frontiers in Built Environment

ORIGINAL RESEARCH
published: 10 June 2021
doi: 10.3389/fbuil.2021.690071

A Correlation Analysis of Construction Site Fall Accidents Based on Text Mining

Xixi Luo, Quanlong Liu and Zunxiang Qiu*

School of Economics and Management, China University of Mining and Technology, Xuzhou, China

Construction site fall accidents are a high-frequency accident type in the construction industry and have received extensive attention from accident causal factor analysis and risk management research, but evaluating the relationship between accident causal factors and unstructured texts remains an area in urgent need of further study. In this paper, an analysis method based on text mining was chosen to analyze and process the collected data of 557 investigation reports of construction site fall accidents in China from 2013 to 2019. First, the accident reports were preprocessed to identify six types and 28 causal factors of fall accidents; subsequently, the 28 causal factors were classified into critical causal factors, subcritical causal factors and general causal factors according to their document frequency. Then, the Apriori algorithm was used to analyze the correlation of construction site fall accidents. Finally, strong association rules were obtained between the accident causal factors and between the causal factors and the types of construction site fall accidents. The results showed that **1)** insufficient safety technology training and untimely elimination of hidden danger in safe production were the most frequent accident causal factors in fall accident reports. 2) There were different degrees of strong and weak correlations among the causal factors of construction site fall accidents, among which the higher the importance was, the stronger the correlation. 3) There were strong potential laws between the causal factors and the types of fall accidents, and the combination of some causal factors was most likely to lead to the occurrence of the corresponding accident types. This study scientifically and logically elucidated the inherent risk factors for fall accidents, which provides a theoretical basis for preventing fall accidents in construction projects.

Keywords: fall accidents, text mining, R language, accident causal factors, correlation analysis

OPEN ACCESS

Edited by:
Yongtao Tan,
RMIT University, Australia

Reviewed by:
Chansik Park,
Chung-Ang University, South Korea
Ahsan Nawaz,
Zhejiang University, China

****Correspondence:***
Quanlong Liu
lxx17602934024@163.com

Specialty section:
This article was submitted to Construction Management, a section of the journal Frontiers in Built Environment

Received: *02 April 2021*
Accepted: *25 May 2021*
Published: *10 June 2021*

Citation:
Luo X, Liu Q and Qiu Z (2021) A Correlation Analysis of Construction Site Fall Accidents Based on Text Mining. Front. Built Environ. 7:690071. doi: 10.3389/fbuil.2021.690071

INTRODUCTION

With the steady advancement of the national economy and urbanization, the development of the construction industry shows vigorous vitality and rapid expansion (Yiu et al., 2019). However, due to the rapid development of engineering construction, the continuous expansion of construction scale and the diversification of structural design have an impact on safe production practice and risk management in the process of project implementation (Choe and Leite, 2016; Hwang et al., 2018; Nawaz et al., 2019). The security problems involved in the process of infrastructure construction adversely affect the main stakeholders of a project, so improvement measures are put forward from the three levels of the government, enterprises and individuals, with the aim of identifying and

TABLE 1 | Literature review of accident cause analysis using data mining techniques.

Literature sources	Industry	Critical accident cause
Cheng et al. (2010)	Construction industry	Most accidents are found to stem from a combination of the failure of management and unsafe acts
Liao and Perng (2008)	Construction industry	The worker's age and service time, project scale and environmental factors clearly influence the safety performance at construction sites
Mirabadi and Sharifian (2010)	Railway industry	Human error, wagon and track are the most common accident causes among all factors
Chong et al. (2004)	Railway industry	The three most important factors in fatal injury are the driver's seat belt usage, the light condition of roadways, and the driver's alcohol usage
Sanmiquel et al. (2015)	Coal industry	The previous causes, place, size, physical activity, preventive organization, experience and age are the genesis of most accidents
Qiao et al. (2018)	Coal industry	Training, attendance, experience and age are the main four factors that affect the frequency of unsafe behavior, with the training factor having the greatest impact on unsafe behavior
Cheng et al. (2013)	Petrochemical industry	The quality of protective devices, pipeline design plans, and implementation of safety management measures are the critical causes of accidents
Zhao et al. (2019)	Petrochemical industry	The political conflicts, economic sanctions and warfare involving oil-producing regions are the most prominent factors
Wang and Yang (2018)	Marine industry	The key factors influencing waterway safety include the type and location of the accident and the type and age of the ship
Huang and Hu (2019)	Marine industry	Crew error and the natural causes are highly correlated with maritime accidents
Xu et al. (2018)	Transportation industry	Crashes associated with serious casualties are a result of complex interactions between road user behavior, vehicle factors, road geometric characteristics, and environmental factors
Chang and Chen (2005)	Transportation industry	The average daily traffic volume and precipitation variables are the key determinants of freeway accident frequencies

preventing security risks (Nawaz et al., 2020). At the government level, a series of laws and regulations, such as the Work Safety Law of the People's Republic of China and the Regulations on the Administration of Work Safety in Construction Projects, have been promulgated to guarantee the implementation of safety management activities in construction. At the construction-unit level, safety problems are addressed by establishing full-time safety management departments and safety management regulations. At the individual level, safe production ideas and safe operating procedures are put into practice through learning safe production skills and regularly participating in training (Liu et al., 2011). Although these measures have been implemented to improve safety in the production process, the construction industry is still one of the most dangerous industries (Pinto et al., 2011; Wu et al., 2013; Zhou et al., 2015). According to the literature (Ministry of Housing and U, 2019), there were 773 production safety accidents nationwide related to housing and municipal engineering during 2019, representing a 5.31% increase in the year-on-year accident rate; additionally, there were 904 deaths, representing a 7.62% increase in the year-on-year rate. Fall accidents accounted for 53.69% of the total number of accidents, making them the most frequently occurring type of engineering accident. In the field of safety management, additional research attention should be given to this type of accident to reduce its occurrence by addressing its causal factors.

With the advancement of the existing research on information technology and the development of multichannel integration, it has become common to acquire, store and analyze data through digital methods, and increasingly modernized technologies provide the ability to analyze unstructured data (Abbaszadegan and Grau, 2015). Previous studies have shown that text mining of unstructured data is an effective method to analyze the causes and prevention of disaster accidents (Lukic et al., 2012). By extracting effective damage information from the text files generated after the accident, the safety management efficiency can be significantly improved, and the accident can be prevented to a certain extent. Currently, text mining technology is widely used in industries with high accident rates to promote safe production management by extracting accident causal factors (Desvignes, 2014), implementing safety monitoring and early warnings (Zhu, 2014), and conducting correlation analyses (Zou et al., 2018). Some research results from the literature are presented in **Table 1**. Cheng et al. conducted a data mining analysis of 1,347 accident files from Taiwan's construction industry, listed the potential hazardous factors leading to the occurrence of the accidents, mined the relationship between these factors and the accident types, and proposed that the database of occupational accidents should be improved (Cheng et al., 2010). Mirabadi applied data mining technology to analyze traffic and railway accidents and identified the common types of hidden dangers in the context of Iranian railway transportation (Mirabadi and Sharifian, 2010). Tan Zhanglu et al. used text mining technology to explore coal safety accidents and revealed the major hidden dangers of coal mine safety and their associated relationships (Tan et al., 2017). Zhao et al. employed text mining technology in the context of risk identification in the oil market, comprehensively and effectively extracted 28 risk factors affecting this market, and constructed a risk factor assessment model for the oil market (Zhao et al., 2019).

