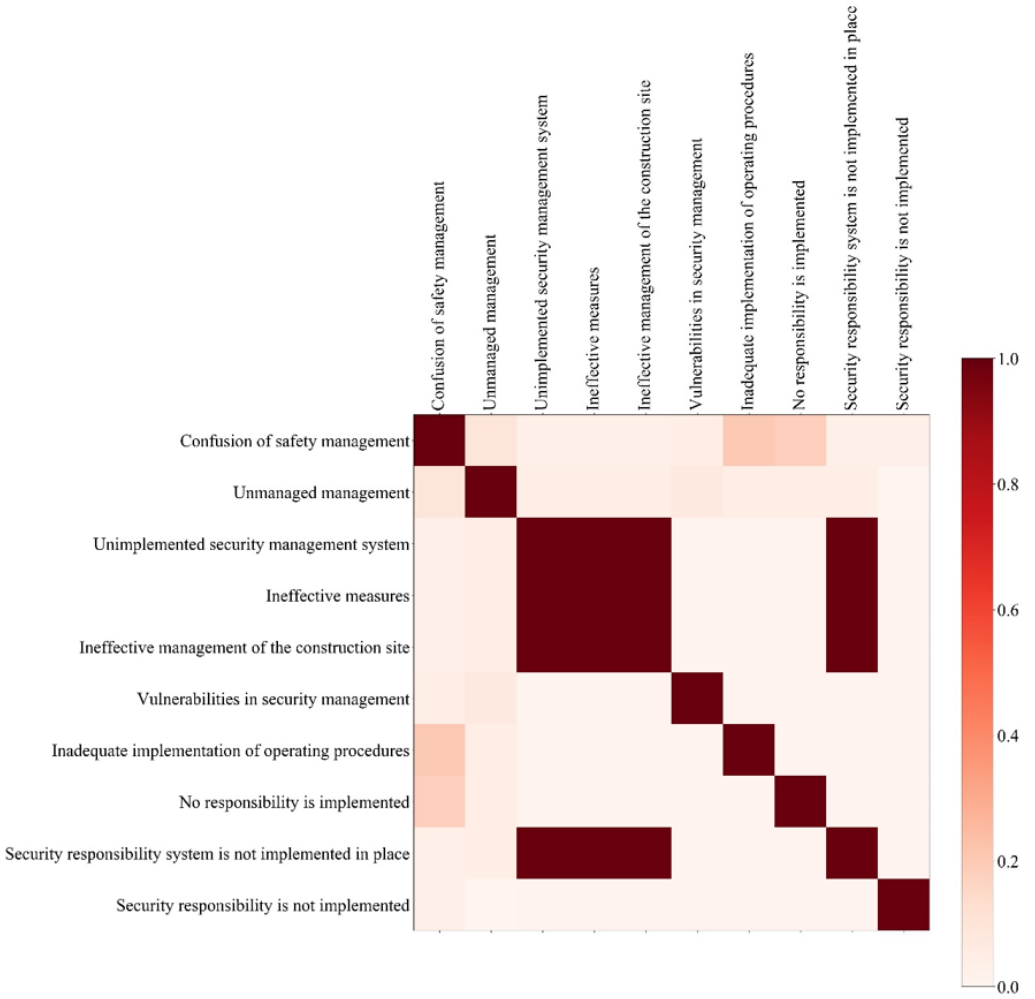
The relationship between unsafe factors and accident types can be ascertained by analyzing the correlations between unsafe factors of building safety and between unsafe factors and types of building accidents. This analysis can provide a realistic reference for the next step in building the relationships between the concepts of the ontology of unsafe factors. We used statistical linear correlation analysis and semantic correlation analysis, along with keyword analysis, to investigate the relationship between unsafe factors, and unsafe factors and types of building accidents.

4.3.1. Pearson correlation coefficient analysis

In the study, we applied the commonly used Pearson correlation coefficient to calculate the correlation of unsafe factors. The correlation coefficient R is used to express the correlation of binary variables, which ranges from -1 to 1. A positive value of R indicates that the two variables exhibit similar growth trends, while a negative value suggests opposite growth trends. The absolute value of the correlation coefficient R is used to indicate the strength of the correlation. To visualize the correlation matrix, we created

a correlation heat map that illustrates the relationships between unsafe factors. Figure 3 displays the heat map, showing a strong correlation between "Lack of management control", "Ineffective measures", "Ineffective management of the construction site", and "Security responsibility system is not implemented in place." Considering the meaning of these words, it becomes apparent that these factors share strikingly similar connotations.

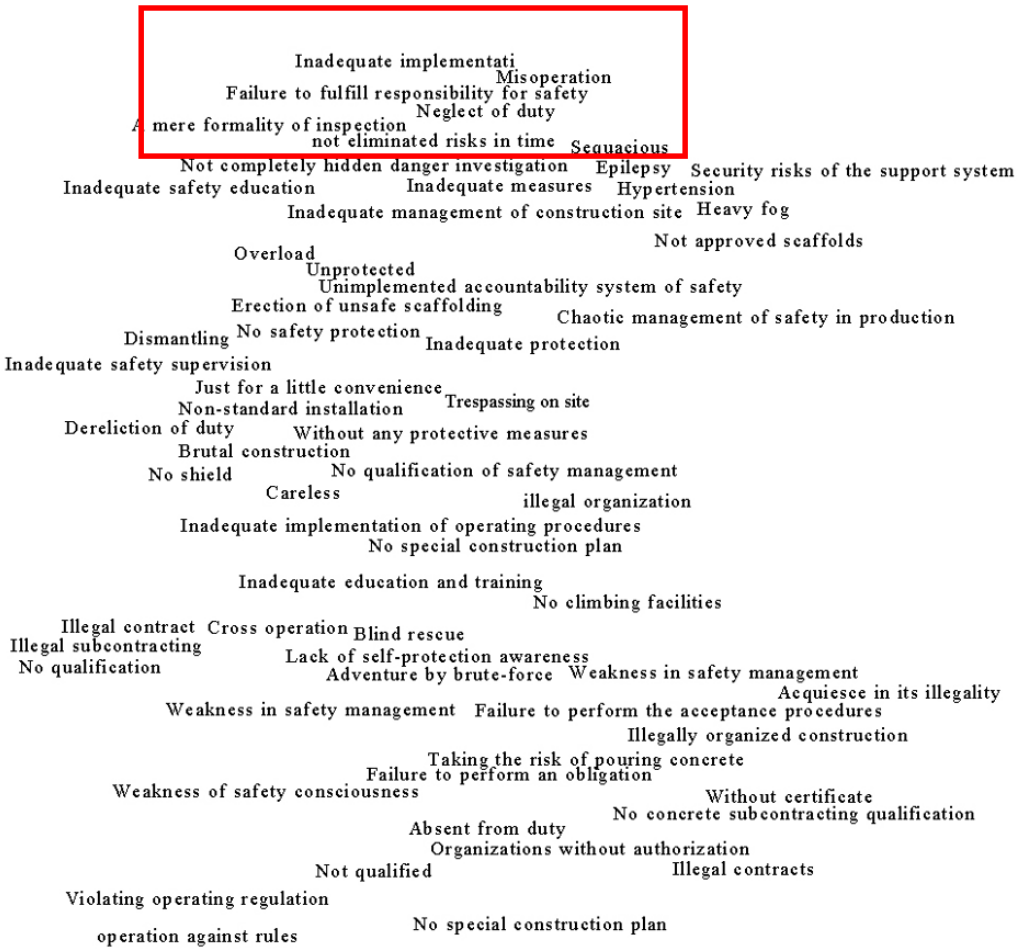
Figure 3: Correlation Heat Map of Unsafe Factors (partial)