Text mining technology is not only applied to extract the causal factors of accidents in the construction industry; for example, Fan proposed the use of text mining technology to establish a vector space model of unstructured text to more quickly retrieve construction accident dispute cases (Fan and Li, 2013). Goh et al. applied different text mining technologies to perform text extraction and classification in the context of

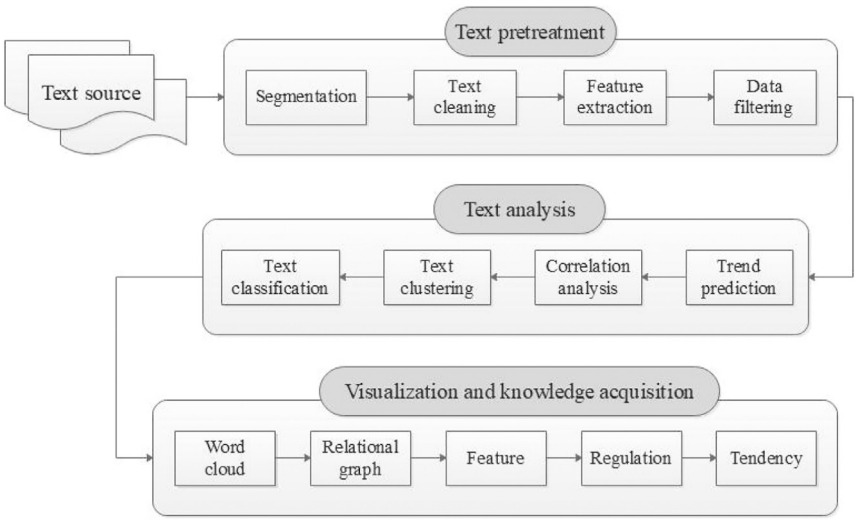

FIGURE 1 | General flow chart of the text data mining process.

construction accidents, which not only accurately classified accident reports but also confirmed that their combined classification model performed better than other models (Goh and Ubeynarayana, 2017). Li Hui performed data mining analysis of the factors associated with high-hazard collapse accidents in the building construction process, constructed a multilevel hierarchical structure model of template collapse accidents, and developed a multilevel accident causation chain (Li et al., 2018). There have been many studies on the influencing factors and improvement paths of building construction safety management; however, most of these studies are based on subjective experiences and targeted construction case data and fail to mine and analyze unstructured sets of big data with universal applicability. This paper proposes a text mining-based method for accident analysis using fall accident cases from construction projects. First, structural processing is carried out on the collected fall construction accident cases to obtain key information about each accident. Then, the association analysis method is used to analyze the obtained information, and the strong association rules between the causal factors, the accident type and the accident causal factors of the construction fall accidents are obtained to explore the internal mechanism of the frequent fall accidents and provide a reference for improving the safety management of projects.

RESEARCH METHODS AND DATA SOURCES

Text Mining Method

Text mining is an important branch of data mining, also known as text data mining, which involves the conversion of unstructured and semistructured data from large-scale text databases into digital data to extract potential useful information. The common functional applications of text mining are text semantic mining, text feature word mining, association rule mining, text clustering analysis and trend prediction (Li et al., 2017). The basic process of text mining comprises three steps: text source collection, structured text processing and text knowledge acquisition. The general process is shown in **Figure 1**. Currently, there are a wide variety of text mining tools available, including commercial tools such as Text Miner, the Intelligent Data Operating Layer (IDOL) Server and Darwin, as well as open-source data mining tools such as the library for support vector machines (LIBSVM) and the ROST content mining (CM) system. Because the R programming language provides users with a free and open development environment, users can realize specific functions by downloading various software packages according to their different usage needs, which can provide outstanding advantages in terms of data processing, statistical analysis and graphical presentation. Therefore, this study utilized R and its corresponding software packages to perform the mining analysis of the accident text (Zhang and Xu, 2012).

Structural Processing of the Accident Text

The purpose of preprocessing a text source is mainly to transform the selected text into structured data that can be processed by text mining tools. This work includes word segmentation and feature processing.

(1) Text segmentation process

When word segmentation is carried out on the collected text sources, appropriate word segmentation tools and methods should be selected to eliminate the unnecessary information in the text and more accurately extract the hidden information. As it is a basic step of text mining, the final product of the word segmentation process will affect the subsequent steps (Zhang et al., 2018a). To avoid segmentation error, which is related to the

use of a professional vocabulary and the interference of stop words, it is necessary to utilize a professional merged thesaurus specific to the field of building construction and a stop words list before word segmentation. The professional thesaurus utilized in this study is derived from professional dictionaries specific to the field of safety engineering, a construction thesaurus and construction industry terminology using the Sogou input method, which combines general words into phrases with domain-specific meanings. The merged thesaurus can identify the synonyms contained in the text, merge words with the same meaning but different expressions, and focus the extraction results. The stop words list is mainly used to remove the unnecessary content in the text and improve the search speed and accuracy of the process. A stop words list can be found on the Internet. After word segmentation, data cleaning is also required, and this step mainly involves removing the numbers, letters, and single words contained in the text. After filtering the results obtained from the initial segmentation, a segmentation list that can be used for feature extraction is obtained. The main program packages that are able to realize Chinese word segmentation in R are jiebaR and Rwordseg. To simplify the process of word segmentation, the jiebaR word segmentation package for R is used for the following operations (CSDN. Bgods.cn, 2015).

(2) Feature processing

After text segmentation, a large number of information vocabularies are generated. To select phrases that reflect the characteristics of the text, the term frequency-inverse document frequency (TF-IDF) method is used for keyword extraction and weight calculation. In 1988, Salton first proposed the statistical method of the TF-IDF algorithm, which can be used to evaluate the importance of different phrases in texts (Salton and Buckley, 1988), and it was later gradually used in the weight calculation of information retrieval and data mining. The TF-IDF algorithm can not only filter out some commonly used meaningless words but also quickly obtain important phrases with a high degree of document discrimination. The TF value in this algorithm represents the frequency of specified keywords appearing in a given document. The IDF value is calculated as the logarithm of the quotient of the total number of documents and the number of documents containing a given phrase, and it indicates the general importance of this phrase to the document. The algorithmic idea of the TF-IDF method is that many phrases may appear very frequently in an article, but a phrase that appears less frequently in the other articles within the entire examined body of text contributes the most to the text overall, as this indicates the ability of that phrase to differentiate the entire text.

$$W_{i,j} = tf_{i,j} \times idf_i, \tag{1}$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{2}$$

$$idf_i = \log\left(\frac{N}{df_i + 1}\right), \tag{3}$$

where $W_{i,j}$ represents the weight of phrase i in text j, $tf_{i,j}$ represents the word frequency of phrase i in text j, idf_i represents the inverse document frequency of phrase i in the entire document set, $n_{i,j}$ represents the number of occurrences of phrase i in text j, $\sum_k n_{k,j}$ represents the total number of words in document j, N represents the total number of documents, and df_i represents the number of documents containing phrase i.