4.3.2. Analysis of Semantic Similarity of Unsafe Factors Based on Word Vector

The vocabulary of unsafe factors is transformed into a numerical semantic space known as the word vector space. This mapping is achieved using the Word2Vec tool, which can train on a tag-free corpus to capture semantic similarity and represent vocabulary in vector form. Consequently, Word2Vec word vectors enable the expression of similarity between terms. In our study, the construction safety accident report serves as the training corpus for Word2Vec. We employ the skip-gram model with negative sampling for model training. The window size is set to 3 and the vector size is set to 20 for training purposes. By utilizing the obtained word vectors, we cluster them into several categories. The visualization of these word vectors is depicted in Figure 4. From the figure, it is evident that "Inadequate implementation", "Misoperation", "Failure to fulfill responsibility for safety", "Neglect of duty", and "A mere formality of inspection", are grouped together closely in the vector space. This cluster is highlighted with a red rectangle for emphasis.

Figure 4: Construction of ontology library of building safety accident knowledge



4.4. Construction of ontology library of building safety accident knowledge

Wang et al. (2018) combined knowledge vectors from a knowledge graph to enhance news recommendations, resulting in improved recommendation effectiveness. In this study, we leverage the ontology library of unsafe factors to generate knowledge vectors with the aim of improving the prediction of building accidents. To improve the prediction of construction accidents, we analyzed relevant industry standards and literature sources such as "Building Construction Safety Inspection Standard" (JGJ59-99), "Building Construction Enterprise Safety Production Management Code" (GB50656 - 2011), and "Safety Production Evaluation Standards for Construction Enter-

prises" (JGJ/T77-2010). These specifications and standards contain a wide range of domain-specific terms, professional concepts, and relevant technical standards.

As described above, we used the new word discovery algorithm, along with the Correlation Heat Map of Unsafe Factors and Visualization of Unsafe Factor Word Vector Space, to analyze specifications, standards, accident reports, research literature, as well as the main unsafe factor concepts and attribute relationships. This analysis enabled us to determine the primary conceptual objects and scope of the building accident ontology. The factors contributing to construction accidents were classified into human factors, physical factors, management factors, technical factors, and environmental factors. Finally, we identified the relevant conceptual level within the field of building safety. Table 1 illustrates the three-layer conceptual hierarchy for certain building accidents.

Once the relevant concepts were determined, we proceeded to define the relationships between concept classes and their attributes. To improve accident prediction, we conducted correlation analysis on the unsafe factors within the dataset and relevant literature. This analysis allowed us to organize the relationship attributes between the concept classes of unsafe factors. The attributes were defined using the Ontology Wed Language (OWL) and included type, definition domain, and value domain. The defined relationship attributes and their descriptions are shown in Table 2. After defining the relationship attributes, we supplemented the concepts relationships with the assistance of ontology relationship inference. This helped further establish relationships between the concepts.

When constructing a building safety accident ontology using Protégé (Musen, 2015), the Class editing module is used to edit the core concept classes. The Properties attribute editing module is used to add and establish attributes and relationships for the concept classes. Additionally, the OntoGraf display module is used to present the hierarchical relationships among the concept classes within the ontology library. Based on the determined concepts of building safety accident knowledge and their relationship attributes, the ontology for building safety accident unsafe factors is developed using the ontology development software Protégé 4.3. The constructed conceptual class of the building accidents ontology is shown in Figure 5, while the relationship attributes are displayed in Figure 6.

The visualization of the ontology library structure using the OntoGraf display module is presented in Figure 7. In this figure, the three levels of

First-level concept	Second-level concept	Third-level concept
Accident types	Object strike, mechanical injury, lifting injury, vehicle injury, falling from a height, collapse, container explosion, electric shock, burn, fire	High-altitude operation fall, demolition work fall, brick strike, formwork strike, earthwork collapse, formwork collapse
Accident symptoms	Horizontal rods and members are subject to continuous bending deformation, and the foundation is subject to continuous settlement and slip deformation	Water seepage, sand burst, cracks, cracks, peeling, settlement, tilt, tension, compression or shear deformation
Predisposing factors	Human factors, physical factors, management defects, environmental factors, technical factors	Insufficient qualifications of employees, insufficient safety awareness of managers, improper organization and command of managers, improper construction workers
Unsafe state	Using unsafe equipment, causing the safety device to fail, not wearing a work hat, or entering a dangerous place	Operation process design or configuration is unsafe, poor ventilation, unsafe storage method, messy operation site
Unsafe behavior	Violation of the command, violation of the physical conditions of the employees, and violation of the regulations	Not wearing safety belts when working at heights. Not wearing safety helmets or safety shoes when entering the construction site.
Cause	Power hand tools, noise, metal parts, pressure vessels, clay, sand, stone, chemicals, ladders, lifting appliances	Portal crane, jack, timber mast crane, crawler crane, forklift, deck crane

Table 1: Concept category of construction safety accident (part).

the concept class in the building safety accident ontology are displayed. The second level focuses on human factors, which serve as the primary factors in the induction of accidents. Once the ontology database was established, we were able to acquire interrelationships and domain knowledge related of unsafe factors.

Relationship attribute name	Class	Range	Explanation
subClassOf	Predisposing factor	class: Management defects	Indicates that the latter is a subclass of the former
InstanceOf	Lifting equipment	class: Forklift	Indicates that the latter is an example of the former
hasReason	Accidents	class: Predisposing factor	Indicates the cause of the accident
hasPhenomenon	Accident Symptoms	class: Accidents	Indicates the phenomenon before the accident
leadTo	Environmental factor	class: Accidents	Indicates the environment that caused the accident
hasInfluence	Environmental factor	class: Predisposing factor	Indicates the impact of the environment on the accident
hasHarmForm	Accidents	class: Damage method	Indicates the way of injury related to safety accident
Affect	Environmental factor	class: Unsafe state	Indicates the impact of the environment on unsafe conditions
relate CaseItem	Accidents	class: Cause	Indicate the cause of safety accident
relate DamageItem	Accidents	class: Harmful objects	Indicates a hazard related to a safety accident
relate InjuredPartyOfBody	Accidents	class: Injured area	Represents injured parts related to accidents
relate InjuryTpye	Accidents	class: Nature of injury	Indicates the nature of the injury related to the safety incident
relate UnsafeState	Accidents	class: Unsafe state	Indicates an unsafe state related to a security incident
relate UnsafeBehavior	Accidents	class: Unsafe behavior	Indicates unsafe behavior related to a security incident

Table 2: Attributes for Unsafe Factors in Construction Accidents.

Figure 5: Concepts in the Ontology of Construction Accidents

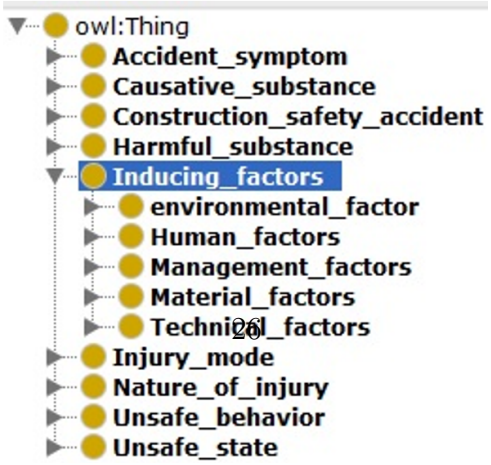


Figure 6: Concept attribute of construction safety accident

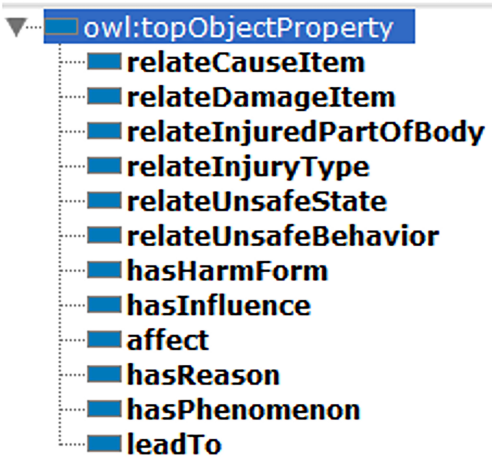
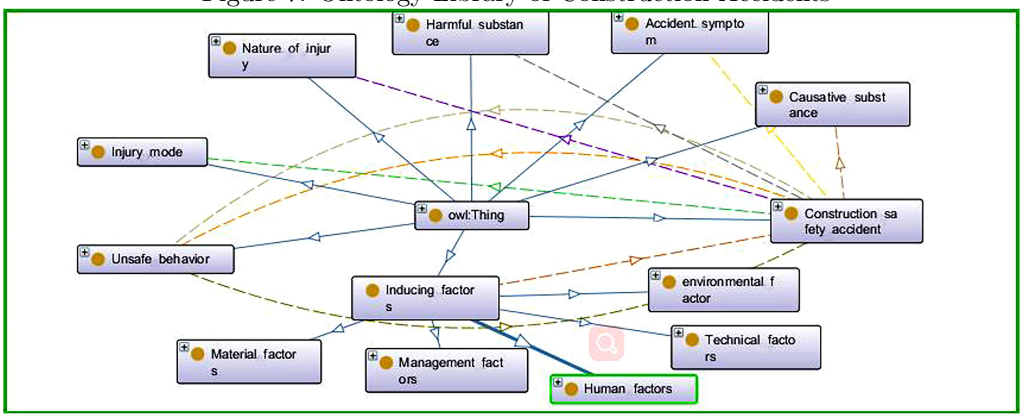


Figure 7: Ontology Library of Construction Accidents



4.5. Data Sets

4.5.1. The data set represented by One-Hot encoding

In the construction safety accident investigation report sample, the summary of the accident cause provided by experts is relatively concise, and the text paragraphs are relatively short. Due to these characteristics, using statistical word frequency as the weight may not sufficiently capture the importance of each factor. Moreover, when directly checking the presence of unsafe factors in actual usage, accident prediction can be carried out without the need for weighted input variables. In this case, a binary representation

can be used, where "1" represents the existence of the unsafe factor and "0" represent its absence. As a result, the commonly used TF-IDF method, which assigns weights to terms based on their frequency and inverse document frequency, is not employed for vectorization. Instead, the sample data from the survey report is vectorized using One-Hot encoding, where each factor is represented as a binary feature.

Initially, 280 of the construction safety accident reports were compiled into 280 records. Subsequently, a Chinese segmentation dictionary was generated using unsafe factors derived from the accident reports and field specification guidelines. The dictionary ensures improved segmentation of unsafe factors and domain concepts within the accident reports. A total of 538 keywords related to unsafe factors were identified after applying word segmentation. These keywords serve as features in the dataset. If a particular feature is present in a report, its corresponding value is set as 1; otherwise, it is represented as 0. Table 3 illustrates the application of one-hot encoding to represent the presence or absence of these features in the dataset.