Association Rule Analysis

Association rule analysis is the process of exploring the potential dependencies and correlations in a collected data set (Feldman and Hirsh, 1997). The process of association rule analysis is mainly divided into two steps: one step is concerned with finding frequent item sets after traversing the whole database being examined, and the other step explores the strong association rules among these frequently occurring item sets. The association rules are expressed as $A \Rightarrow B$, where A and B are, respectively, the left-hand side (LHS) and right-hand side (RHS) of the association rule, and the strength of these association rules is measured by three index values, namely, support, confidence and lift (Cui and Bao, 2016).

(1) Support

$I = \{i_1, i_2, \cdots, i_m\}$ is the item set, $D = \{t_1, t_2, \cdots, t_n\}$ is the transaction data set, and D is composed of multiple transactions. Each transaction $t_i (i = 1, 2, \cdots, n)$ contains one or several items from the item set I, which is a nonempty subset. The support degree Support$(A \Rightarrow B)$ in rule $A \Rightarrow B$ is the probability that a given item set contains both A and B, which can be expressed by the probability value $P(A \cup B)$. A high degree of support indicates that the mining results appear consistently, and the provided rules are effective association rules. A low degree of support indicates that the mining results appear only occasionally, and the provided rules have little research value. By setting a support threshold, it is possible to quickly screen out sporadic and inefficient association rules.

(2) Confidence

The confidence degree Confidence$(A \Rightarrow B)$ of rule $A \Rightarrow B$ is the probability of B where A is included in a given item set, which is represented by the conditional probability $P(B|A)$ and is expressed as follows.

$$Confidence(A \Rightarrow B) = \frac{Support(A \Rightarrow B)}{Support(A)} = \frac{P(A \cup B)}{P(A)},$$

where Support(A) is the support degree of A and $P(A)$ is the probability of A.

The confidence threshold indicates the probability that the occurrence of transaction A affects the occurrence of transaction B, and its value describes the reliability of deriving subsequent events from lead events. When the results calculated by the association rule algorithm meet the set support and confidence thresholds, the results provide strong association rules and are valuable for further mining.

(3) Lift

The lift degree $\mathrm{Lift}(A \Rightarrow B)$ of rule $A \Rightarrow B$ is expressed by the ratio of the probability that A and B are both included and the probability of the inclusion of B. The calculation is as follows.

$$\mathrm{Lift}(A \Rightarrow B) = \frac{\mathrm{Confidence}(A \Rightarrow B)}{\mathrm{Support}(B)} = \frac{P(B|A)}{P(B)}.$$

The lift degree reflects the correlation between A and B in the association rules. When the lift degree is 1, the occurrence of A is not related to that of B. When the lift degree is greater than 1, there is a positive correlation between A and B. When the lift degree is less than 1, there is a negative correlation between A and B.

Commonly used association rule algorithms include the Apriori algorithm, the Frequent Pattern (FP)-Tree algorithm, the Equivalence Class Clustering and bottom-up Lattice Traversal (ECLAT) algorithm and the gray association method; among these, Apriori is the most classical algorithm for mining frequent item sets in the context of association rule analysis (Han and Kamber, 2006; Hong et al., 2020). The Apriori algorithm was developed by Agrawal and Srikant to identify frequent item sets (Agrawal and Srikant, 2000). This algorithm employs a bottom-up search method to gradually form a candidate item set in a grid and continuously prunes the generated item set to ensure that all the subsets of the candidate set are currently known frequent item sets. Finally, an item set that satisfies the given support and confidence thresholds is obtained.

Data Sources

The construction site fall accident reports collected in this study were obtained from the Ministry of Housing and Urban-Rural Development of the People's Republic of China, the State Administration of Work Safety, the websites of various administrative departments, the safety management network, the municipal government and various safety supervision bureaus (Ministry of Housing and U, 2020; Safety management network, 2020). A total of 557 cases of construction site fall accidents in China from 2013 to 2019 were collected, focusing on major production safety accident reports and general production safety accident reports, and the cases judged as accidental casualties (nonsafety production accidents) were excluded from the analysis.

RESULTS AND DISCUSSION

The use of the text mining method to analyze construction site fall accidents encompasses two steps: extraction of accident feature information and correlation analysis of the relevant information pertaining to the accident. Text data can be transformed into structured data through extraction of accident feature information. In this process, the key information of the accident is determined mainly through the calculation of the weight parameters. On this basis, the correlation analysis between the relevant information of the accident is further explored.

Information Extraction of Construction Site Fall Accidents

Segmentation Results of the Accident Reports

After collecting 557 cases of construction fall accident reports in text format, the reports were imported into the R language processing platform. First, the custom and disabled dictionaries are configured. Then, the jiebaR package is downloaded, and the worker function is configured with a hidden Markov model (HMM) word segmentation method designed for text segmentation. Finally, 2,567 initial eigenvalues are obtained. To vividly and clearly display the results of the scored words, the Wordcloud package is downloaded to create a Wordcloud image of the resulting words, as shown in **Figure 2**.

It can be seen from the word cloud diagram in **Figure 2** that in addition to the phrases "indirect cause", "timely discovery" and "construction field", which have a high probability of occurrence but no practical research significance, the word segmentation results mainly distinguish between two types of information, namely, the causal factors and the types of construction site fall accidents.

Feature Items and Their Weight Calculations

According to the calculation steps in **Eqs 1–3**, three index values can be calculated from the phrases obtained after the word segmentation of the accident reports, namely, the TF, document frequency (DF) and term frequency-inverse text frequency (TF-IDF). The top 20 phrases in the weight calculation results are listed in **Table 2**.

The results in **Table 2** show the following:

(1) The phrase characteristic index calculation method of TF-IDF can be used to evaluate the importance of phrases to the whole document set by comprehensively considering the frequency of the phrases and the differentiation degree of the text. For example, the phrase "indirect cause", which appears the most frequently in the text, has a TF-IDF value of only 0.051, indicating that the information that this phrase contributes is not important and does not require much research. However, the phrase "hoisting machinery", which has a general frequency of occurrence, has the highest TF-IDF value, indicating that this phrase can differentiate the acquired text well and is the most important phrase for the purpose of the research.

(2) Comparing the phrases "protective equipment", "managerial staff", "safety measures" and "precaution", which have approximately equal TF-IDF values, it can be observed that the overall frequency of each phrase is consistent with its DF when the importance of the phrase is the same. It can also be said that the phrase with higher term frequencies also has higher document frequencies.

(3) The results can be roughly divided into two key groups: the causal factors of fall accidents and the types of fall accidents. By using the final weight calculation results to extract feature words, phrases that are key to the study of construction site

FIGURE 2 | Word segmentation results of the construction site fall accident reports.

TABLE 2 | Analysis results of the characteristic indexes of the phrases.

Number	Keywords	TF	TF-IDF	DF	Number	Keywords	TF	TF-IDF	DF
1	Hoisting machinery	201	0.402	87	11	Safety awareness	260	0.178	216
2	Scaffold	268	0.332	88	12	Fatality	134	0.177	116
3	Construction field	357	0.245	206	13	Security Check	102	0.173	76
4	Safety belt	338	0.221	217	14	Rules and regulations	104	0.169	89
5	Accident potential	217	0.198	148	15	Safety measures	79	0.167	64
6	Management system	233	0.192	167	16	Precaution	76	0.167	59
7	Protective equipment	187	0.188	126	17	Technical regulations	64	0.161	43
8	Managerial staff	205	0.188	145	18	Confused management	61	0.157	50
9	Operating instruction	192	0.184	140	19	Safety regulation	60	0.156	49
10	Hidden danger	162	0.181	109	20	Operation certificate	44	0.154	29

fall accidents can be scientifically obtained, which provides effective support for the rational exploration of hidden information in accident reports.

Extraction Results of the Accident Report Information

After calculating the TF-IDF value of the initial feature words in the accident reports, two types of implicit information on the causal factors and the types of fall accidents can be obtained. First, combined with the description of the phrase in the accident report and the statistical study of the accident causative mechanism (Zhang et al., 2018b), the 28 extracted causal factors of fall accidents are classified into six categories: organization and behavior, technical management, safety configuration, contract management, safety education and emergency measures. The classification results are shown in **Figure 3**.

To highlight the different degrees of importance that each causal factor has in the context of construction site fall accidents, the 28 extracted factors are divided into three levels of importance according to their document frequencies: critical causal factors, subcritical causal factors and general causal factors. **Figure 4** shows the document frequencies of the accident causal factors. The ratio of the DF of each causal factor to the total number of documents is used for classification. Based on the research results in the existing literature (Zhang et al., 2019), factors with a ratio greater than or equal to 30% are categorized as critical causal factors, factors with a ratio between 10 and 20% are identified as subcritical causal factors, and factors with a ratio below 10% are

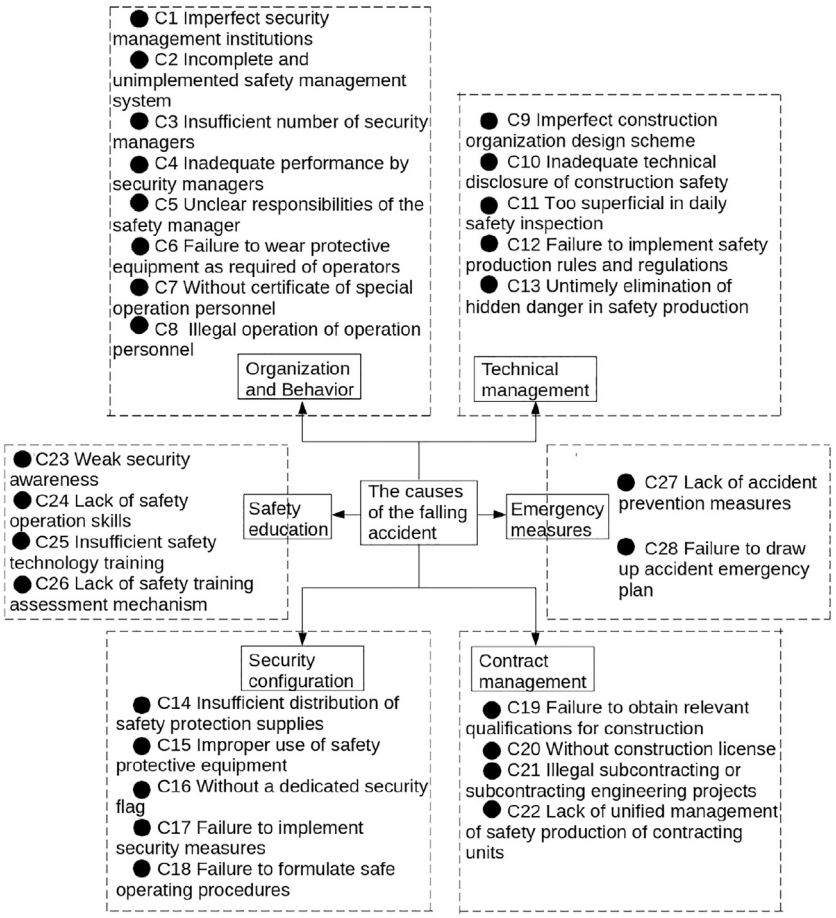

FIGURE 3 | Causal factors of construction site fall accidents.

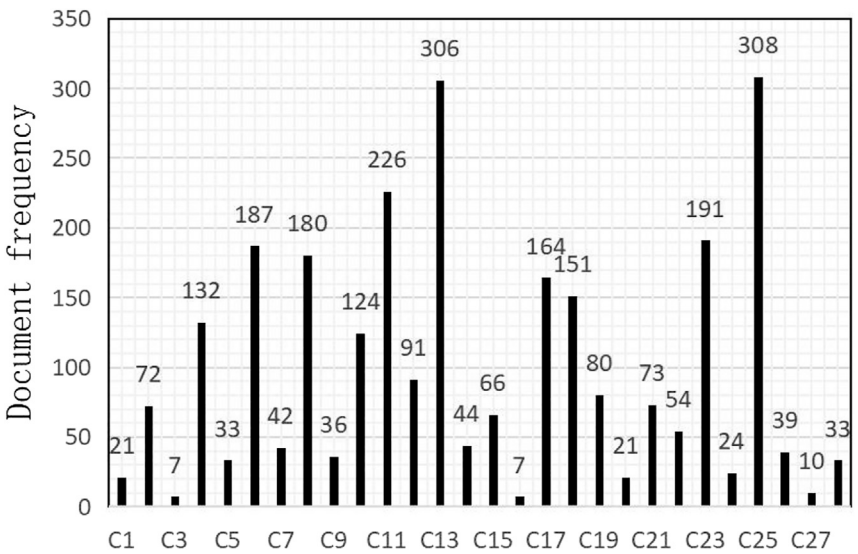

FIGURE 4 | Document frequency of the accident causal factors.

regarded as general causal factors. The following factors were identified as critical causal factors: C25 (insufficient safety technical training), C13 (untimely elimination of hidden danger in safety production), C11 (too superficial in daily safety inspection), C23 (weak safety awareness), C6 (failure to wear protective equipment as required of operators), and C8 (illegal operation by operation personnel). The subcritical causal factors were identified as follows: C17 (failure to implement security measures), C18 (failure to formulate safe operating procedures), C4 (inadequate performance by security managers), C10 (inadequate technical disclosure of construction safety), C12 (failure to implement safety production rules and regulations), C19 (failure to obtain relevant qualifications for construction), C21 (illegal subcontracting or subcontracting engineering projects), C2 (incomplete and unimplemented safety management system), and C15 (improper use of safety protective equipment). The remaining factors were classified as general causal factors. To reduce the possibility and severity of fall accidents during construction, the construction crew, supervisors, and operators should formulate preventive measures that correspond to the critical causal factors while taking into account the subcritical and general causal factors to create the most effective environment for safe production projects.