Pier column	**Exterior wall**	**Retaining wall**	**...**	**Not targeted**	**Weak pertinence**	**Accident type**
0	0	0		0	0	0
0	0	0		0	0	1
0	0	0		0	0	0
0	0	0		0	0	1
0	0	0		0	0	0
0	1	0		0	1	0
0	0	0		0	0	0

Table 3: The data set represented by One-Hot encoding.

After removing empty and duplicated data, there were 259 remaining records. Due to the limited amount of data, a binary classification approach was applied to predict accidents involving high–altitude falls, which is the most common accident type. The goal was to identify whether an accident falls into this category, with a value of 1 indicating such an accident, and a value of 0 indicating otherwise. The dataset consisted of 129 records labeled as 0 (indicating accidents other than high-altitude falls) and 130 records labeled as 1 (indicating high-altitude falls). To address the issue of data imbalance, the up sampling method was utilized to balance the dataset. Consequently, a total of 260 samples were obtained, with an equal number of 130 samples for each class (0 and 1).

4.5.2. The data set with word vectors

In order to make a comparison, the dataset with word vectors was also used to test different models. Word vectors are commonly employed in text mining as they capture semantic information. By training word vectors for unsafe factors, we can obtain correlations between these factors. This information can then be used for dimensionality reduction and addressing data sparsity issues. Four different datasets were created using word2vec, each with word vectors of varying dimensions: 20, 50, 100, and 200 dimensions, respectively. These datasets enable the evaluation and comparison of models using different word vector dimensions.

4.5.3. The data set with ontology

The Resource Description Framework (RDF) serves as a formal method for describing resources and is commonly used in ontology to represent knowledge. RDF is structured as triples, specifically Subject-Predicate-Object (SPO), which capture relationships between entities. In the context of unsafe factors, concepts are represented as entities, while attribute relationships are represented as relationships between entities. In this study, the unsafe factors knowledge within the ontology was expressed as triples, forming a dictionary of concepts and their relationships. The TransH model is capable of learning various types of relationships, including one-to-one, reflexive, many-to-one, one-to-many, and many-to-many relationships. The vector training was performed using a 50-dimensional vector, which is proved to be effective by Wang et al. (2014). Following, the model training, a total of 912 50-dimensional vectors representing unsafe factors and 12 50-dimensional vectors representing relationships were obtained. Ultimately, the dataset with 50-dimensional vectors representing unsafe factors was created.

5. Experiments and Analysis

In our study, we used the feature selection methods available in the Scikit-learn library of machine learning. Specifically, we employed the SelectKBest method with chi2 for the feature selection. From a pool of 538 different unsafe factors, we selected the optimal number of features.

To establish benchmark comparison models for predicting the construction safety accident types, we employed five classic machine learning algorithms: Naive Bayes, Support Vector Machine, Logistic Regression, Random

Forest, and Multi-layer Perceptron. Cross-validation and the grid method were used to find the optimal model parameters.

Because every record for the datasets with a word vector or ontology is a two-dimensional matrix, it is not suitable as the input of traditional machine learning models. To address this limitation, we leveraged the FastText algorithm (Joulin et al., 2016), a text classification method known for its superior performance compared to other methods. In FastText, sentence vectors are formed by directly summing and averaging the word vectors. Hence, we adopted the method the FastText approach to transform the dimensionality each sample data to match the dimension of the word vector.

Finally, we applied both the traditional machine learning model and the TextCNN model to datasets with One-Hot vector, word vector, and ontology.

The matrix dataset with ontology, which includes the vector representation of unsafe factors, is treated similarly to the word vector dataset and used as input for the model. As the number of unsafe factors in each sample is limited to 100, the length of the vector is set to 100. The input matrix size of the model is defined as 50x100. After tuning, the optimal parameters for the improved model are shown in Table 4.

Parameters	**Explanation**	**Value**
embedding_dim	Vector dimension size	50
filter_sizes	Convolution kernel size	3,4,5
num_filters	The number of convolution-kernels	200
dropout_keep_prob	Probability of holding a random dropout	0.5
batch_size	Number of training samples in batch	64
Epochs	Number of data training	200
L2_reg_lambda	L2 regular	0.0

Table 4: TextCNN Model Parameters.

After feature selection, the model identified the optimal number of features as 106, resulting in an accuracy rate of 88%, surpassing the 86% achieved by the dataset with word vectors. This result indicates that deep learning models incorporating ontologies can yield superior prediction results. The learning curve of the model is shown in Figure 8, where the horizontal axis represents the number of training epochs, and the vertical axis represents the accuracy rate. The blue curve represents the accuracy rate of the training set, while the orange curve represents the accuracy rate of the testing set. The loss value curve is shown in Figure 9, with the horizontal axis representing the number of training epochs and the vertical axis representing

the loss value. The blue curve corresponds to the loss value of the training set, while the orange curve represents the loss value of the testing set.

Figure 8: TextCNN model learning curve.

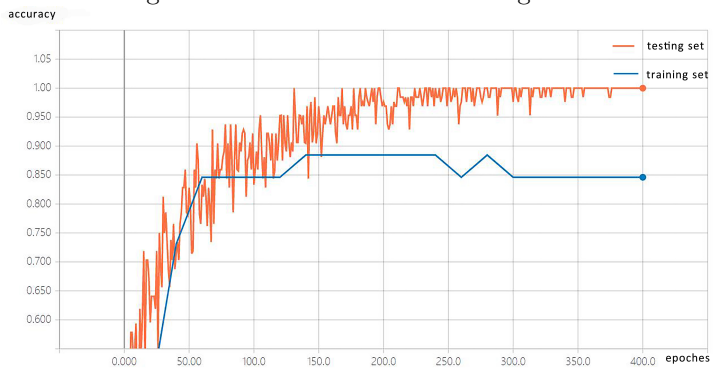
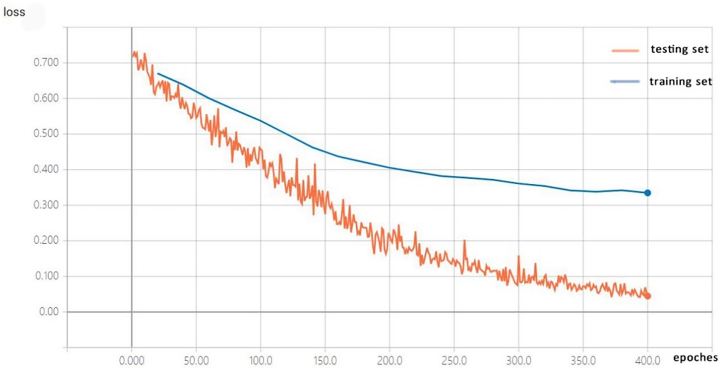


Figure 9: Ontology Library of Construction Accidents.