Second, the specific words describing the type of accident in the accident reports are extracted. According to the different locations of the accidents, construction site fall accidents can be divided into six types: T1 hoisting machinery equipment operations (accidents caused by lifts, hoists and other mechanical equipment), T2 fringe operations (accidents caused by operations at the edge), T3 scaffolding operations (accidents caused by operations on scaffolding), T4 opening operations (accidents caused by operations at openings), T5 suspended operations (accidents caused by operations at suspended heights) and T6 roofing operations (accidents caused by work at the roof cornice) (Kines, 2003).

After the first step of structured text information processing and feature information extraction is complete, the specific and comprehensive causal factors and types of construction site fall accidents are obtained, which provides basic support for the establishment of a correlation analysis model for text mining.

Correlation Analysis of Construction Site Fall Accidents

The above analysis process extracts and calculates the weights of the key information in the accident reports, and the extracted accident causal factors are classified. Based on these structured data, the Apriori algorithm is applied to conduct correlation analysis using R. Finally, the correlation analysis results of the fall accidents are obtained, and a visualization diagram is drawn for display.

Correlation Analysis of the Causal Factors of Construction Site Fall Accidents

The arules program package for R was downloaded; subsequently, the minimum support was set to 0.01, and the minimum confidence was set to 0.3. A series of strong association rules are obtained through a correlation analysis of the 28 causal factors of construction site fall accidents. The 10 association rules listed in **Table 3** are partial results with a confidence of one and a higher support value.

The following conclusions can be drawn from the results of the correlation analysis of the causal factors of construction site fall accidents:

(1) Limited by the length of paper, three representative rules are selected from the obtained association rules for detailed analysis. Rule 1: inadequate technical disclosure of construction safety (C10), failure to implement safety production rules and regulations (C12), and untimely elimination of hidden danger in safety production (C13) → insufficient safety technology training (C25); the degree of support for this rule is 3.23% and the confidence level is 100%. Specifically, when the reports of the construction site fall accidents include C10, C12 and C13, the problem of insufficient safety technology training is bound to exist. Rule 2: too superficial in daily safety inspection (C11), failure to formulate safe operating procedures (C18) and lack of unified management of safety production of contracting units (C22) → untimely elimination of hidden danger in safety production (C13); the degree of support for this rule is 3.05% and the confidence level is 100%. Specifically, when a construction site fall accident report includes C11, C18, and C22, the problem of untimely elimination of hidden danger in safe production is also inevitable. Rule 4: failure to implement safety production rules and regulations (C12), untimely elimination of hidden danger in safety production (C13), failure to formulate safe operating procedures (C18) and failure to wear protective equipment as required of operators (C6) → too superficial in daily safety inspection (C11); the degree of support for this rule is 2.69% and the confidence level is 100%. Specifically, when a construction site fall accident report includes C12, C13, C18 and C6, the problem of being too superficial in the daily safety inspection is also likely to exist.

(2) Some of the association rules listed in **Table 3** have a confidence level of 100%; in these cases, it is clear that the examined accident causal factors are correlated, namely, when a combination of causal factors appears in an accident report, the causal factors must be closely related. For example, the combined accident causal factors in Rule one consist of problems at the technical management level. When these problems exist, it can be inferred that there must be loopholes in the safety technical training of the construction organization. The analysis results not only verify the theoretical value of the study on the potential correlation between accident causal factors but also provide practical guidance for the rapid search for possible combinations of factors in the investigation of accident causal factors.

(3) The right-hand-side factors in the observation table are all extracted critical causal factors, and the left-hand-side factors are highly likely to be the associated combination of the

TABLE 3 | Results of the correlation analysis among the accident causal factors (part).

Number	LHS	Association	RHS	Support	Confidence	Lift
1	{C10,C12,C13}	=>	{C25}	0.0323	1	1.8084
2	{C11,C18,C22}	=>	{C13}	0.0305	1	1.8203
3	{C13,C19,C6}	=>	{C25}	0.0305	1	1.8084
4	{C12,C13,C18,C6}	=>	{C11}	0.0269	1	2.4646
5	{C18,C26}	=>	{C25}	0.0233	1	1.8084
6	{C11,C19,C23}	=>	{C25}	0.0233	1	1.8084
7	{C12,C23,C5}	=>	{C13}	0.0215	1	1.8203
8	{C13,C18,C26}	=>	{C25}	0.0215	1	1.8084
9	{C12,C23,C25,C5}	=>	{C13}	0.0215	1	1.8203
10	{C13,C15,C18,C25}	=>	{C11}	0.0197	1	2.4646

TABLE 4 | Association analysis results between the accident causal factors and the accident types (part).

Number	LHS	Association	RHS	Support	Confidence	Lift
1	{C19,C2}	=>	{T1}	0.0108	0.375	3.2637
2	{C10,C13,C18}	=>	{T1}	0.0108	0.3	2.6109
3	{C17,C6}	=>	{T2}	0.0341	0.3393	1.6724
4	{C17,C25,C6}	=>	{T2}	0.0215	0.3529	1.7397
5	{C13,C17,C6}	=>	{T2}	0.0215	0.3429	1.6900
6	{C11,C12,C13,C18}	=>	{T2}	0.0197	0.3667	1.8074
7	{C12,C13,C6}	=>	{T2}	0.0180	0.4348	2.1431
8	{C10,C11,C4}	=>	{T3}	0.0126	0.3333	2.4430
9	{C17,C9}	=>	{T4}	0.0126	0.4118	4.3274
10	{C17,C23,C4}	=>	{T4}	0.0126	0.3182	3.3439
11	{C17,C8,C9}	=>	{T4}	0.0108	0.5	5.2547
12	{C13,C25}	=>	{T5}	0.1131	0.3316	1.1471
13	{C11,C13,C19,C6}	=>	{T5}	0.0197	1	3.4596
14	{C14,C25}	=>	{T6}	0.0233	0.3611	2.7553
15	{C17,C19,C21}	=>	{T6}	0.0108	0.6	4.5781

extracted subcritical causal factors. This result not only indicates that the critical causal factors in the accident reports appear more frequently than the subcritical factors but also indicates that the critical causal factors have more complex and close connections with other causal factors.