The six models with the best experimental results were evaluated using different datasets with varying numbers of feature dimensions as inputs. The corresponding values of Accuracy, Precision, Recall, F1-score and AUC are shown in Table 5.

In this Table, the models are represented as follows: Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Multi-layer Perceptron (MLP), and TextCNN. The datasets include: One_hot_157 which represents the one-hot encoding dataset with 157 features; Word_vec_100_118, denoting the dataset with 100-dimensional word vectors and 118 selected features; and Factor_vec_50_80, indicating the dataset with ontology with a 50-dimensional unsafe vector with 80 features.

Model	Data	Accuracy	Precision	Recall	F1-score	AUC
NB	One_hot_157	0.85	0.78	**0.98**	0.86	0.85
	Word_vec_100_118	0.82	0.86	0.78	0.81	0.87
	Factor_vec_50_80	0.82	0.87	0.76	0.80	0.89
SVM	One_hot_62	0.79	0.75	0.88	0.81	0.88
	Word_vec_50_122	0.85	0.88	0.82	0.84	0.91
	Factor_vec_50_180	0.80	0.82	0.77	0.79	0.86
LR	One_hot_160	0.85	0.85	0.86	0.85	0.91
	Word_vec_200_122	0.85	0.88	0.81	0.84	0.90
	Factor_vec_50_80	0.77	0.83	0.70	0.75	0.87
RF	One_hot_141	0.83	0.81	0.77	0.81	0.86
	Word_vec_100_103	0.81	0.88	0.74	0.80	0.88
	Factor_vec_50_80	0.77	0.81	0.74	0.76	0.83
MLP	One_hot_198	0.82	0.80	0.81	0.79	0.90
	Word_vec_50_120	0.84	0.86	0.80	0.83	0.91
	Factor_vec_50_160	0.82	0.81	0.76	0.78	0.87
Text-CNN	One_hot_538_80	0.86	0.83	0.93	0.87	0.91
	Word_vec_50_60	0.86	0.87	0.86	0.86	0.90
	Factor_vec_50_106	**0.88**	**0.92**	0.85	**0.88**	**0.92**

Table 5: Classification results of the models.

5.1. Comparative analysis for data sets

Based on the results, the dataset using One-Hot encoding yielded accuracy rates of 82% or higher for the four models. In particular, the NB and LR models achieved accuracy rates of 85%, while the TextCNN model had the highest accuracy rate at 86%. However, it is worth noting that although the NB model had the highest recall rate at 98%, its accuracy rate was the lowest among the five remaining models at 78%. Regarding the overall performance, both LR and MLP models stood out with AUC values of 0.91 and 0.90, respectively.

The SVM model and the MLP model demonstrated better performance in terms of accuracy rates when using the dataset with word vectors compared to the dataset with One-Hot encoding. The LR model and TextCNN model achieved identical results for both the word vector and One-Hot encoding datasets. As a result, the models utilizing the word vector dataset generally outperformed those using One-Hot encoding. However, it is worth noting that the improvement in performance for the models using the word vector dataset was only marginal. This may be attributed to the limited number of construction safety accident investigation reports in this study, which could have hindered significant improvements in the models leveraging the word

vector dataset. Nonetheless, the TextCNN model using the word vector dataset yielded the best results, suggesting that directly processing the two-dimensional matrix data using a CNN model is more effective than using the traditional learning methods after averaging simple word vectors. Once again, these findings highlight the superior performance of the TextCNN models.

Among the traditional machine learning models, the SVM model exhibits superior performance in both the word vector and ontology datasets compared to the One-Hot encoding dataset. The MLP model, on the other hand, demonstrates similar prediction results across all three the datasets. However, the other three models generally perform better in the One-Hot encoding dataset, except for the NB model. Notably, with the exception of the NB model, the remaining models tend to perform better when utilizing the word vector dataset rather than the ontology dataset. The TextCNN model using the ontology dataset achieves the best classification results, while the traditional models using the ontology dataset yield relatively poorer results. This discrepancy may be attributed to the process of transferring the vector data into inputs accessed by the traditional machine learning models. The simple summation method used in this process may result in the loss of important information contained within vector data.

5.2. Comparative analysis for models

In terms of the performance of traditional machine learning models, the LR model demonstrates the highest accuracy rates. It achieves an accuracy of 85% on both the One-Hot encoded and word vector datasets. The MLP and RF models follow closely behind with accuracy rates of 84% on the word vector dataset and 83% on the One-Hot encoding dataset. In terms of stability, both NBs and MLPs show reliable performance, consistently yielding accuracy rates above 80% across all three datasets. Notably, most traditional models perform better when applied to the One-Hot encoded dataset compared to the other datasets. For instance, both NB and LR algorithms achieve an accuracy rate of 85%, with the LR model demonstrating an AUC value of 0.91.

Among the various models evaluated, the deep learning model TextCNN stands out as the top performer when compared with the traditional machine learning methods across all three data sets. Notably, TextCNN achieves its best performance when using the unsafe factor vectors. In this configuration,

TextCNN yields an impressive accuracy rate of 88%, with a precision of 92% and an F1 value of 0.88. Additionally, it attains an AUC value of 0.92.

6. Discussion

Ontology, as a semantic technique, serves as a valuable tool for representing domain knowledge through formalized descriptions. By identifying core concepts, attributes, and relationships among these concepts, ontology facilitates knowledge exchange and sharing. This technique enables the conversion of textual knowledge within a specific domain into a structured, comprehensible format, thereby enhancing its usability and applicability (Studer et al., 1998).