Correlation Analysis Between the Causal Factors and the Types of Construction Site Fall Accidents

A correlation analysis of the six types of construction site fall accidents is conducted as follows, and a series of association rules are obtained. **Table 4** shows the representative partial association results of each type of accident.

From **Table 4**, we can summarize the more prominent rules between the different types of fall accidents and their causal factors:

(1) The analysis of the fall accidents caused by lifting machinery or equipment resulted in the following combinations: C19, C2, C10, C13, and C18→T1. The first combination of accident causal factors emphasizes that the examined construction unit is not qualified for their assigned work. The scale of the related company and the completeness of their rules and regulations are not sufficient to contract the project. In this case, accidents caused by hoisting machinery equipment are prone to occur. The second combination of accident causal factors emphasizes that if operators do not have adequate education and full proficiency in equipment operation, sites with potential safety hazards are likely to facilitate fall accidents caused by lifting machinery or equipment.

(2) The analysis of the fall accidents caused by fringe operations resulted in the following combinations: C17, C6, C12, C13, and C6→T2. These two combinations of accident causal factors correspond to the two situations with the highest degrees of support and confidence levels. The first combination of causal factors emphasizes that if inadequate protective measures are employed by a construction unit or if protective equipment is incorrectly worn by operators during fringe operations, fall accidents could easily occur. Such direct factors cause a relatively large number of accidents. The second combination of causal factors corresponds to a situation in which personnel protection is not complete, and the management personnel of the given construction unit have not implemented the management content of their safe production system training. In this situation, fall accidents can easily occur, as the work is carried out without the proper precautions to mitigate safety hazards. Moreover, accidents are more likely to occur when these three loopholes exist simultaneously, as the accident probability in this case is

FIGURE 5 | Rule diagram based on the grouping matrix.

43.48%. The existence of accident cause sets {C17, C25, C6}, {C13, C17, C6}, {C11, C12, C13, C18} is also a common accident cause combination that triggers fall accidents in fringe operations.

(3) The analysis of the fall accidents caused by scaffolding operations resulted in the following combination: C10, C11, and C4→T3. Specifically, most of the fall accidents that occurred during scaffolding operations are due to an insufficient technical disclosure of the relevant work types of the operators, the skills of the operators being substandard, and the safety management personnel failing to meticulously perform daily safety inspections.

(4) The analysis of the fall accidents caused by opening operations resulted in the following combinations: C17, C9, C17, C8, and C9→T4. These two combinations of accident causal factors correspond to the two situations with the highest degrees of support and confidence levels. The first combination of causal factors shows that most of the fall accidents that occur during opening operations are due to inadequate safety protection measures at the construction sites and the safety hazards that are present in the construction organization design during the early stages of the projects. The second combination of causal factors shows that when there are two accident causal factors as in the previous combination and an illegal operation conducted by the operators at the mouth of a cave, there is a very high probability of high falls, and the probability of accidents is 50%. There are other accident causal factors that are also common combinations that lead to fall accidents during opening operations. For example, {C17, C23, C4}, i.e., during the construction of engineering projects without safety protection measures, if the safety management personnel fail to take their safety management work seriously, the workers who work at the height of the hole and have less safety awareness are prone to fall accidents.

(5) The analysis of the fall accidents caused by suspended operations resulted in the following combinations: C13, C25, C11, C13, C19, C6→T5. These two combinations of accident causal factors correspond to the two situations with the highest degrees of support and confidence levels. The first combination of accident causal factors shows that most fall accidents are due to a lack of complete safety technical training for operators and the failure to eliminate hidden dangers in operating areas in a timely manner. The confidence level of the second combination of causal factors is 100%. For construction enterprises with no qualifications corresponding to their contracted projects, if the management level does not eliminate potential safety hazards in a timely manner, conduct a safety inspection, and require their operators to wear the necessary protective equipment during a suspension operation, the probability of a fall accident can reach 100%.

(6) The analysis of the fall accidents caused by roofing operations resulted in the following combinations: C14, C25, C17, C19, and C21→T6. These two combinations of accident causal factors also correspond to the two situations with the highest degrees of support and confidence levels. The first combination of causal factors shows that most of the fall accidents caused by roofing operations originate from two factors, namely, an inadequate distribution of protective equipment and a lack of safety technical training. The second combination of causal factors indicates that when the construction unit contracted for a project engages in illegal subcontracting and has insufficient construction qualifications, its ability to carry out safe production is very limited. In such a case, if safety protection measures are not strengthened, there is a high probability (up to 60%) of fall accidents during roofing operations.

Based on the association rule results of the causal factors and the types of fall accidents, it can be more clearly concluded that when a set of partial causal factors appears, it is most likely to lead to the corresponding accident types. In the process of practical safety management, managers and operators can be guided to make reasonable predictions of possible safety risks through a comparison of the collection of causal factors and the actual situation and further develop targeted risk control measures to reduce the possibility of accidents.

Visualization of the Resulting Association Rules

After the integration of all the information factors, a graphic display is created; this visual output is based on the grouping matrix constructed according to the obtained association rules, enabling a more in-depth observation of the rules and of the commonalities existing among them (Niu et al., 2020). The rules are grouped according to the clustering method, and the minimum support is set to 0.1 while the minimum confidence is set to 0.3. Frequent item sets containing 118 rules are obtained, and the arulesViz package is utilized to create a graphical visualization as shown in **Figure 5**. In this graphic display, the LHS group forms the columns, and the RHS group forms the rows. The sizes of the circles in the figure indicate the support degrees of the groups after aggregation. The depth of the circles color represents the groups lift; these colors gradually become lighter from the upper left corner to the lower right corner, indicating that the lift degree gradually decreases (Xu et al., 2018). From the visualization results in the figure, it can be seen that the rules with the highest degrees of support among the strong association rules are C4 and C18→C11, which indicates that there is a strong positive relationship between these three causative factors. The rules with the highest degrees of support are C13 and C10→C25; these factors are also critical causal factors, indicating that there is a frequent symbiotic association between the accident causal factors with a high degree of importance and that it is thus necessary to better control critical causal factors.

CONCLUSION

Fall accidents are the most frequently occurring accidents in construction, and multisource exploration of their accident causal factors and promotion paths is urgently required. In this paper, the texts of accident reports related to construction site fall accidents are obtained by collecting a sample of relevant cases; subsequently, text mining technology is used to extract text information and analyze the relevance of these accident data. Finally, the following conclusions can be drawn:

(1) The jiebaR program package and the TF-IDF statistical method are used to process the word segmentation and extract the feature words from the collected accident texts in R. Through the calculation of the TF, DF and TF-IDF, 28 accident causal factors and six accident types are identified. Based on the accident causation mechanism, the causal factors are aggregated and classified, and the 28 causal factors are finally summarized into six groups, namely, organization and behavior, technical management, security configuration, contract management, safety education and emergency measures. According to the frequency of occurrence in the text, the factors are categorized into three groups, namely, critical accident causal factors, subcritical accident causal factors and general accident causal factors.