As demonstrated in subsections 2.1–2.3, ontology finds application across diverse domains, including agriculture, medicine, education, environment, the military, the semantic web, and e-commerce (Oberle, 2014). Ontology-based risk identification techniques have also been deployed in various practical engineering scenarios to enhance accident risk prediction, assessment, and prevention. Notable applications encompass metro/subway and railway construction, chemical industry hazard evaluation, electrical grid infrastructure and natural hazard management, pipeline corridor development, oil refinery operations, and wind farm construction. Within the construction industry, ontology-based risk identification methods have been employed to enhance the comprehension and prediction of accident risks associated with construction projects. A common thread among these studies is the limited utilization of expert knowledge contained within the existing corpus of construction accident reports and occasional utilization of advanced, cutting-edge artificial intelligence methods, including deep learning, in risk identification and prediction.

Hence, construction safety remains a significant concern. The existing construction accident reports contain rich knowledge, which, if properly extracted and leveraged, can provide valuable insights for risk management. This paper describes a novel approach that focuses on extracting knowledge from these reports and building an ontology for predicting construction accidents. We collected numerous technical specifications, safety standards, and accident cases related to the field. Next, we used text mining technology to analyze unsafe factors and the relationships between them. This analysis served as the foundation for constructing an ontology that encapsulates domain knowledge regarding unsafe factors in construction accidents. Finally,

we used the TextCNN model to predict the type of accidents to demonstrate the effectiveness of our approach when compared to other commonly used models.

The results of our analysis reveal the following key findings. First, the dataset with word vectors contains more semantic information than the dataset with One-Hot encoding. Furthermore, the dataset incorporating unsafe factor concepts captures even richer knowledge than the word vector dataset. These unsafe factor concepts are represented as vectors, which are generated using the TransH model trained on triples within the ontology. Second, the deep learning TextCNN model outperforms traditional machine learning models. Its ability to harness the comprehensive knowledge contained in the complex reports enables superior performance. Notably, when combined with the dataset incorporating the ontology, the TextCNN model achieves higher accuracy rates in predicting building safety accidents, thus enhancing efficiency within the prediction process.

From a practical point of view, the use of a construction accident ontology can provide a structured and systematic approach to categorizing, understanding and preventing accidents in the construction industry. Below we have identified ten areas where such an ontology can be applied:

1. Predictive analysis: Using historical data and the structured knowledge from the ontology, predictive models can be developed to predict potential accidents, thereby enabling proactive safety measures (Halder and Batra, 2023).
2. Root cause analysis: By following the structure provided by the ontology, investigators can trace back to the root causes of accidents, enabling them to identify systemic problems rather than merely surface symptoms (Basaran and Yilmaz, 2016).
3. Incident classification: An ontology can help in classifying accidents based on various parameters such as the cause, type of injury, location on the site, etc. This classification can help systematic reporting and incident tracking (Teizer et al., 2022).
4. Data integration: Construction sites often use multiple tools and systems to collect data. An ontology can provide a common framework for integrating data from different sources, simplifying the analysis and extraction of insights (Qiu et al., 2023).
5. Knowledge sharing: An ontology can facilitate knowledge sharing among different construction projects or even different companies. The lessons

learned from one project can be methodically applied to another, reducing the learning curve and improving safety practices (Gao et al., 2022).

6. Decision support: Decision support systems (DSS) can use the ontology to provide recommendations for safety actions. For example, based on the nature and frequency of recorded accidents, the system could propose specific safety training or interventions (Zhu, 2013).
7. Policy and procedure development: The structured knowledge approach can guide policymakers in crafting safety policies and procedures that are tailored to address specific issues identified by the ontology (Shen et al., 2022).
8. Stakeholder communication: An ontology can serve as a shared language for all stakeholders, ensuring that everyone from site workers to senior management has a clear and consistent understanding of construction accidents and their consequences (Jin et al., 2019).
9. Safety Training: Educate workers using the ontology to make them aware of the different types of accidents, their causes, and preventive measures. This can make safety training more comprehensive and targeted (Pedro et al., 2023).
10. Research and development: Academics and industry researchers can use the ontology to structure their research on construction safety, ensuring that their findings are relevant and can be seamlessly integrated into industry practice (Johansen et al., 2023).

It should also be emphasized that, due to the evolving nature of knowledge, the practical implementation of a construction accident ontology should involve the engagement of experts and regular updates. In other words, it is imperative to ensure that the ontology is developed and refined with input from safety experts, site supervisors, and even workers with firsthand experience. Furthermore, as practices evolve and new types of risks emerge, the ontology should be regularly updated to reflect the current state of the industry.

Last but not least, promoting the use of the ontology can ensure that all stakeholders, especially those on site, are aware of the ontology and its benefits. Providing training or workshops to promote its consistent use, can in a broader context, guide the construction industry towards a more proactive and informed safety approach, thus reducing risks and safeguarding the well-being of its workers.

7. Conclusions

Current and previous research on construction accidents predominantly rely on existing safety measures and regulations elicited from experts and manuals, Moreover, these approaches heavily depend on expert evaluations for risk control. These methods often suffer from drawbacks such as low efficiency, insufficient intelligence, and subjectivity. Notably, a commonality among these studies is the lack of utilization of expert knowledge within existing accident reports. In contrast, construction accident reports contain valuable expert opinions that can serve as a valuable resource for extracting knowledge about accidents and improving safety management. By using this expert knowledge, we can construct a more robust, comprehensive and informed ontology. To achieve this, we intend to enhance the ontology by incorporating vector-based concepts enabling the inclusion of richer information. Additionally, we aim to crawl a larger volume of construction accident reports, further improving the accuracy rates of accident prediction.

Furthermore, we plan to employ knowledge graph techniques to analyze the relationships between unsafe factors. This analysis will facilitate the inference of accident causes, offering deeper insights into accident prevention and mitigation strategies.

In future research, our focus will be on strengthening the ontology, expanding the dataset, enhancing prediction accuracy, and utilizing knowledge graph techniques to gain a deeper understanding of the relationship between unsafe factors and accident causation. By undertaking these efforts, we aim to advance the field of construction safety management and improve accident prevention practices.

References

Aneziris, O.N., Topali, E., Papazoglou, I.A., 2012. Occupational risk of building construction. Reliability Engineering & System Safety 105, 36–46.