(2) The Apriori algorithm is used to analyze the association rules of the extracted accident causal factors. After setting the minimum support degree and the confidence degree, an association analysis of the 28 causal factors of construction site fall accidents is carried out. It is observed from the results that there is a strong correlation between the causal factors. The occurrence of each accident is the result of the synergistic effect of the causal factors, and the more critical factors among the factors causing a given accident are more strongly correlated with the other factors.

(3) By analyzing the association rules of the extracted accident causal factors and accident types, a series of association rules are obtained; from these rules, the occurrence rules of the examined accidents can be obtained by analyzing the more critical rules. From the two aspects of the number and probability of project portfolios, the causal factors of fall accidents caused by lifting machinery and equipment, fringe operations, scaffolding operations, opening operations, suspended operations and roofing operations are explored to systematically analyze the selected construction site fall accidents and provide a theoretical reference for the safe production management of high-altitude operations in construction enterprises.

Combined with the research methods of text mining and association analysis, the final research results not only provide a theoretical reference with which to explore the probability of accident occurrence but also proved that the application of the automatic association analysis process can be more scientific and systematic in the study of construction accidents and broaden the methodological scope of accident causation research. This study has some limitations. First, because the combination of text mining and association analysis to analyze the text of construction site fall accidents is still in the preliminary research stage, there are still shortcomings in the use of the method, such as the dictionary of accident causal factors being incomplete. Second, reports of construction site fall accidents in China from 2013 to 2019 were chosen in the selection of data samples, with a large time span. However, the study did not take into account the continuous changes in the external environment and construction technology, which leaded to the problem that the process of accident causal factor identification did not reflect the temporal characteristics. In the follow-up study, a more detailed association analysis can be carried out after stage division according to the environmental characteristics or construction stage to render the research results more comprehensive and scientific.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

The authors gratefully acknowledge the Fundamental Research Funds for the Central Universities (Grant no. 2020ZDPYMS29). This project funding is granted by the state and all funds can be used for open access publication fees incurred in the scientific research process.

ACKNOWLEDGMENTS

We thank the Editor and Reviewers for their thoughtful suggestions during the review process.

REFERENCES

Abbaszadegan, A., and Grau, D. (2015). Assessing the Influence of Automated Data Analytics on Cost and Schedule Performance. *Proced. Eng.* 123, 3–6. doi:10.1016/j.proeng.2015.10.047

Agrawal, R., and Srikant, R. (2000). "Fast Algorithms for Mining Association Rules," in Proc. 20th Int. Conf. Very Large Data Bases VLDB, San Jose, CA (Santiago de Chile, Chile: IBM), 1215.

Chang, L.-Y., and Chen, W.-C. (2005). Data Mining of Tree-Based Models to Analyze Freeway Accident Frequency. *J. Saf. Res.* 36 (4), 365–375. doi:10.1016/j.jsr.2005.06.013

Cheng, C.-W., Lin, C.-C., and Leu, S.-S. (2010). Use of Association Rules to Explore Cause-Effect Relationships in Occupational Accidents in the Taiwan Construction Industry. *Saf. Sci.* 48 (4), 436–444. doi:10.1016/j.ssci.2009.12.005

Cheng, C.-W., Yao, H.-Q., and Wu, T.-C. (2013). Applying Data Mining Techniques to Analyze the Causes of Major Occupational Accidents in the Petrochemical Industry. *J. Loss Prev. Process Industries* 26 (6), 1269–1278. doi:10.1016/j.jlp.2013.07.002

Choe, S., and Leite, F. (2016). Assessing Safety Risk Among Different Construction Trades: Quantitative Approach. *J. Construction Eng. Manag.* 143 (5), 04016133. doi:10.1061/(ASCE)CO.1943-7862.0001237

Chong, M. M., Abraham, A., and Paprzycki, M. Traffic Accident Analysis Using Decision Trees and Neural Networks (2004). Available online at: http://arxiv.org/ftp/cs/papers/0405/0405050.pdf [Accessed June 2004].

CSDN. Bgods.cn (2015). jiebaR Chinese Word Segmentation and Word Cloud (R Language). https://blog.csdn.net/songzhilian22/article/details/49184047 [Accessed October 16, 2015].

Cui, Y., and Bao, Z. Q. (2016). Summary of Association Rule Mining. *Appl. Res. Comput.* 33 (2), 330–334. doi:10.3969/j.issn.1001-3695.2016.02.002

Desvignes, M. (2014). "Requisite Empirical Risk Data for Integration of Safety with Advanced Technologies and Intelligent Systems,". [dissertation/master's thesis] (Boulder, CO: University of Colorado at Boulder).

Fan, H., and Li, H. (2013). Retrieving Similar Cases for Alternative Dispute Resolution in Construction Accidents Using Text Mining Techniques. *Automation in Construction* 34 (sep), 85–91. doi:10.1016/j.autcon.2012.10.014

Feldman, R., and Hirsh, H. (1997). "Finding Associations in Collections of Text," in *Machine Learning and Data Mining: Methods and Applications*. Editors R. S. Michalski, I. Bratko, and M. Kubat (New York, NY: J. Wiley), 223–240.

Goh, Y. M., and Ubeynarayana, C. U. (2017). Construction Accident Narrative Classification: An Evaluation of Text Mining Techniques. *Accid. Anal. Prev.* 108, 122–130. doi:10.1016/j.aap.2017.08.026

Han, J., and Kamber, M. (2006). *Data Mining: Concepts and Techniques*. 2nd ed. USA: Morgan Kaufmann Publishers.

Hong, J., Tamakloe, R., and Park, D. (2020). Application of Association Rules Mining Algorithm for Hazardous Materials Transportation Crashes on Expressway. *Accid. Anal. Prev.* 142, 105497. doi:10.1016/j.aap.2020.105497

Huang, C. H., and Hu, S. P. (2019). Factors Correlation Mining on Maritime Accidents Database Using Association Rule Learning Algorithm. *Cluster Comput.* 22 (2), 4551–4559. doi:10.1007/s10586-018-2089-z

Hwang, B.-G., Shan, M., and Phuah, S. L. (2018). Safety in green Building Construction Projects in singapore: Performance, Critical Issues, and Improvement Solutions. *KSCE J. Civ Eng.* 22 (2), 447–458. doi:10.1007/s12205-017-1961-3

Kines, P. (2003). Case Studies of Occupational Falls from Heights: Cognition and Behavior in Context. *J. Saf. Res.* 34 (3), 263–271. doi:10.1016/s0022-4375(03)00023-9

Li, H., Zhang, Y. B., and Qi, S. J. (2018). Cause Analysis and Countermeasure of Building Construction Collapse Accident. *Construction Econ.* 039 (008), 53–57. doi:10.14181/j.cnki.1002-851x.201808053

Li, J., Wang, J. P., Xu, N., and Zhou, Z. (2017). Analysis of Risk Factors for Subway Construction Safety Risk Accidents Based on Text Mining. *Tunnel Construction* 2, 160–166. doi:10.3973/j.issn.1672-741X.2017.02.006