Aziz, A., Ahmed, S., Khan, F.I., 2019. An ontology-based methodology for hazard identification and causation analysis. Process Safety and Environmental Protection 123, 87–98.

Basaran, I., Yilmaz, S., 2016. Developing rail safety competencies based on accident and incident investigations: Using root cause taxonomies to

learn from accidents, in: 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), IEEE. pp. 481–485.

Bashar, M.A., Li, Y., 2018. Interpretation of text patterns. Data Mining and Knowledge Discovery 32, 849–884.

Benjaoran, V., Bhokha, S., 2010. An integrated safety management with construction management using 4d cad model. Safety Science 48, 395–403.

Cao, T., Mu, W., Gou, J., Peng, L., 2020. A study of risk relevance reasoning based on a context ontology of railway accidents. Risk analysis 40, 1589–1611.

Carbonari, A., Giretti, A., Naticchia, B., 2011. A proactive system for real-time safety management in construction sites. Automation in construction 20, 686–698.

Chen, W.T., Bria, T.A., 2022. A review of ontology-based safety management in construction. Sustainability 15, 413.

Chen, Y., 2015. Convolutional neural network for sentence classification. Master's thesis. University of Waterloo.

Cheng, E.W., Ryan, N., Kelly, S., 2012. Exploring the perceived influence of safety management practices on project performance in the construction industry. Safety science 50, 363–369.

Eardley, W., Ashe, D., Fletcher, B., 2016. An ontology engineering approach to user profiling for virtual tours of museums and galleries. International Journal of Knowledge Engineering 2, 85–91.

Farghaly, K., Soman, R.K., Collinge, W., Mosleh, M.H., Manu, P., Cheung, C.M., 2022. Construction safety ontology development and alignment with industry foundation classes (ifc). Journal of Information Technology in Construction 27, 94–108.

Fernández-López, M., Gómez-Pérez, A., Juristo, N., 1997. Methontology: from ontological art towards ontological engineering, in: Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series.

Fu, Y., Wen, P., Wu, J., Shu, Y., . Knowledge graph-based policy analysis from a hybrid prospect of external attributes and internal characteristics under carbon peaking and carbon neutrality goal. Available at SSRN 4384948 .

Gao, S., Ren, G., Li, H., 2022. Knowledge management in construction health and safety based on ontology modeling. Applied Sciences 12, 8574.

Ge, J., Zhang, Y., Chen, S., Xu, K., Yao, X., Li, J., Liu, B., Yan, F., Wu, C., Li, S., 2022. Accident causation models developed in china between 1978 and 2018: Review and comparison. Safety science 148, 105653.

Gruninger, M., 1995. Methodology for the design and evaluation of ontologies, in: Proc. IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing, pp. 85–91.

Hai, N., Gong, D., Liu, S., 2021. Ontology knowledge base combined with bayesian networks for integrated corridor risk warning. Computer Communications 174, 190–204.

Halder, A., Batra, S., 2023. Application of predictive analytics in built environment research: A comprehensive bibliometric study to explore knowledge domains and future research agenda. Archives of Computational Methods in Engineering , 1–26.

Huang, J.H., Powers, D., 2003. Chinese word segmentation based on contextual entropy, in: Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation, pp. 152–158.

Jiang, X., Tan, A.H., 2009. Learning and inferencing in user ontology for personalized semantic web search. Information sciences 179, 2794–2808.

Jiang, X., Wang, S., Wang, J., Lyu, S., Skitmore, M., 2020. A decision method for construction safety risk management based on ontology and improved cbr: Example of a subway project. International journal of environmental research and public health 17, 3928.

Jin, R., Zou, Y., Gidado, K., Ashton, P., Painting, N., 2019. Scientometric analysis of bim-based research in construction engineering and management. Engineering, Construction and Architectural Management 26, 1750–1776.

Johansen, K.W., Schultz, C., Teizer, J., 2023. Hazard ontology and 4d benchmark model for facilitation of automated construction safety requirement analysis. Computer-Aided Civil and Infrastructure Engineering .

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2016. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 .

Kamel, M.N., Lee, A.Y., Powers, E.C., 2007. A methodology for developing ontologies using the ontology web language (owl)., in: ICEIS (4), pp. 261–268.

Kastrati, Z., Imran, A.S., Yayilgan, S.Y., 2019. The impact of deep learning on document classification using semantically rich representations. Information Processing & Management 56, 1618–1632.

Kim, H.M., Fox, M.S., Gruninger, M., 1995. An ontology of quality for enterprise modelling, in: Proceedings 4th IEEE Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE'95), IEEE. pp. 105–116.

Knight, K., Chander, I., Haines, M., Hatzivassiloglou, V., Hovy, E., Iida, M., Luk, S.K., Whitney, R., Yamada, K., 1995. Filling knowledge gaps in a broad-coverage machine translation system. arXiv preprint cmp-lg/9506009 .

Knight, K., Luk, S.K., 1994. Building a large-scale knowledge base for machine translation, in: AAAI, pp. 773–778.

Le, Q.T., Lee, D.Y., Park, C.S., 2014. A social network system for sharing construction safety and health knowledge. Automation in Construction 46, 30–37.

Le, Q.T., Pedro, A., Park, C.S., 2015. A social virtual reality based construction safety education system for experiential learning. Journal of Intelligent & Robotic Systems 79, 487–506.

Leu, S.S., Chang, C.M., 2013. Bayesian-network-based safety risk assessment for steel construction projects. Accident Analysis & Prevention 54, 122–133.

Liu, M., Huang, R., Xu, F., 2023. Research on the construction of safety information ontology knowledge base and accident reasoning for complex hazardous production systems-taking methanol production process as an example. Sustainability 15, 2568.

Liu, X., Gao, J., He, X., Deng, L., Duh, K., Wang, Y.Y., 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval .

Lu, Y., Li, Q., Zhou, Z., Deng, Y., 2015. Ontology-based knowledge modeling for automated construction safety checking. Safety science 79, 11–18.

Luo, G., 2009. Design and evaluation of the imed intelligent medical search engine, in: 2009 IEEE 25th International Conference on Data Engineering, IEEE. pp. 1379–1390.

Macêdo, J.B., das Chagas Moura, M., Aichele, D., Lins, I.D., 2022. Identification of risk features using text mining and bert-based models: Application to an oil refinery. Process Safety and Environmental Protection 158, 382–399.

Mohamed, E., Seresht, N.G., AbouRizk, S., 2023. Context-driven ontology-based risk identification for onshore wind farm projects: A domain-specific approach. Advanced Engineering Informatics 56, 101962.

Musen, M.A., 2015. The protégé project: a look back and a look forward. AI matters 1, 4–12.

Noy, N.F., Fergerson, R.W., Musen, M.A., 2000. The knowledge model of protege-2000: Combining interoperability and flexibility, in: Knowledge Engineering and Knowledge Management Methods, Models, and Tools: 12th International Conference, EKAW 2000 Juan-les-Pins, France, October 2–6, 2000 Proceedings 12, Springer. pp. 17–32.

Oberle, D., 2014. How ontologies benefit enterprise applications. Semantic Web 5, 473–491.

Park, C.S., Kim, H.J., 2013. A framework for construction safety management and visualization system. Automation in Construction 33, 95–103.

Park, M., Lee, K.w., Lee, H.s., Jiayi, P., Yu, J., 2013. Ontology-based construction knowledge retrieval system. KSCE Journal of Civil Engineering 17, 1654–1663.

Pedro, A., Baik, S., Jo, J., Lee, D., Hussain, R., Park, C., 2023. A linked data and ontology-based framework for enhanced sharing of safety training materials in the construction industry. IEEE Access .

Pedro, A., Pham-Hang, A.T., Nguyen, P.T., Pham, H.C., 2022. Data-driven construction safety information sharing system based on linked data, ontologies, and knowledge graph technologies. International journal of environmental research and public health 19, 794.

Peraketh, B., Menzel, C., Mayer, R.J., Fillion, F., Futrell, M.T., DeWitte, P., et al., 1994. Ontology capture method (idef5). Knowledge Based Systems, Incorporated Technical report .

Phengsuwan, J., Shah, T., Sun, R., James, P., Thakker, D., Ranjan, R., 2022. An ontology-based system for discovering landslide-induced emergencies in electrical grid. Transactions on Emerging Telecommunications Technologies 33, e3899.

Pinto, A., 2014. Qram a qualitative occupational safety risk assessment model for the construction industry that incorporate uncertainties by the use of fuzzy sets. Safety Science 63, 57–76.

Pulido, J., Herrera, R., Arechiga, M., Block, A., Acosta, R., Legrand, S., 2006. Identifying ontology components from digital archives for the semantic web. IASTED Advances in Computer Science and Technology (ACST) , 1–6.

Qiu, Q., Xie, Z., Zhang, D., Ma, K., Tao, L., Tan, Y., Zhang, Z., Jiang, B., 2023. Knowledge graph for identifying geological disasters by integrating computer vision with ontology. Journal of Earth Science 34, 1418–1432.

Shen, Q., Wu, S., Deng, Y., Deng, H., Cheng, J.C., 2022. Bim-based dynamic construction safety rule checking using ontology and natural language processing. Buildings 12, 564.

Studer, R., Benjamins, V.R., Fensel, D., 1998. Knowledge engineering: Principles and methods. Data & knowledge engineering 25, 161–197.

Swartout, B., Patil, R., Knight, K., Russ, T., 1996. Toward distributed use of large-scale ontologies, in: Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems, p. 25.

Teizer, J., Johansen, K., Schultz, C., 2022. The concept of digital twin for construction safety, in: Construction Research Congress 2022, pp. 1156–1165.

Uschold, M., Gruninger, M., 1996. Ontologies: Principles, methods and applications. The knowledge engineering review 11, 93–136.

Wachter, J.K., Yorio, P.L., 2014. A system of safety management practices and worker engagement for reducing and preventing accidents: An empirical and theoretical investigation. Accident Analysis & Prevention 68, 117–130.

Wang, H., Zhang, F., Xie, X., Guo, M., 2018. Dkn: Deep knowledge-aware network for news recommendation, in: Proceedings of the 2018 world wide web conference, pp. 1835–1844.

Wang, W.C., Liu, J.J., Chou, S.C., 2006. Simulation-based safety evaluation model integrated with network schedule. Automation in construction 15, 341–354.

Wang, Z., Zhang, J., Feng, J., Chen, Z., 2014. Knowledge graph embedding by translating on hyperplanes, in: Proceedings of the AAAI conference on artificial intelligence, pp. 1112–1119.

Wu, H., Zhong, B., Li, H., Love, P., Pan, X., Zhao, N., 2021. Combining computer vision with semantic reasoning for on-site safety management in construction. Journal of Building Engineering 42, 103036.

Wu, H., Zhong, B., Medjdoub, B., Xing, X., Jiao, L., 2020. An ontological metro accident case retrieval using cbr and nlp. Applied Sciences 10, 5298.

Wu, L., Rao, Y., Yu, H., Wang, Y., Ambreen, N., 2019. A multi-semantics classification method based on deep learning for incredible messages on social media. Chinese Journal of Electronics 28, 754–763.

Xing, X., Zhong, B., Luo, H., Li, H., Wu, H., 2019. Ontology for safety risk identification in metro construction. Computers in Industry 109, 14–30.

Xu, Q., Xu, K., 2021. Analysis of the characteristics of fatal accidents in the construction industry in china based on statistical data. International journal of environmental research and public health 18, 2162.

Ye, Y., Yang, D., Jiang, Z., Tong, L., 2008. Ontology-based semantic models for supply chain management. The International Journal of Advanced Manufacturing Technology 37, 1250–1260.

Zhang, S., Boukamp, F., Teizer, J., 2014. Ontology-based semantic modeling of safety management knowledge, in: Computing in Civil and Building Engineering (2014), pp. 2254–2262.

Zhang, S., Boukamp, F., Teizer, J., 2015. Ontology-based semantic modeling of construction safety knowledge: Towards automated safety planning for job hazard analysis (jha). Automation in Construction 52, 29–41.

Zhao, L., Ichise, R., 2014. Ontology integration for linked data. Journal on Data Semantics 3, 237–254.

Zhu, Y.L., 2013. The construction safety accident emergency decision support system based on ontology and cbr. Applied Mechanics and Materials 423, 2149–2153.