Liao, C.-W., and Perng, Y.-H. (2008). Data Mining for Occupational Injuries in the Taiwan Construction Industry. *Saf. Sci.* 46 (7), 1091–1102. doi:10.1016/j.ssci.2007.04.007

Liu, H., Lu, M., and Shang, M. (2011). The Problems and Innovation Strategies of Technological Innovation in China's Construction Industry. *J. Eng. Manag.* 25 (004), 359–363. doi:10.3969/j.issn.1674-8859.2011.04.001

Lukic, D., Littlejohn, A., and Margaryan, A. (2012). A Framework for Learning from Incidents in the Workplace. *Saf. Sci.* 50 (4), 950–957. doi:10.1016/j.ssci.2011.12.032

Ministry of Housing and Urban-Rural Development of the People's Republic of China. Announcement of the Ministry of Housing and Urban-Rural Development on the Production Safety Accidents of Housing and Municipal Engineering in 2019 (2019). http://www.mohurd.gov.cn/wjfb/202006/t20200624_246031.html [Accessed June 19, 2020].

Ministry of Housing and Urban-Rural Development of the People's Republic of China (2020). Safety Accident Report of Ministry of Housing and Urban and Rural Construction. http://www.mohurd.gov.cn/zlaq/cftb/zfhcxjsbcftb/index.html [Accessed October 25, 2020].

Mirabadi, A., and Sharifian, S. (2010). Application of Association Rules in Iranian Railways (RAI) Accident Data Analysis. *Saf. Sci.* 48 (10), 1427–1435. doi:10.1016/j.ssci.2010.06.006

Nawaz, A., Su, X., Din, Q. M. U., Khalid, M. I., Bilal, M., and Shah, S. A. R. (2020). Identification of the H&S (Health and Safety Factors) Involved in Infrastructure Projects in Developing Countries-A Sequential Mixed Method Approach of OLMT-Project. *Ijerph* 17 (2), 635. doi:10.3390/ijerph17020635

Nawaz, A., Waqar, A., Shah, S., Sajid, M., and Khalid, M. (2019). An Innovative Framework for Risk Management in Construction Projects in Developing Countries: Evidence from pakistan. *Risks* 7 (1), 24. doi:10.3390/risks7010024

Niu, Y., Li, Z. M., and Fan, Y. X. (2020). Research on Correlation Analysis of Influencing Factors of Highway Truck Traffic Accidents Based on Data Mining. *Saf. Environ. Eng.* 11 (4), 495–498. doi:10.13578/j.cnki.issn.1671-1556.2020.04.024

Pinto, A., Nunes, I. L., and Ribeiro, R. A. (2011). Occupational Risk Assessment in Construction Industry - Overview and Reflection. *Saf. Sci.* 49 (5), 616–624. doi:10.1016/j.ssci.2011.01.003

Qiao, W., Liu, Q., Li, X., Luo, X., and Wan, Y. L. (2018). Using Data Mining Techniques to Analyze the Influencing Factor of Unsafe Behaviors in Chinese Underground Coal Mines. *Resour. Pol.* 59, S0301420718302253. doi:10.1016/j.resourpol.2018.07.003

Safety management network (2020). http://www.safehoo.com/Case/Case/Drop/ [Accessed September 15, 2020].

Salton, G., and Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* 24 (5), 513–523. doi:10.1016/0306-4573(88)90021-0

Sanmiquel, L., Rossell, J. M., and Vintró, C. (2015). Study of Spanish Mining Accidents Using Data Mining Techniques. *Saf. Sci.* 75, 49–55. doi:10.1016/j.ssci.2015.01.016

Tan, Z. L., Chen, X., Song, Q. Z., and Chen, X. C. (2017). Coal Mine Safety Hazards Analysis Based on Text Mining. *J. Saf. Environ.* 4, 1262–1266. doi:10.13637/j.issn.1009-6094.2017.04.009

Wang, L., and Yang, Z. (2018). Bayesian Network Modelling and Analysis of Accident Severity in Waterborne Transportation: a Case Study in china. *Reliability Eng. Syst. Saf.* 180, 277–289. doi:10.1016/j.ress.2018.07.021

Wu, W., Yang, H., Li, Q., and Chew, D. (2013). An Integrated Information Management Model for Proactive Prevention of Struck-By-Falling-Object Accidents on Construction Sites. *Automation in Construction* 34, 67–74. doi:10.1016/j.autcon.2012.10.010

Xu, C., Bao, J., Wang, C., and Liu, P. (2018). Association Rule Analysis of Factors Contributing to Extraordinarily Severe Traffic Crashes in china. *J. Saf. Res.* 67, 65–75. doi:10.1016/j.jsr.2018.09.013

Yiu, N. S. N., Chan, D. W. M., Shan, M., and Sze, N. N. (2019). Implementation of Safety Management System in Managing Construction Projects: Benefits and Obstacles. *Saf. Sci.* 117, 23–32. doi:10.1016/j.ssci.2019.03.027

Zhang, F., Fleyeh, H., Wang, X. R., and Lu, M. H. (2018). Construction Site Accident Analysis Using Text Mining and Natural Language Processing Techniques. *Automation in Construction* 99, 238–248. doi:10.1016/j.autcon.2018.12.016

Zhang, W., and Xu, X. (2012). A Review of Text Mining Tools. *Libr. Inf. Serv.* 56 (8), 26–55.

Zhang, W., Zhu, S. N., Cao, C. X., and Wu, X. G. (2018). The Cause Mechanism and Case Data Statistics of Construction Safety Accidents. *J. Eng. Manag.* 032 (003), 92–96. doi:10.13991/j.cnki.jem.2018.03.017

Zhang, W., Zhu, S. N., Zhang, X., and Zhao, T. S. (2019). System Model and Empirical Analysis of the Cause of Construction Safety Accidents. *Chin. Saf. Sci. J.* 29 (06), 60–66.

Zhao, L.-T., Guo, S.-Q., and Wang, Y. (2019). Oil Market Risk Factor Identification Based on Text Mining Technology. *Energ. Proced.* 158, 3589–3595. doi:10.1016/j.egypro.2019.01.906

Zhou, Z., Goh, Y. M., and Li, Q. (2015). Overview and Analysis of Safety Management Studies in the Construction Industry. *Saf. Sci.* 72, 337–350. doi:10.1016/j.ssci.2014.10.006

Zhu, H. (2014). "Research on Early Warning of Operational Risk Based on Association Rules Apriori Algorithm,". [dissertation/master's thesis] (Jilin: . Jilin University).

Zou, Y., Xiao, Z., Zhang, L., Zio, E., Liu, J., and Jia, H. (2018). A Data Mining Framework within the Chinese Npps Operating Experience Feedback System for Identifying Intrinsic Correlations Among Human Factors. *Ann. Nucl. Energ.* 116, 163–170. doi:10.1016/j.anucene.2018.02.038

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.